SINGLE-CELL TRANSCRIPTOMIC VISUALIZATION AND ANALYSIS OF GENETIC CHARACTERIZATION OF ALZHEIMER'S DISEASE

PROJECT OF MAT 6493

Shuang Ni **Nishka Katoch** **Timur Zhanabaev**

December 14, 2022

1 Introduction

Alzheimer's disease (AD) is a progressive neurological disorder that causes the brain to shrink and brain cells to die, which impairs cognition, memory, and language and causes dementia. The genetic basis of AD is complex, with both inherited and non-inherited risk factors contributing to the development of the disease. To understand the disease, the well-studied biochemistry needs to be integrated into the complex cellular context of the brain.

1.1 Motivation

The genetic basis of AD is complex, with both inherited and non-inherited risk factors contributing to the development of the disease. Further research is needed to fully understand the genetic and environmental factors that contribute to the development of AD and to identify potential therapeutic targets for the treatment and prevention of this devastating disease. Single-cell RNA sequencing (scRNA-seq) provides a way to study the cellular heterogeneity of the brain, by profiling tens of thousands of individual cells.

The motivation for this project came from the analysis of single-cell transcriptomics for Alzheimer's disease [6]. While previous Alzheimer's disease analysis focused on neurons and microglia (the immune cells of the brains made to respond to pathogens and damage), this paper studies the transcriptional signatures across six celltypes - excitatory neurons (EX), inhibitory neurons (In), astrocytes(Ast), oligodendrocytes (Oli), oligodendrocyte precursor cells (Opc), and microglia (Mic). They also delve into the 40 distinct subpopulation of cells. Overall the paper resulted in observing the complexity of glial-neuronal interactions in AD pathalogy by. Hence by studying the various cell-types we could infer more about the pathology of AD.

Dimensionality reduction and data visualization is an essential task in high-dimensional data analysis, which facilitates description, exploration and understanding from raw data and analysis results. Our motivation was to analyze the dataset in [6] by different visualization methods discussed further on to explore some pathology of AD from the data.

1.2 Goals

The goal of this project is to characterizing the complex cellular changes in AD brain pathology by visualization. We will analyze 80,660 single-nucleus transcriptomes from the prefrontal cortex of 48 individuals with varying degrees of Alzheimer's disease pathology. The dataset includes the metadata of eight major brain cell types, sub-types, sex, diagnosis, course of disease, and so on. Our goal is to reduce the dimension of data using multiple algorithms and visualize them by labeling with different metadata. With analyzing the generated figures, we can zoom in to some part of interest and do more analysis in that subset. For the feature analysis component, we want to find some parts of interest that make a conclusion of what kind of genes or cells are more important in the pathology of AD.

In this project, multiple visualizing methods were implemented, including kernal PCA [10], Isomap [11], t-SNE [13], UMAP [7], PHATE [8] and Multiscale PHATE [5]. These methods are described in detail in Section 3. The data preprocessing procedure is described in Section 2. And the analyzing results based on the preprocessed data for each method are shown in Section 4.

1.3 Data Source

Single-cell transcriptomes are used to study cellular heterogeneity of the brain. Transcriptomics is the full range of messenger RNA, or mRNA, molecules expressed by an organism. Unlike the genome's stability, transcriptomes actively change and hence can provide useful information. In this project, single-nucleus RNA-seq (snRNA-seq) profiling of prefrontal cortex are used, since this region displays the major traits affected by AD. The snRNA-seq dataset is post-mortem human brain samples came from 48 participants in the Religious Order Study (ROS) or the Rush Memory and Aging Project (MAP), two longitudinal cohort studies of ageing and dementia [1], collectively known as ROSMAP. The data are available under controlled use conditions set by human privacy regulations. It is available on The Rush Alzheimer's Disease Center (RADC) Research Resource Sharing Hub at `https://www.radc.rush.edu/docs/omics.htm` (snRNA-seq PFC). Hansruedi Mathys et al. profiled droplet-based single-nucleus cortical transcriptomes across 48 individuals with varying degrees of AD pathology and both sexes[6]. It is published on ROSMAP data compendium, which can be accessed at `https://www.synapse.org/#!Synapse:syn3157322`.

2 Data Preprocessing and Explanation

The original dataset contained 82159 cells with 32738 genes, which were combined into a single dataset. A large-scale single-cell gene expression data analysis package, SCANPY[15], was used to preprocessing the dataset.

Cells with a high ratio of mitochondrial relative to endogenous RNAs had low starting amounts of RNA, which might indicate that source cells were dead or stressed and thus result in RNA degradation[6]. Therefore, we removed cells with fewer than 200 detected genes and cells with an abnormally high ratio of counts mapping to mitochondrial genes relative to the total number of detected genes. Since only counts associated with protein-coding genes were considered, mitochondrially encoded genes and genes detected in fewer than 2 cells were excluded.

Additionally, some common data preprocessing steps were performed on the single-cell data. First, normalize each cell by total counts over all genes, so that every cell has the same total count after normalization. Then, the normalized dispersion is obtained by scaling with the mean and standard deviation of the dispersions for genes falling into a given bin for mean expression of genes. This means that for each bin of mean expression, highly variable genes are selected. Next, the data of each cell was transformed to log space by computes $X = \ln(X + 1)$. Finally, regress out mostly unwanted sources of variation. A total of 3,188 highly variable genes were identified based on dispersion and mean, the technical influence of the total number of counts was regressed out, and the values were rescaled. These preprocessing steps were performed by using the functions normalize_per_cell, filter_genes_dispersion, log1p, and regress_out in SCANPY. The cells that have a distinctly high number of total counts and mixed expression of markers from different cell types were tagged as potential doublets and not considered for downstream analyses. After applying these preprocessing steps, the dataset included 17,926 genes profiled in 70,634 nuclei.

The dataset also includes several metadata for each cell, such as cell types, subclusters, and clinical data, such as diagnosis, braak stages, APOE genotype, CERAD score, death age, and sex of the patients. The cell types and subclusters were profiled by Mathys et. al [6]. They assigned each pre-cluster a celltype label using statistical enrichment for sets of marker genes. There are 8 celltypes: Excitatory neurons (Ex), Inhibitory neurons (In), Oligodendrocytes (Oli), Microglia (Mic), Astrocytes (Ast), Oligodendrocyte progenitor cell (Opc), endothelial cells (End), and pericyte cells (Per). To be noted that, endothelial cells (End), and pericyte cells (Per) have low cell counts. The braak staging is used to classify the degree of pathology in AD, which was described by Braak in 1991 [2]. Six stages of disease propagation can be distinguished with respect to the location of the tangle-bearing neurons and the severity of changes: transentorhinal stages 1–2: clinically silent cases; limbic stages 3–4: incipient AD; neocortical stages 5–6: fully developed AD [3]. The apolipoprotein E (APOE) genotype is the clearest genetic risk factor that has been associated with non-familial or sporadic for AD. The APOE gene plays a role in the metabolism of cholesterol and other lipids, and the epsilon 4 allele is associated with an increased risk of developing AD. It is among the more common genotypes, which includes 6 genotypes (e22, e23, e24, e33, e34, e44). The Consortium to Establish a Registry for Alzheimer's Disease (CERAD) was funded by the National Institute on Aging in 1986 to develop standardized, validated measures for the assessment of AD [9]. A value of 1 means that the patient definitely has AD, 2 means probable, 3 means possible, and 4 means no AD diagnosis. For the gender of the patients, 1 represents male and 0 represents female. For diagnosis, value of 1 means the patient was diagnosed with AD and 0 means not diagnosed with AD.

3 Methodologies

A large number of dimensionality reduction and visualization methods have been proposed. In this project, we implemented both classical and state-of-the-art visualization methods for biomedical high-dimensional data.

3.1 t-SNE

In 2008, t-distributed stochastic neighbor embedding (t-SNE) [13] was proposed by Maaten and Hinton, which could capture both the local and the global structures. And in 2014, Maaten further proposed an acceleration technique for the t-SNE by first constructing a K-nearest neighbor (KNN) graph of the data points and then projecting the graph into low-dimensional spaces with tree-based algorithms [12].

The process of t-SNE requires the computation of similarities between pairs of data points in K dimensions using a distribution (*Student t distribution in this case*), with points closer to each other generating higher probabilities. The Kullback–Leibler divergence is formulated between the distributions of K and 2 dimensions, the lower dimension being the target. The minimization of KL then learns the 2-D embedding for visualizations purposes.[13]

3.2 Kernel Principle Component Analysis (kPCA)

Following the kernel SVM ideas of lifting the features to a higher dimensional feature space to create a linearly separable data set, a kernel trick is used in the eigen decomposition of the kernel matrix for PCA. Defining the feature mapping $\Phi(\vec{x})$, the kernel matrix is,

$$K = k(\vec{x}, \vec{y}) = (\Phi(\vec{x}), \Phi(\vec{y}))$$

With the eigenvector equations to solve; λ being the eigenvalues and $\vec{u}$ the eigenvectors and $N$the number of samples,

$$N\lambda\vec{u} = K\vec{u}$$

All we need again is only the kernel to find the reduced space of the dataset [10], with the rest of the PCA methodology following as before. An important thing to note is that there is no assumption of scaled feature maps, since it is never computed to calculate the means for scaling [10]. Although, the scaling will be performed on the original data set.

3.3 Isomap

The last basic method, before dwelling into UMAP and PHATE, is the Isomap method. It can be generally viewed as an extension to kernel PCA and Multidimensional Scaling (MDS). The general procedure for Isomap involves a graph construction from k nearest neighbors for each point, where the edges are then formed by distance metrics being below a certain threshold with the edge weight generated from a distribution function. Therefore, the generated geodesic distance matrix from Isomap can be seen as the kernel that the eigenvalue decomposition is performed on, as in MDS and k-PCA.

3.4 UMAP

Uniform Manifold Approximation and Projection (UMAP) [7] is a manifold learning technique for dimension reduction, which is proposed by McInnes et al. in 2018. Compared with t-SNE, it preserves more of the global structure with superior runtime performance. It is based on the mathematical foundation of Laplacian eigenmaps. This method constructs a high-dimensional representation and this representation is optimized in the low-dimensional space.

The high-dimensional representation is constructed by approximating the manifold and then constructing the fuzzy simplicial set of the manifold. This simplicial representation is a graph made by extending the radius of each point and connecting the points where the radius overlaps. This radius is chosen locally based on the distance of each point n^{th} nearest neighbor. The graph is then made fuzzy by decreasing the likelihood of the connection as the radius grows. These connections balance the preservation of local structure versus the preservation of the global structure.

The high-dimensional representation is optimized in a similar fashion to t-SNE with a few differences that make it faster in computation and preserves more space.

3.5 PHATE

PHATE is a dimensionality reduction method used to generate a denoised low-dimensional embedding of high-dimensional data [8]. This embedding preserves the local and global structure of the original dataset without imposing assumptions on the structure of the data. Methods such as PCS and Isomap fail to preserve the local structure of data while other methods such as t-SNE tend to scramble the global structure of data. PHATE provides a way to preserve the local and global structure in the data for better visualization and inference. The local structure is preserved by converting pairwise distances of points to affinities. These affinities are inversely proportional to distance so it helps to encode the local structure even if it is not sampled uniformly in the manifold. The global structure of the data is encoded by heat diffusion. Heat diffusion also denoises the data.

The steps involved in PHATE algorithm are:

a) Pairwise distances often Euclidean metric of the data is calculated. When using distance metrics like Euclidean metric, with a large number of dimensions the clusters can appear equidistant and compromises on the local structure and meaningful patterns.

b) Affinities are introduced to solve the dimensionality issue stated above. Affinities are calculated by kernel functions, in this algorithm α - decay kernel is used to calculate affinities.

c) This diffusion is performed by transforming affinities to the probability that measures the probability of one data point to another in a single step of the walk. These manifold distances are presented as multistep transition probabilities called diffusion probabilities.

This probability is calculated as:

$$P = \frac{k(x, y)}{\sum_{z \in X}(x, z)}$$

where k is the calculated affinities. These probabilities are considered to be 'global context' and a distance called potential distance is used to describe the relationship of each point to both near neighbors and distant points. This distance is calculated as:

$$d_{ij} = \sqrt{|| \log P_i - \log P_j ||^2}$$

d) Calculated potential distance information is now embedded into low dimensions using metric Multidimensional Scaling (MDS).

3.6 Multiscale PHATE

Biomedical datasets are usually high-throughput and high-dimensional, requiring multi-level analyzes ranging from coarse, high-level summaries to fine-grained, detailed representations of data subsets. The key to understanding complex data, such as biomedical data, is to create meaningful representations that reveal structure at all resolutions or scales. Most of tools for dimensionality reduction and data exploration only show a single level of granularity of the data. Multiscale PHATE [5] provides a possibility of sweeping through all levels of data granularity to learn abstracted biological features directly predictive of disease outcome. It is based on a dynamic topological process called diffusion condensation, which slowly condenses data points toward local centers of gravity to form natural, data-driven groupings across granularities. And PHATE is used for visualizing a series of iterations in this dynamic condensation process.

Given data matrix $X = \{\mathbf{x}_i \in \mathbb{R}^d\}_{i=1,\cdots,N}$, kernel parameter ε ,merge threshold ζ, and gradient parameter ϵ, the Multiscale PHATE algorithm can generate embeddings in T resolutions $J = J_1, J_2, \cdot, J_T$, and can select scales of visualization S. The procedure can be described as follow:

a) Creating an initial diffusion potential U_0 and resolution J_0 using PHATE.

b) For t-th level, where $t \in [0, T]$, compute pairwise Euclidean distance matrix D_t from U_t.

c) Compute affinity matrix K_t from D_{t,ε_t}.

d) Compute a Markov transition matrix P_t (diffusion operator) from row normalized K_t.

e) Update $U_{t+1} \leftarrow P_t U_t$.

f) Merge data points x_i, x_j if $||U_{t+1}(i) - U_{t+1}(j)|| < \zeta$, where $U_{t+1}(i)$ is the i-th row of $U_{t+1}(i)$.

g) Compute pairwise distance matrix D_{t+1} from U_{t+1}.

h) Apply metric MDS to D_{t+1} to get the resolution J_{t+1}.

i) Compute gradient g_{t+1} from U_{t+1}, U_t, and update kernel parameter ε_{t+1}.

j) Repeat step (b) to (g) for all $t \in [0, T]$.

k) For $i \in [1, T-1]$, add i to visualization scale set S if g_i is a local minimum.

4 Results

4.1 Analyzing with t-SNE

The t-SNE method is used from the `sklearn` framework, under `manifold.TSNE()`. For t-SNE it was recommended to first perform PCA before computing t-SNE, reducing the computation time of pairwise distances in the

lower dimension, and preserving some global structure [4]. Hence in the project, from `sklearn` framework, the `decomposition.TruncatedSVD()` method is used as it is implemented for a sparse data structure, as well as it will not require the dataset normalization prior to performing SVD (compared to PCA).

One of the big shortcomings in t-SNE is the loss of global geometry of the data, and only performing well for local clustering. As well, the clusters are often broken apart from the overall structure and shuffled around. The method is also sensitive to parameter choice, often creating false clusters when the perplexity (choice of nearest neighbors) parameter is too low. Only with a range of parameter values that t-SNE's graphical representation can be fully interpreted.

For the AD single-cell data, the following procedure is done:

a) Truncated SVD with `n_components=100` is performed first on the pre-processed AD data, of 70634 samples and 17926 features.

b) t-SNE is performed on the entire data set of decomposed data, $R^{(70634\times100)}$ with parameters; `perplexity = [30,100], learning rate = 'auto', iter = 1000, init = 'random'`. (the same can be achieved by using `init = 'pca'` on a dense matrix)

The resulting figures for the different labeling of the AD dataset is then presented for perplexities of 30 and 100, common values from the literature being used are between 5 to 50.[13] (see Figure 1 below and S12 in the Appendix). For the cell-type subcluster embeddings, the same parameters are used with each subcluster sample size n_c being restricted to 5000 cells (kernels being 5000×5000). (see Figure S13 and S14 in Appendix)

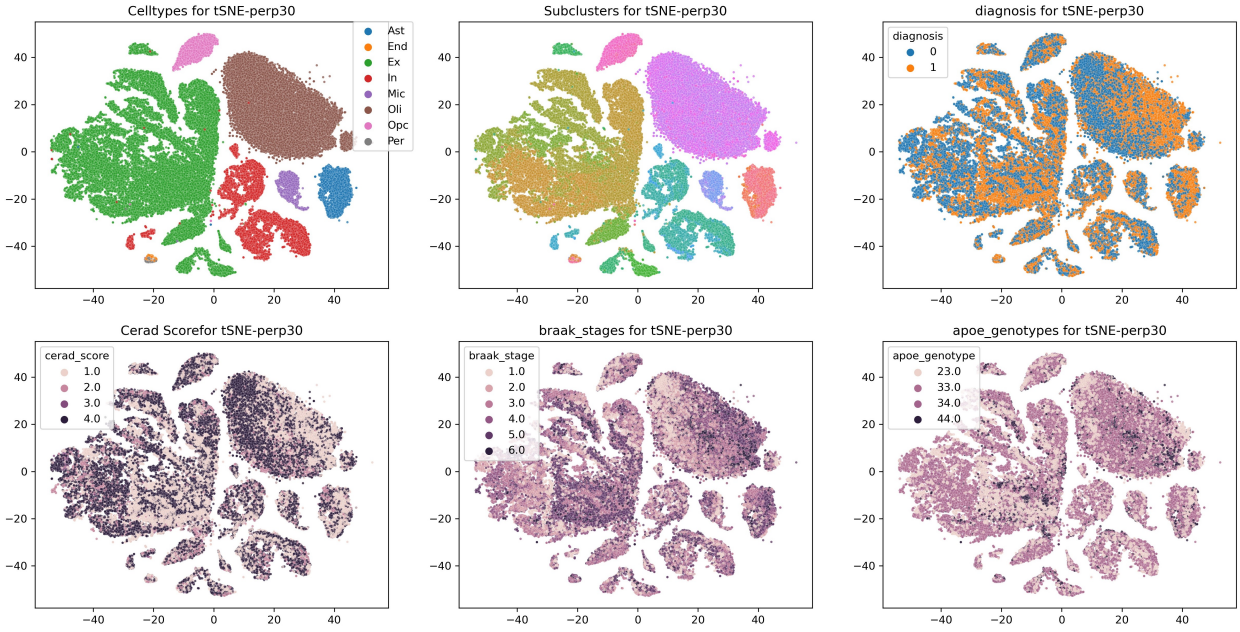


Figure 1: Alzheimer's Disease single cell data representation (cell-types, subclusters, diagnosis, cerad score, braak stages, APOE genotype); t-SNE with perplexity = 30

Interpretation

We know that t-SNE tends to break apart the trajectories in data, or in this context possibly a tree branching structure, where multiple cell types diverge from a common point. For the AD dataset, t-SNE method performs well, as we observe distinct clusters produced for cell types, even though the positional information of each cluster and their apparent size loses meaning. Moreover, the clusters are rotated from figures found for different perplexities.

When performing t-SNE on each cell type subcluster, even more fine-grained structural information is revealed, as the method continue to perform well. The method is able to cluster cell types and even their subclusters without knowing the labels prior, showing that there is an intrinsic geometric difference between such cell families. (Figures S13 and S14).

However, the method tends to break away some parts of each cell type cluster into smaller islands close to the large "landmass", as it is an expected artifact for t-SNE.

For the AD pathology, we consider CERAD score, Braak stages as the levels of AD diagnosis. For majority of cell-type clusters the method does not seem to reveal any separability for AD diagnosis marked cells. The AD-labelled cells seem to be generally evenly distributed among the various cell groups and their sub-clusters. The only notable observation is for the *Oli* cell type, where the cell AD diagnosis forms a gradient in the geometry, showing some separability between the 2 labels. Meaning that *oligodendrocytes* cell type can be potentially used as a good predictor for classification tasks AD diagnosis. The same gradient is visible for CERAD score and Braak stages.

The t-SNE method was shown to be superior to other basic methods; kernel PCA and Isomap. As seen in the later sections of the paper, even resembling to the superior results from multi-scale PHATE method. The most notable aspect was also the use of `TruncatedSVD()` facilitating the computation speed at minimal information loss, when comparing to PHATE method.

4.2 Analyzing with Kernel PCA

The next method for AD single-cell data visualisation is kernel PCA method, using the `sklearn` framework, under `decomposition.KernelPCA()`. Instead if linear PCA, since the `sklearn` package does not implement PCA for sparse matrices, the `truncatedSVD` is used as both a visualization (leaving 2 dimensions, see Figure S15) and pre-processing method for kernel PCA (leaving 100 dimensions).

For the Alzheimer's single-cell data, the kernel PCA procedure is as follows,

a) A random sample of a $n = 17658$ of observations is taken from the AD dataset, as to reduce the size of the kernel to being only 17658×17658, instead of 70634×70634 (most importantly is that our machines do not have 18.6 Gb of RAM required to store a 70634×70634 kernel).

b) The reduced data set is then preprocessed with the `truncatedSVD` to 100 dimensions, as to reduce the kernel computation time.

c) kernel PCA is then used with the radial basis function kernel, i.e with parameters; `n_components=2, kernel = 'rbf', gamma=60/n`. (see Figure 2 in below for the cell type plots labeled by: diagnosis, cerad score, braak stages and APOE genotype)

For the cell-type subcluster embeddings, the same parameters are used with each subcluster sample size n_c being restricted to 5000 cells (kernels being 5000×5000). (see Figures S16 and S17 in Appendix)

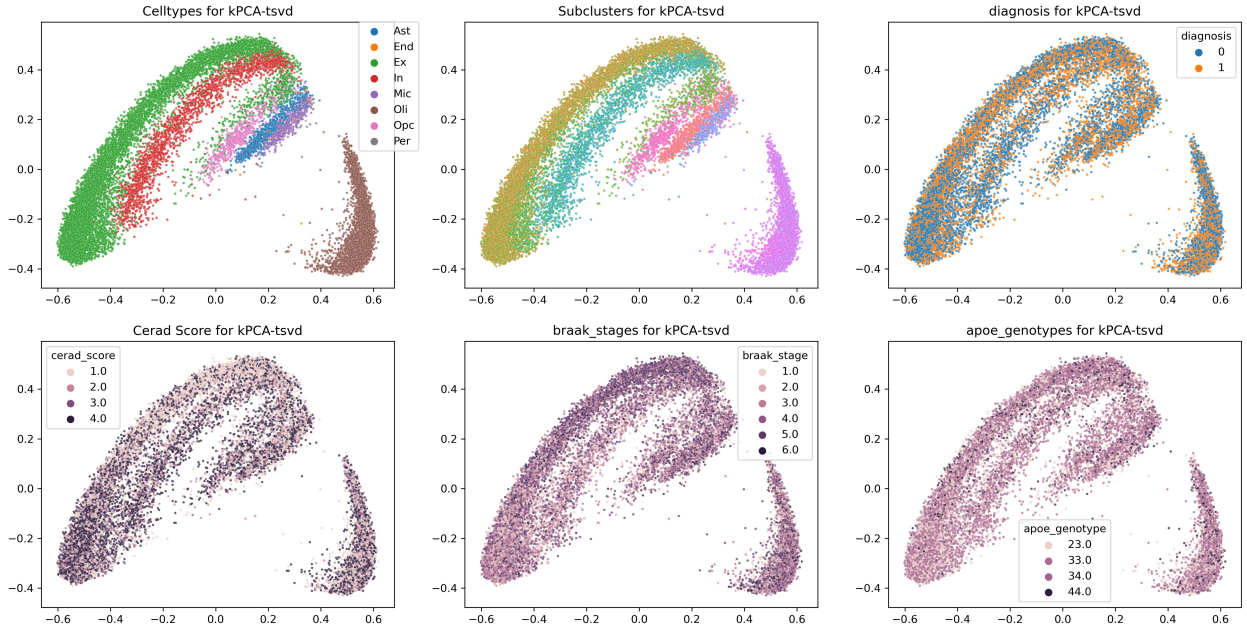


Figure 2: Alzheimer's Disease Single-Cell Data Representation (cell-types, subclusters, diagnosis, cerad score, braak stages, APOE genotype); kernel PCA (RBF kernel, $\gamma = \frac{60}{n}$)

Interpretation

One of the pitfall of PCA (and Isomap) is the inability to capture the fine-grain structure of the data. We are then left only with the general global geometry of the data. The same is visible for the kernel PCA dimensional reduction of AD data. On the bright side, the method's produced clusters are not broken away at times into smaller islands as in t-SNE.

Moreover, The linear T-SVD method shows similar blobs produced by kernel-PCA, for which there is more curvature, but the same separability. This indicates that our original data might not have required to be linearly separated with a kernel to begin with.

While the cell types are now linearly separable and the global positions are preserved, following similar positions (and order) as in t-SNE and PHATE, although the fine grain structure of each supposed cluster is noisy, often mixing in together with other clusters and the information contained within, i.e the labels: diagnosis, cerad score, braak stages and APOE genotype.

For other labels: diagnosis, cerad score, braak stages and apoe genotype, the data remains to be mixed. Compared to healthy subject cells, the AD diagnosis does not revealing a different cell data geometry. i.e the cell degeneration does not seem to produce different clusters for their own AD marked cells. The method is not deemed to be a success, as the clusters are to noisy with each subcluster giving no meaningful insights into the AD pathology.

4.3 Analyzing with Isomap

The Isomap method is also used from the `sklearn` framework, under `manifold.Isomap()`. As for t-SNE, the dataset is dimensionally reduced from it's 17926 to 100 dimensions using `decomposition.TruncatedSVD()`, to reduce noise and fasten the Isomap computation (graph construction and geodesic distances). Moreover, euclidean is used being a better distance metric from our experimentation.

Then, for the AD's single-cell data, the Isomap is applied using the parameters `n_neighbors = 10, metric = 'l2'`, on a subsampled (20%) AD data set (see Figure 3).

For the cell-type subcluster embeddings, the same parameters are used with each subcluster sample size n_c being restricted to 5000 cells (to keep the kernel below max system memory). (For figures, see S18 and S19)

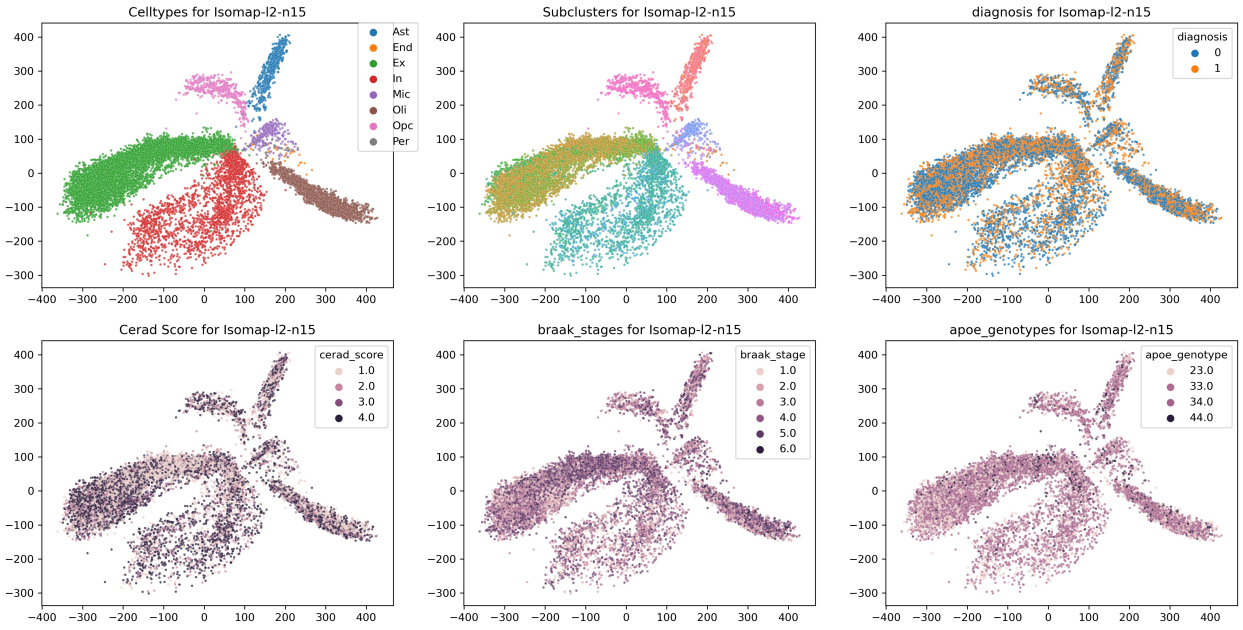


Figure 3: Alzheimer's Disease Single-Cell Data Representation (cell-types, subclusters, diagnosis, cerad score, braak stages, APOE genotype); Isomap with Manhattan distances and n_neighbours = 15

Interpretation

Isomap is generally a considered to be a good method to capture global geometry as well. Having tried with several metrics; cosine, Manhattan distance and Euclidean distance. Only Manhattan distance gave sensible visualizations, with a nearest neighbor parameter of $n = 15$ for the whole AD dataset, and $n = 10$ for local subclusters.

The method is successful in producing the same cell-type clusters in the global dataset as other methods. Although, within the sub-clusters, we are left with noisy images as in kernel PCA, failing to provide some meaningful geometry for AD pathology inference.

4.4 Analysing with UMAP

UMAP is implemented using the `umap` framework. We create the UMAP operator as `umap.UMAP` and then run `fit_transform` to run the operator with different parameters such as:

- `n_neighbors`: This parameter controls the n nearest neighbors. This balances local versus global structure in the data. The lower the value, the more concentrated the local structure. The higher the value, the more focus is on the global structure.
- `min_dist`: The parameter controls how tightly the graph is allowed to pack points together. It provides the minimum distance apart that points are allowed to be in the low-dimensional representation.
- `n_components`: This parameter describes the low dimension in which the data visualize itself as.
- metric: This parameter controls the metric by which the distance is calculated by.

The AD's single cell data is visualized with UMAP with different `n_neighbors` with the values 10, 50, and 100 seen in S20, 4, S21.

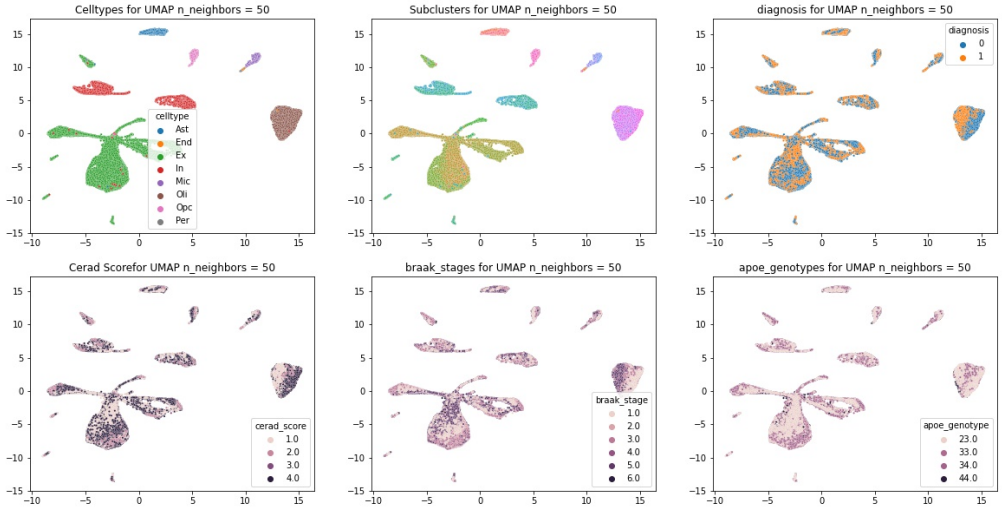


Figure 4: Alzheimer's Disease single-cell data representation; UMAP with 50 nearest neighbours

We are also able to visualize this single-cell data with respect to the Manhattan metric in S23. We can also see the connection made in the low dimensional representation graph of AD's single cell data when UMAP with `n_neighbors=100` is applied in S22

The AD's single-cell data is analyzed with Supervised UMAP. This is done by passing the diagnosis label in `fit_transform`, doing this we get a visualization of the data with a clustering with respect to their class label which in this case is the diagnosis seen in 5, we can also see clustering in the subcluster in S24. We also supervise UMAP clustering on the basis of sex in 6.

Interpretation

UMAP is known to have a balance between it's local structure and global structure. The balance is maintained by having an appropriate value for the `n_neighbors` parameter. The visualization with different values such as 10, 50, and 100 nearest neighbors can be seen. If `n_neighbors` value is too low such as 10, we see how the local structure is prioritized more with no proper clustering in AD's single cell type data. For example, the 'Ex' cell type (green cluster) is not fully merged and forms tiny local clusters instead. While for a higher value such as 100 we see a bit stray from

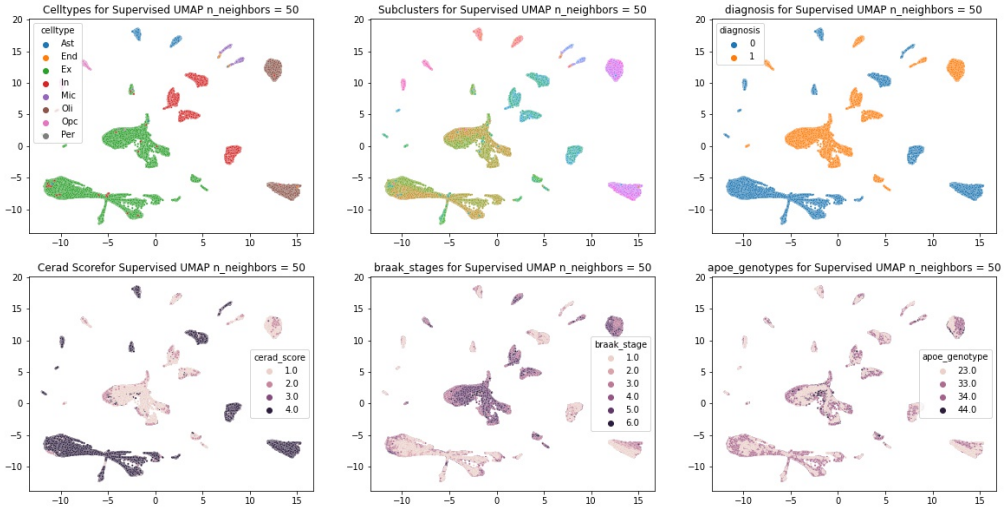


Figure 5: Alzheimer's Disease single-cell data representation; Supervised UMAP with 50 nearest neighbors on the basis of Diagnosis

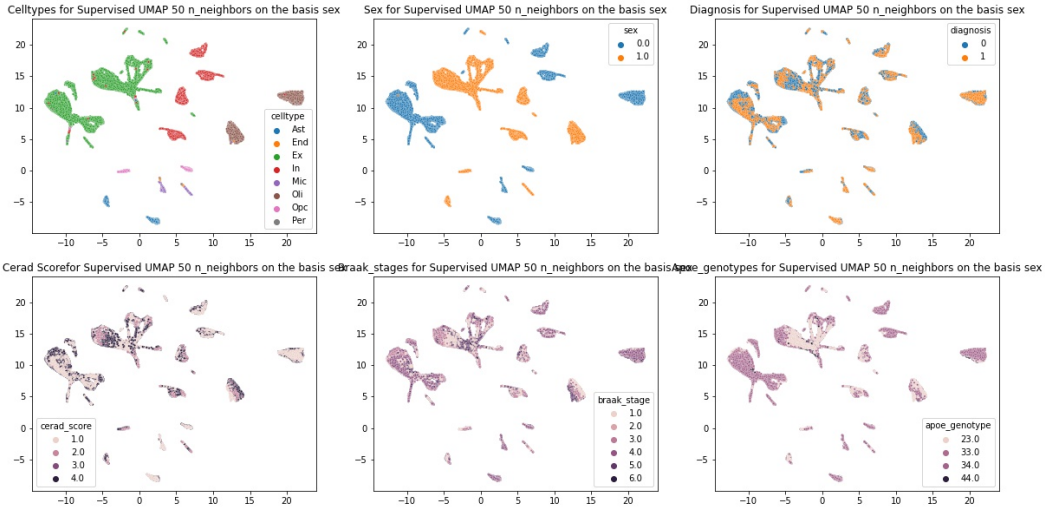


Figure 6: Alzheimer's Disease single-cell data representation; Supervised UMAP with 50 nearest neighbors on the basis of Sex

the local structure with the global structure given more priority. For example- 'In' cell types sees a break from the local structure with data points straying from the cluster. UMAP with `n_neighbors=50` sees a more balanced approach. with clusters being well formed while having a global structure too.
We also are able to analyze the cell types based on the diagnosis class by applying supervised dimension reduction in UMAP. We are able to visualize well-clustered data based on class and diagnosis and can analyze the graphical representation and cluster of data based on `diagnosis = {0,1}` and on the basis of sex. We see that in the clustering based on diagnosis, females are more likely to be diagnosed with AD. We also see that there is a higher number of female diagnosed patients in the Ex cell type. In this cluster, we observe lower CREAD scores and a higher Braak stage

in comparison to patients that do not have AD. This cluster also visualizes the Apoe genotype of value 44.0, which is not seen much in other clusters.

4.5 Analysing with PHATE

The PHATE method is used from the `phate` framework. We create the PHATE operator as `phate.PHATE` and then run `fit_transform` to run the operator with different parameters such as:

- `knn`: To specify the number of nearest neighbors.
- `decay`: This parameter specifies the alpha decay for the kernel.
- `t`: This parameter provides the number of times to power the operator. Equivalently it is the amount of smoothing done to the data.

For the AD's single-cell data, PHATE is applied using varying parameters like `knn=5` and `knn=10`can be seen in 7. We experiment by varying the smoothness by `t=30` and `knn=4` and the resulting figure is S25

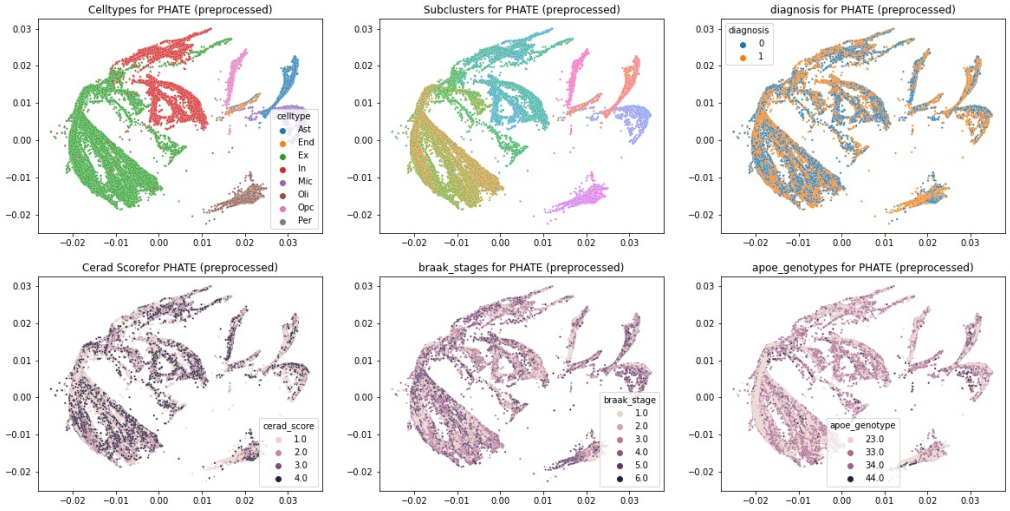


Figure 7: Alzheimer's Disease single cell data representation; PHATE for KNN=5

Interpretation

PHATE on AD single-cell data gives us a more interconnected visualization of data in contrast to UAMP or t-SNE. With classes like different cell types interacting with one another, we get a more general global structure while still having the local structure maintained. We are able to infer how the cell-type clusters and sub-cluster interact with each other. While increasing the smoothing and decreasing the nearest neighbor, the local structure stretches a bit further but still holds. We observe a higher diagnosis of AD in Ex and In cell types.

4.6 Analysing with Multiscale PHATE

Multiscale PHATE produces multigranular visualizations and clusters of high-dimensional biological data. Multiscale PHATE visualization of Alzheimer's Disease single-cell data identifies all major cell types, as shown in Figure 9. The Figure 9a shows the gradient of each iteration and marks the local minimum, which are used to generate multi-level of resolutions. From Figure 9b to 9f show some chosen different resolutions from finest to coarsest. with the iteration going, some closed points are merged into a single point. Since endothelial cells (End), and pericyte cells (Per) have low cell counts, the points in End and Per celltypes are merged into other clusters, therefore, there are only six cluster shown in the visualization in coarse resolution, as shown in Figure 9e.

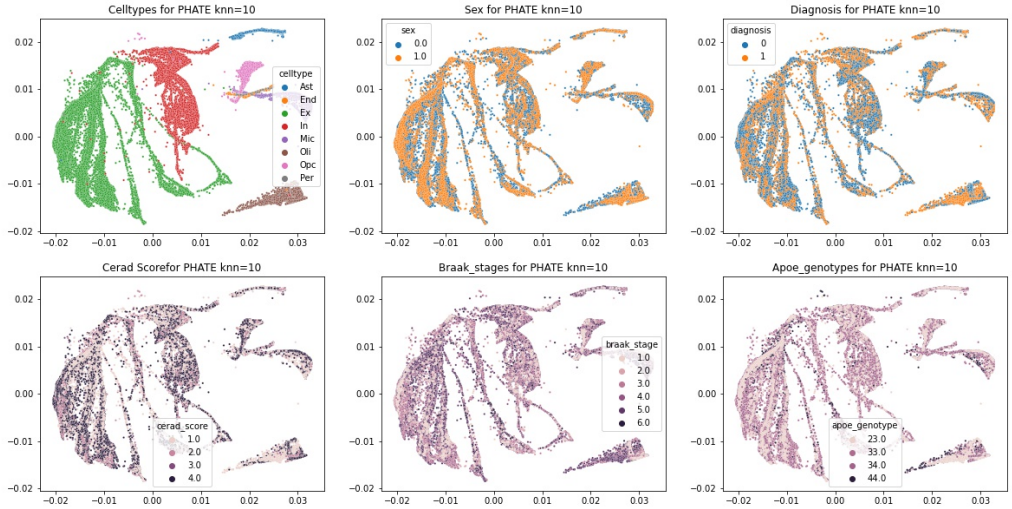
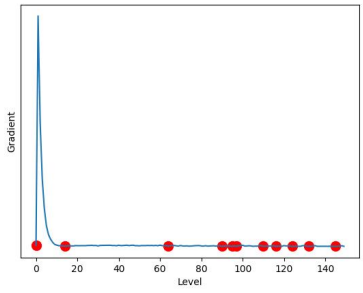
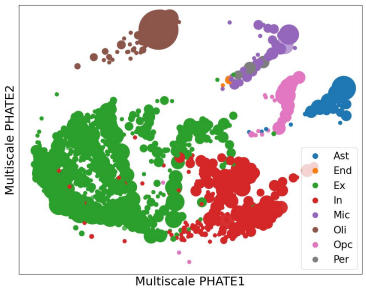


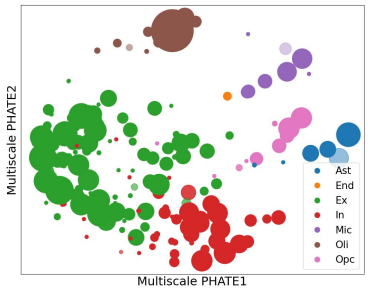
Figure 8: Alzheimer's Disease single cell data representation; PHATE for KNN=100



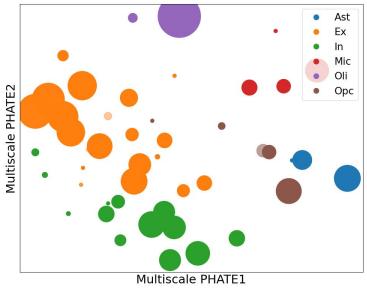
(a) Gradient analysis identifies a range of levels for visualization.



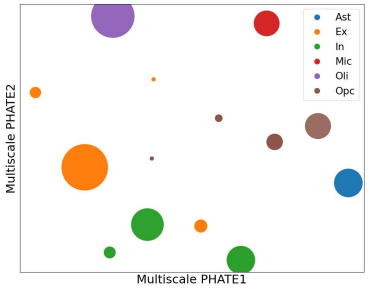
(b) Finest resolution



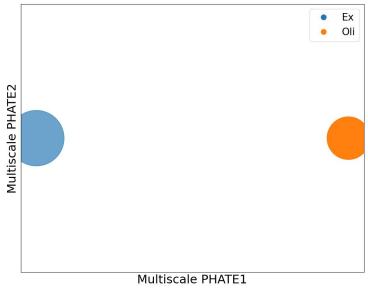
(c) Fine resolution (level = 64)



(d) Medium resolution (level = 97)



(e) Coarse resolution (level = 116)



(f) Coarsest resolution

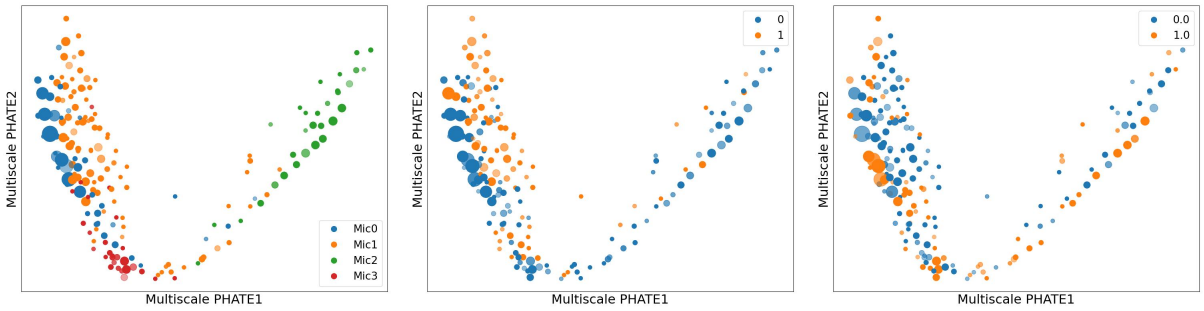
Figure 9: Visualizations with multiscale PHATE in different resolutions.

In order to find some correlation between features and diagnosis, we choose a medium resolution, which keeps both the global structure and local details, to visualize the embeddings by labeling in six features, including celltype, diagnosis, sex, CERAD score, Braak stage, and APOE genotype, as shown in Supplementary Figure S26. For each merged point, the label was assigned by majority vote. From the figure S26b and S26d, we can see that the point with a diagnosis of 1 has a CERAD score of 1, which meets the definition of the CERAD score. From the figure S26f, we can found that

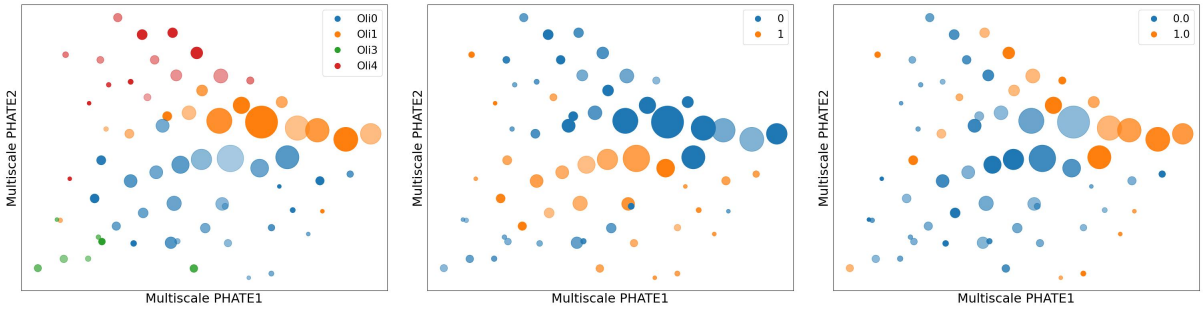
most of the APOE genotypes are e33. However, from these figures, we couldn't find any potential correlation between diagnosis and other features from these figures, so we need more analysis.

Single-cell RNA-sequencing technologies suffer from many sources of technical noise, including under-sampling of mRNA molecules, often termed 'dropout', which can severely obscure important gene-gene relationships. Considering the AD single-cell data might also suffers from this kind of "droupout", we applied MAGIC (Markov Affinity-based Graph Imputation of Cells) [14], a method that shares information across similar cells, via data diffusion, to denoise the cell count matrix and fill in missing transcripts. After applying MAGIC, we visualized the imputed data with Multiscale PHATE, which is shown in Supplementary Figure S27. Compared with the visualization of original data in Figure 9, the visualization of MAGIC imputation couldn't cluster the cell type well, which means it is not a suitable method for AD single-cell data.

Then, we analyzed each celltype separately. We visualized with Multiscale PHATE on each cell type in several levels and choose the clearest level to analyze the relationship among sub-cluster, diagnosis and sex, as shown in Figure 10 and Supplementary Figure S28. We didn't analyze for CERAD score, Braak stage in this case because they are metrics that were defined as associated to diagnosis of AD. But the impact of sex and celltype on the diagnosis of AD is waiting for us to discover. For Microglia celltype (Mic), comparing the three figures in Figure 10a, the cells belong to Mic1 sub-cluster are mostly being diagnosis as AD, and their gender are mostly female, which means Mic1 is one of the AD-pathology-associated cell subpopulations. In Figure 10b, the cells that belong to sub-cluster Oli1 are almost non-diagnosis, which means Oli1 is one of the no-pathology subpopulations.



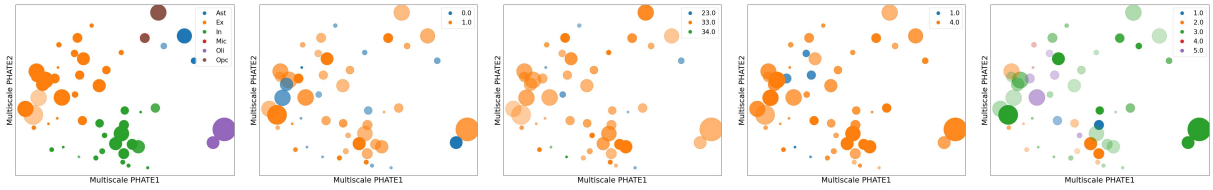
(a) Mic (level = 53)



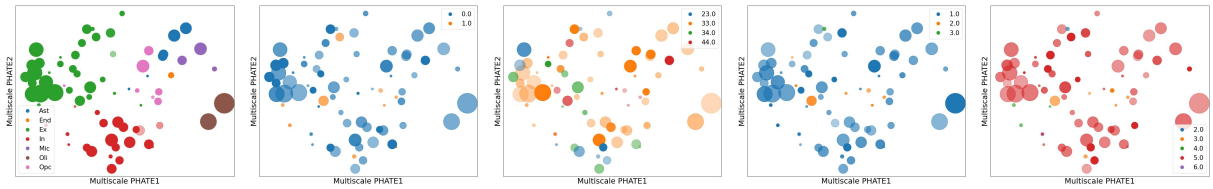
(b) Oli (level = 86)

Figure 10: Visualizations of Microglia and Oligodendrocyte progenitor celltype in best resolution and labeling in sub-cluster, diagnosis and sex. In each sub-figure, the legend is: LEFT: sub-clusters, MIDDLE: Diagnosis, RIGHT: Sex

For further analysis, we focus on visualizing for the data of each gender, as shown in Supplementary Figure S29, and diagnostic result, as shown in Figure 11. We can found from Figure 11 that most of the diagnosis cells are from female patients, and the same pattern shown in Supplementary Figure S29a. For non-diagnosis cells, most of the APOE genotype are e33, but for diagnosis cells, there are several types mixed. For CERAD score, most of diagnosis cells have a value of 1 and most of non-diagnosis cells have a value of 4. For Braak stages, most of diagnosis cells have a value of 5 and most of non-diagnosis cells have a value of 1-3.



(a) Non-diagnosis



(b) Diagnosis

Figure 11: Visualizations with multiscale PHATE for non-diagnosis and diagnosis in level 77 and labeling in celltype, sex, APOE genotype, CERAD score and Braak stages (left to right).

5 Conclusion

In this project, we visualized the dataset of single-nucleus transcriptomes of Alzheimer's disease from the prefrontal cortex of 48 individuals by six different methods.

Using initial Truncated TSVD methods, kernel PCA and Isomaps we have shown that cell-types have a distinct glabal geometry that is separable using even simpler methods, although producing noisy local structure. Then with t-SNE, we have been able to reveal local structures, separating the cell types and then the sub-clusters of each cell type. The fine-grain structure then reveals the geometrical structure of the AD pathology, with CERAD score and Braak stages showing correlations between AD diagnosis labels. As well as being separable for some cell-types (Oli), allowing further classification methods to applied for AD diagnosis inference from the the local structure.

For visualization of single-cell transcriptomes using UMAP and PHATE, our visualizations gave us valuable input about the local and global structure. PHATE seems to have a greater preservation of global structure and UMAP was very flexible to give us more local structures and clustering on the basis of the provided labels like sex and diagnosis. We were able to infer valuable details by being able to cluster our data with respect to cell type and diagnosis, by showing a higher chance of AD in females. We were also able to see the effect it had with respect to CERAD score and Braak stages, AD affects patients with lower CREAD scores and higher Braak stages.

For Multiscale PHATE, we visualized the whole dataset in several levels and labeled them by majority vote in different features. Then, we further analyzed the relationship of several clinical features by visualizing in subsets of data. From the visualization, we can easily found that the cells with AD diagnosis have a low CERAD score and high Braak stages. Some subclusters of the celltype may have pathology-associated property, we could find from the visualisations that Mic1 is one of the AD-pathology-associated cell subpopulations and Oli1 is one of the no-pathology subpopulations. Besides, although the original dataset is sampling from 48 patients, including 24 of both sexes with 12 of each is AD-pathology and 12 of each is no-pathology, the visualization results shows that the number of diagnosis females is more than diagnosis males, which is gender biased. This might cause by the preprocessing procedure that we deleted more cells from male than female.

By implementing the various visualization methods we were able to infer similar results mentioned by Hansruedi et al [6]. In future, we are going to explore more specific relationship among these genetic and clinical features, to figure out more AD-pathology-associated cell populations.

6 Collaboration and individual contributions

This project is collaborated by Shuang, Nishka, Timur as a project of MAT 6493. In collaborating on this project, each of us has a clear division of contributions.

Shuang was the initiator and leader of this project. She was responsible for providing project ideas, planning, explanation, related papers, downloaded data, and writing framework. Besides, she performed data preprocessing, and wrote the

section of data sources, data preprocessing and explanation. Additionally, she performed the visualization with Multiscale PHATE, tried MAGIC imputation on the dataset, visualized multiple subsets of data with labeling different features, and wrote the corresponding methodology part and results analysis.

Nishka was responsible for the implementation of UMAP and PHATE. She was also involved in writing the motivations and methodologies, inference, figures, etc of the UMAP and PHATE.

Timur was responsible for the sections involving kernel PCA, t-SNE and Isomap analysis, written sections, and figures.

References

[1] David A Bennett, Aron S Buchman, Patricia A Boyle, Lisa L Barnes, Robert S Wilson, and Julie A Schneider. Religious orders study and rush memory and aging project. *Journal of Alzheimer's disease*, 64(s1):S161–S189, 2018.

[2] Heiko Braak and Eva Braak. Neuropathological stageing of alzheimer-related changes. *Acta neuropathologica*, 82(4):239–259, 1991.

[3] Heiko Braak and EVA Braak. Staging of alzheimer's disease-related neurofibrillary changes. *Neurobiology of aging*, 16(3):271–278, 1995.

[4] Kobak Dmitry and Berens Philipp. The art of using t-sne for single-cell transcriptomics. *Nature Communications*, 2019.

[5] Manik Kuchroo, Jessie Huang, Patrick Wong, Jean-Christophe Grenier, Dennis Shung, Alexander Tong, Carolina Lucas, Jon Klein, Daniel B Burkhardt, Scott Gigante, et al. Multiscale phate identifies multimodal signatures of covid-19. *Nature Biotechnology*, 40(5):681–691, 2022.

[6] Hansruedi Mathys, Jose Davila-Velderrain, Zhuyu Peng, Fan Gao, Shahin Mohammadi, Jennie Z Young, Madhvi Menon, Liang He, Fatema Abdurrob, Xueqiao Jiang, et al. Single-cell transcriptomic analysis of alzheimer's disease. *Nature*, 570(7761):332–337, 2019.

[7] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[8] Kevin R Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel B Burkhardt, William S Chen, Kristina Yim, Antonia van den Elzen, Matthew J Hirn, Ronald R Coifman, et al. Visualizing structure and transitions in high-dimensional biological data. *Nature biotechnology*, 37(12):1482–1492, 2019.

[9] John C Morris, Albert Heyman, Richard C Mohs, JP Hughes, Gerald van Belle, GDME Fillenbaum, ED Mellits, and C Clark. The consortium to establish a registry for alzheimer's disease (cerad): I. clinical and neuropsychological assessment of alzheimer's disease. *Neurology*, 1989.

[10] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International conference on artificial neural networks*, pages 583–588. Springer, 1997.

[11] Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[12] Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. *The Journal of Machine Learning Research*, 15(1):3221–3245, 2014.

[13] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[14] David Van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J Carr, Cassandra Burdziak, Kevin R Moon, Christine L Chaffer, Diwakar Pattabiraman, et al. Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3):716–729, 2018.

[15] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):1–5, 2018.

A Supplementary Figures

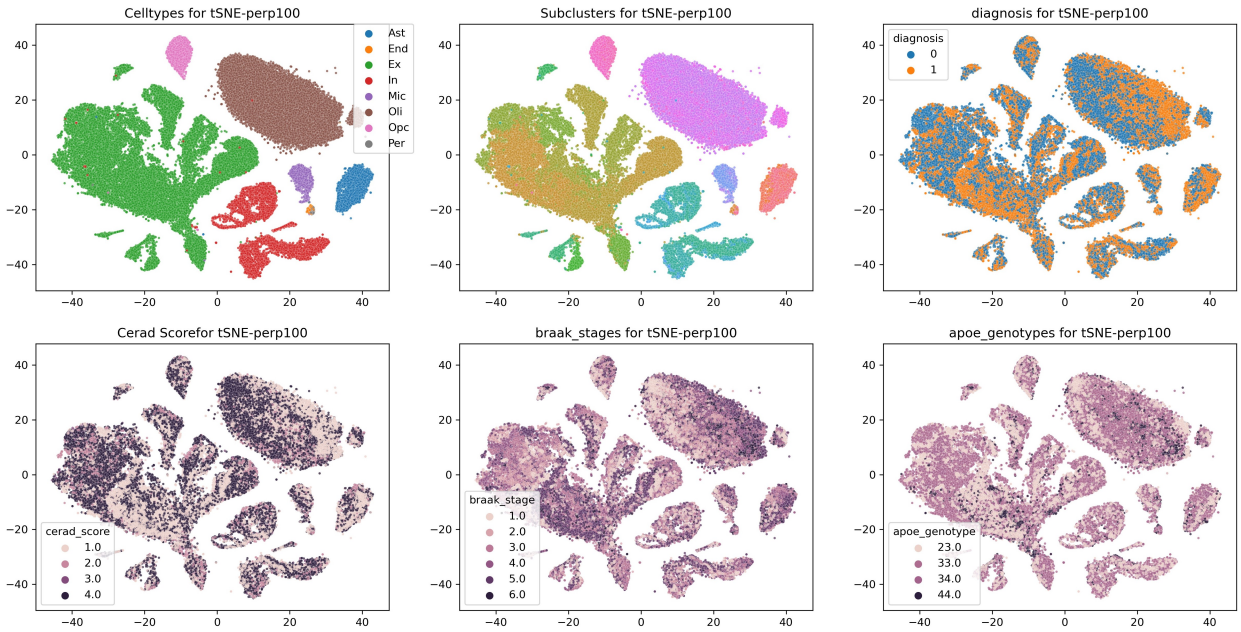


Figure S12: Alzheimer's Disease Single-Cell Data Representation (cell-types, subclusters, diagnosis, cerad score, braak stages, APOE genotype); t-SNE with perplexity = 100

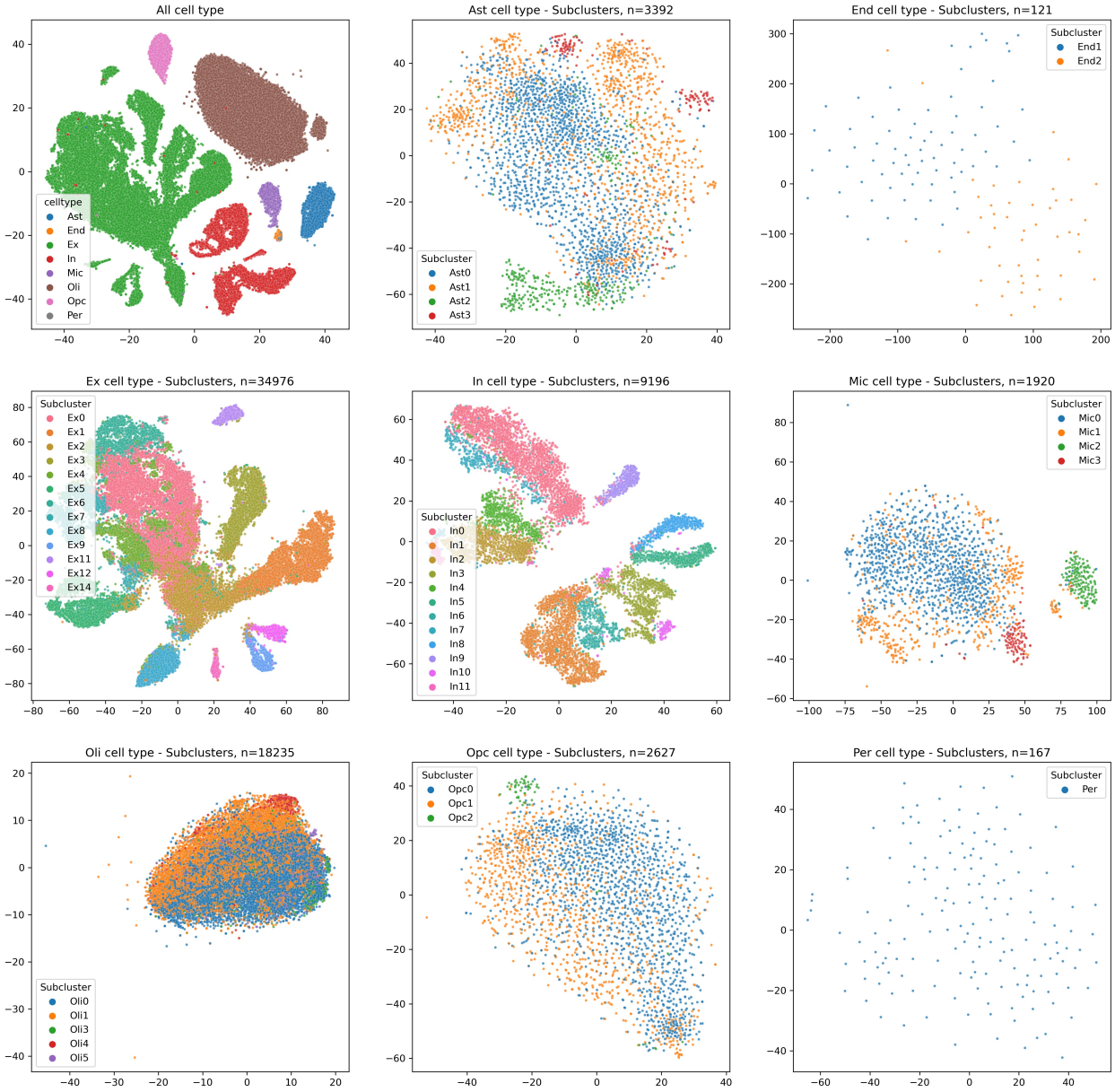


Figure S13: Alzheimer's Disease Subcluster Single-Cell Data Representation; t-SNE with perplexity = $n_c/100$ (and restricted to $[5, 100]$, n_c is the number of cells per cell-type)

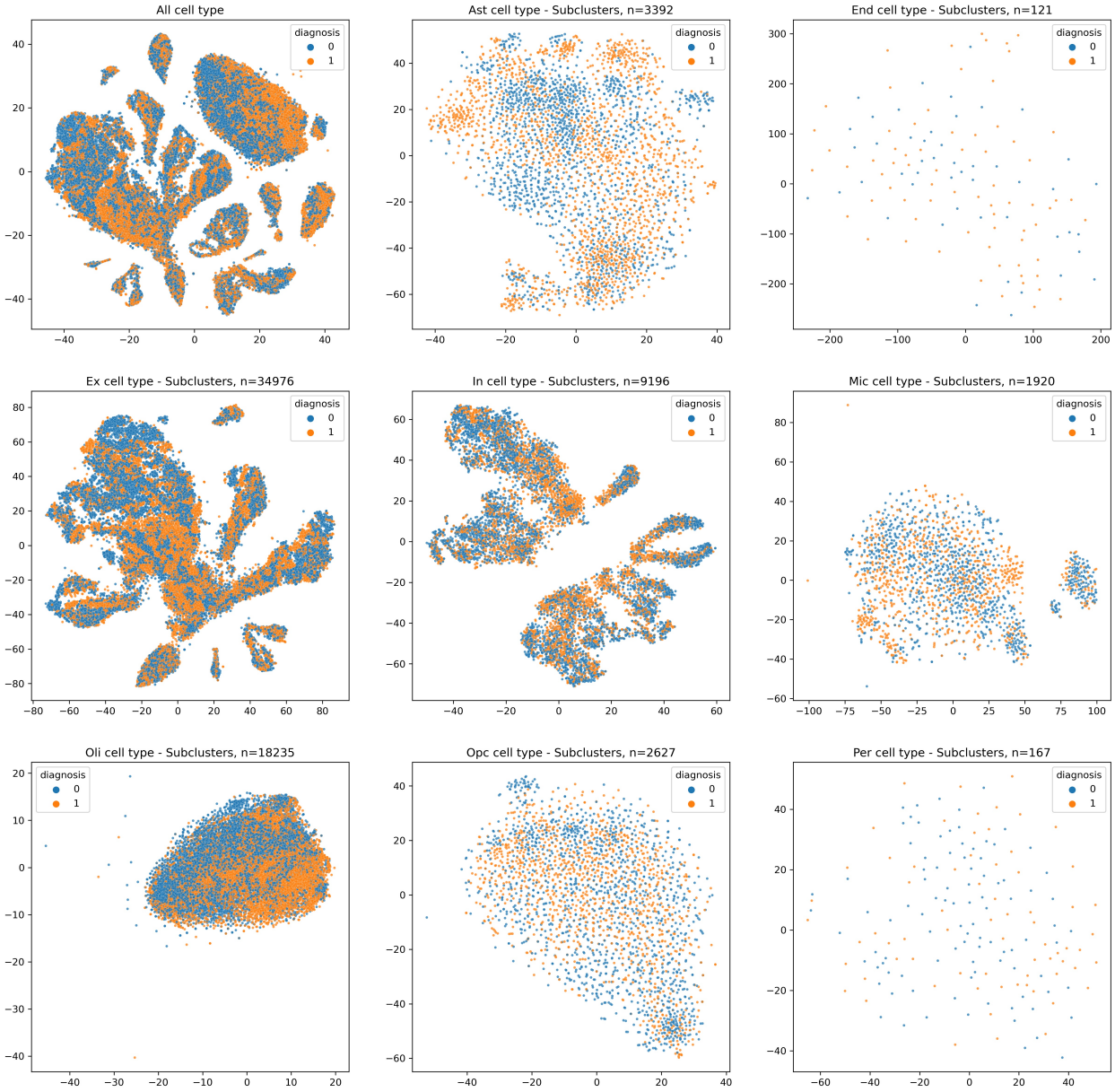


Figure S14: Alzheimer's Disease Subcluster Single-Cell Data Representation (AD diagnosis labels); t-SNE with perplexity = $n_c/100$ (and restricted to $[5, 100]$, n_c is the number of cells per cell-type)

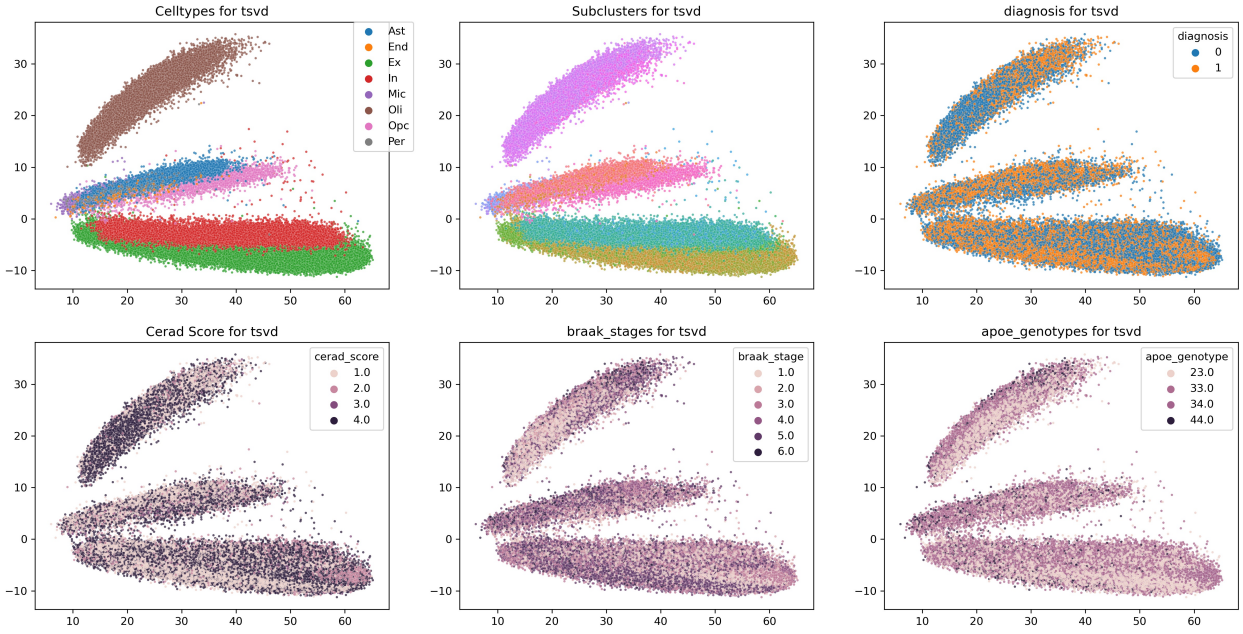


Figure S15: Alzheimer's Disease Subcluster Single-Cell Data Representation (cell-types, subclusters, diagnosis, cerad score, braak stages, APOE genotype); Truncated SVD, $d = 2$

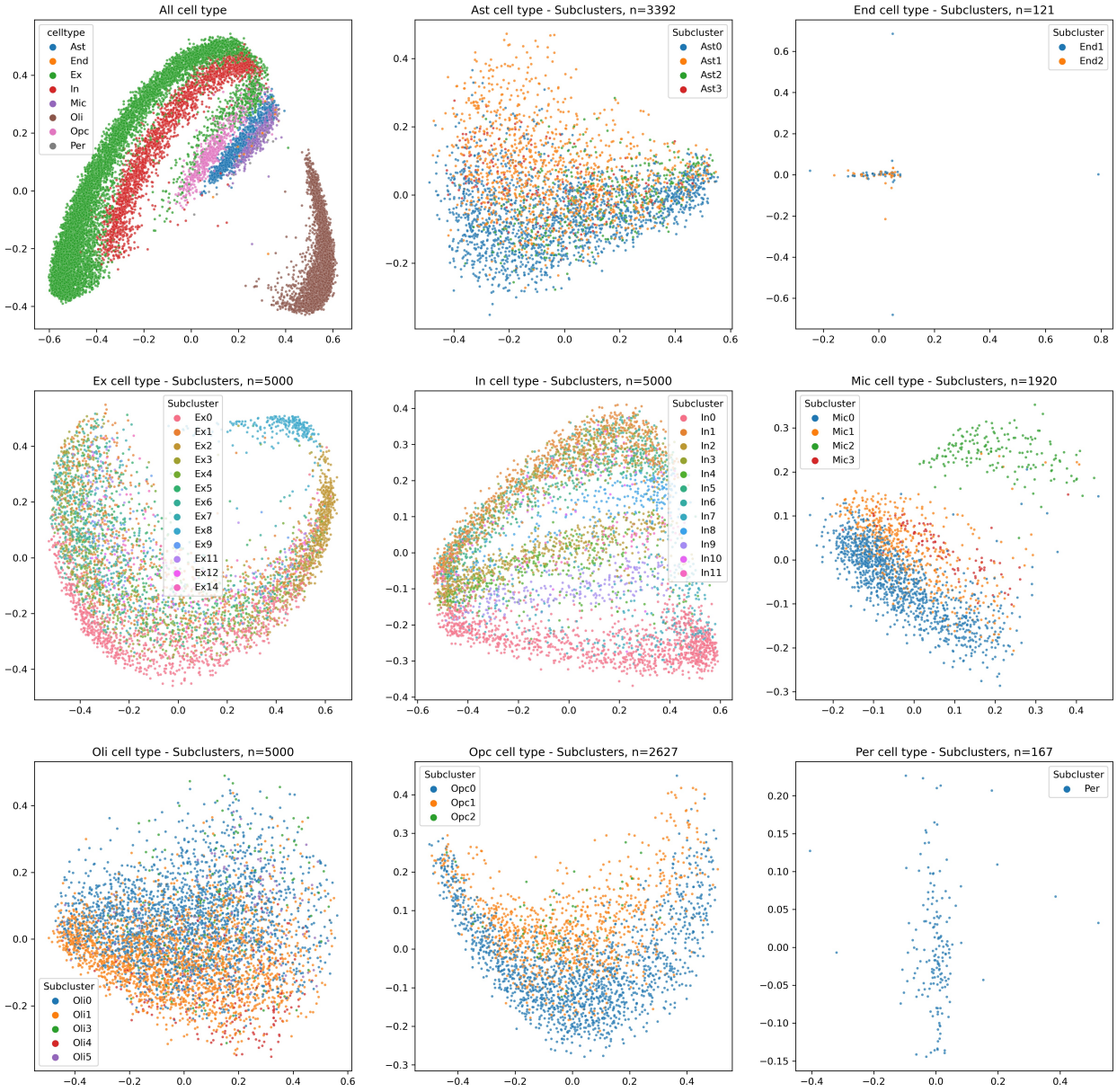


Figure S16: Alzheimer's Disease Subcluster Single-Cell Data Representation; RBF kernel PCA, $\gamma = \frac{1}{n_c}$

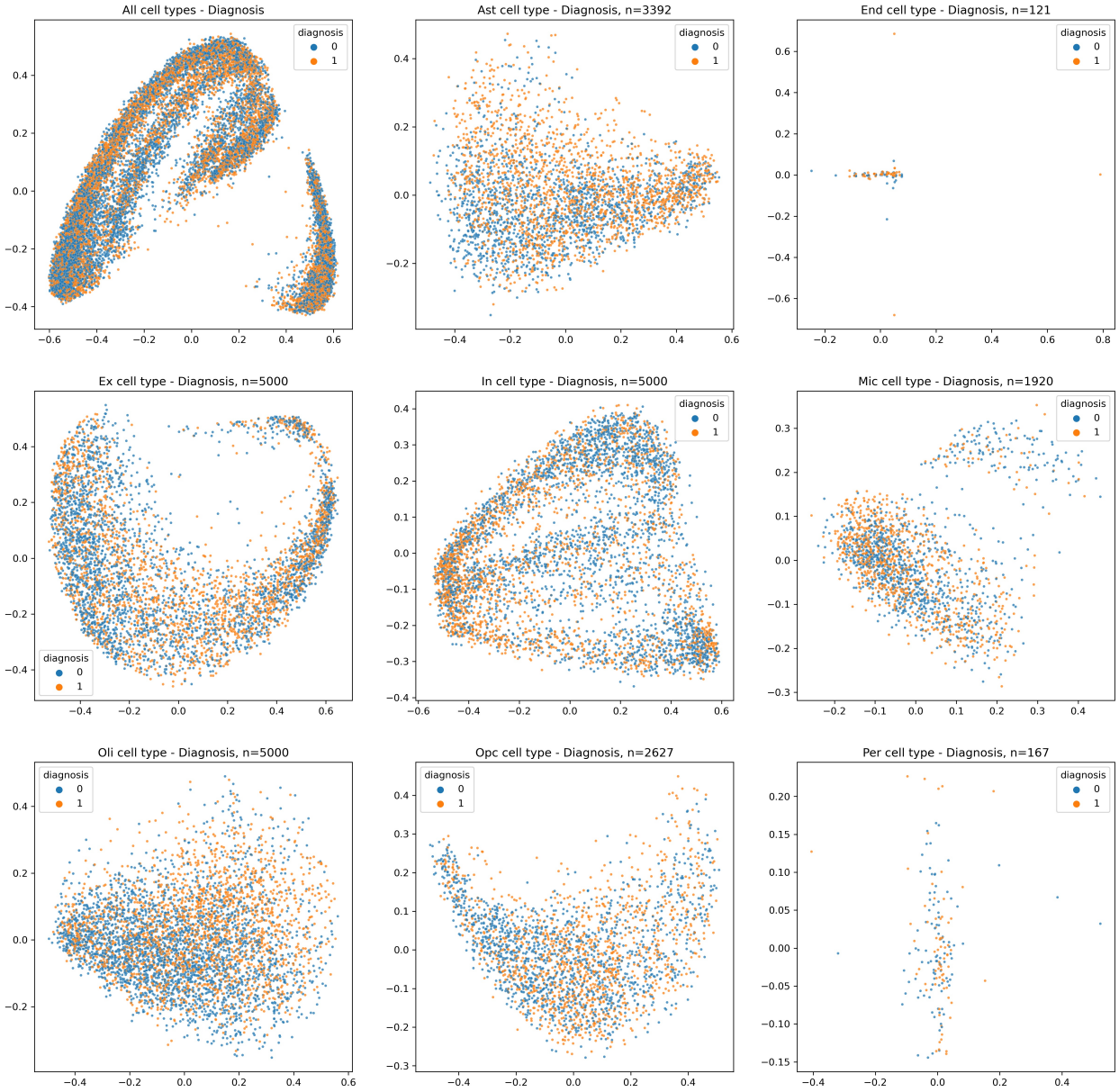


Figure S17: Alzheimer's Disease Subcluster Single-Cell Data Representation (AD diagnosis labels); RBF kernel PCA, $\gamma = \frac{1}{n_c}$

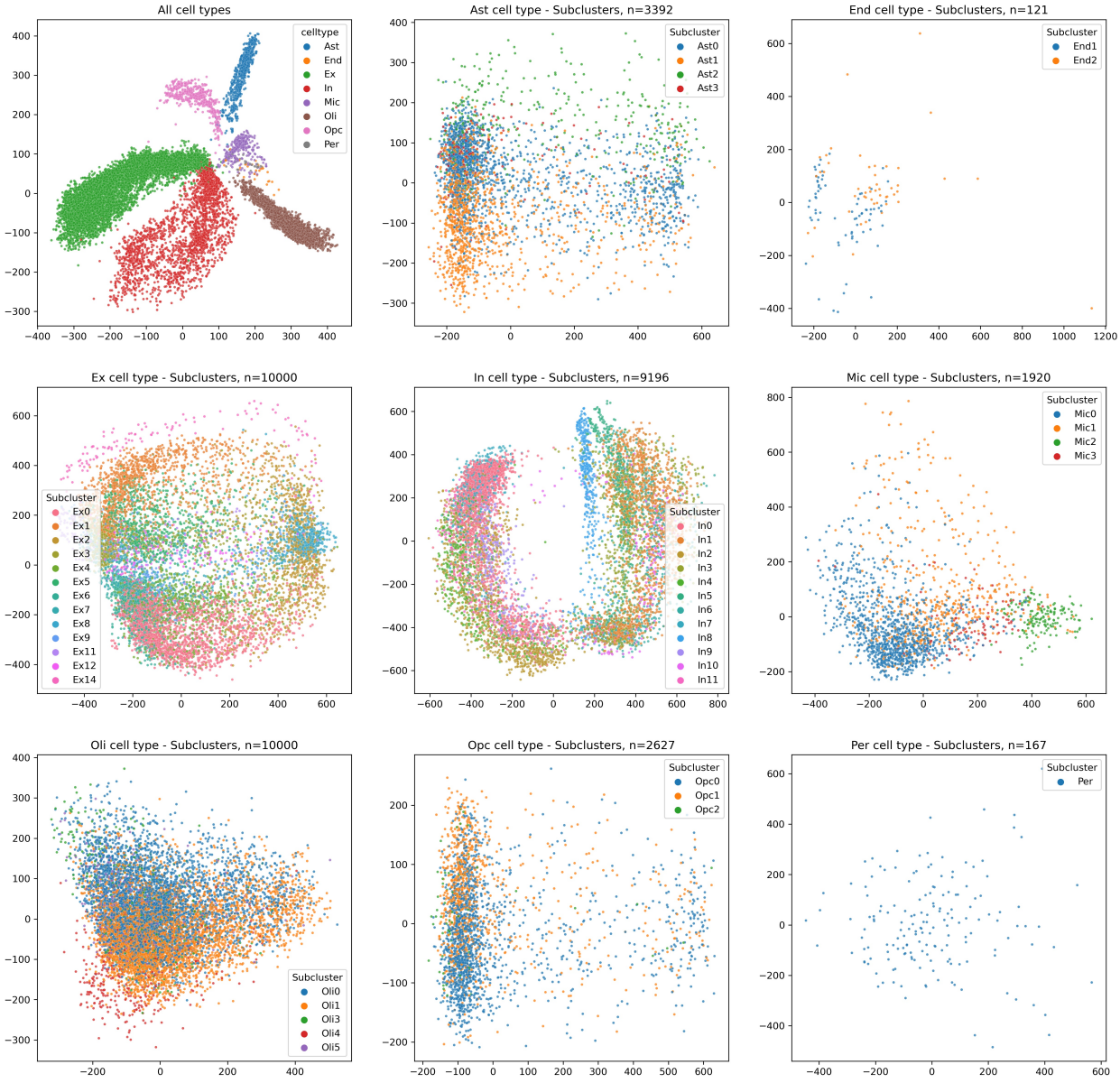


Figure S18: Alzheimer's Disease Subcluster Single-Cell Data Representation; Isomap with Manhattan distances and n_neighbours = 10 (for subclusters)

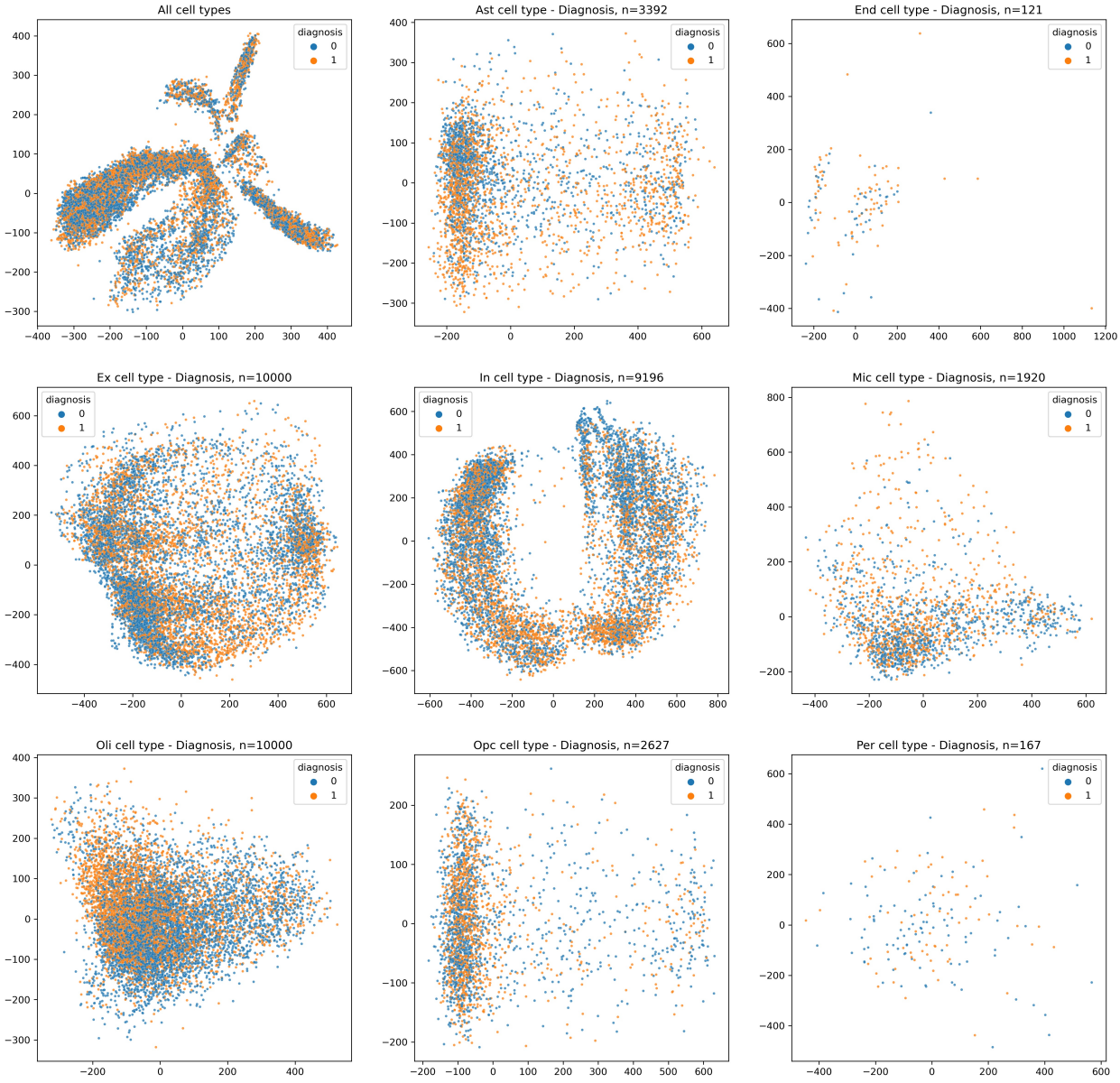


Figure S19: Alzheimer's Disease Subcluster Single-Cell Data Representation (AD diagnosis labels); Isomap with Manhattan distances and n_neighbours = 10 (for subclusters)

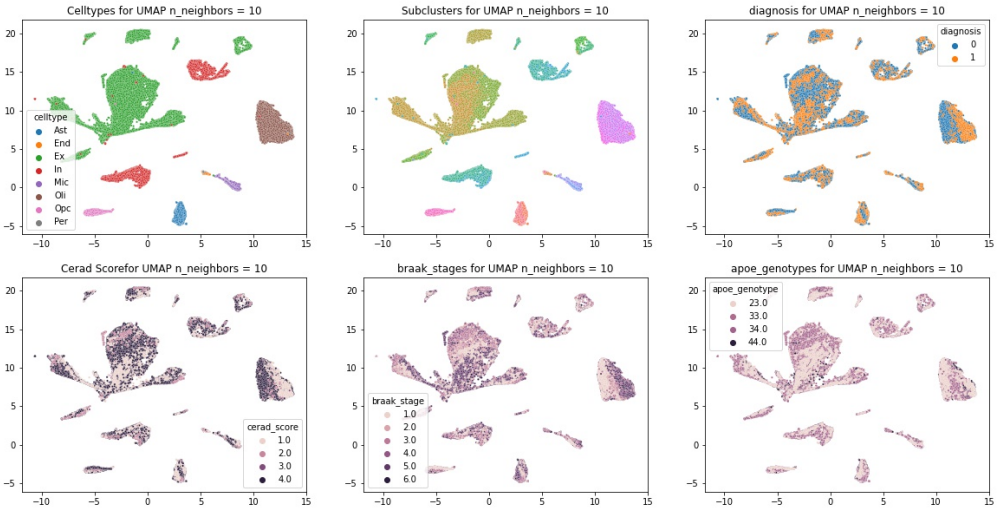


Figure S20: Alzheimer's Disease single cell data representation; UMAP with 10 nearest neighbours

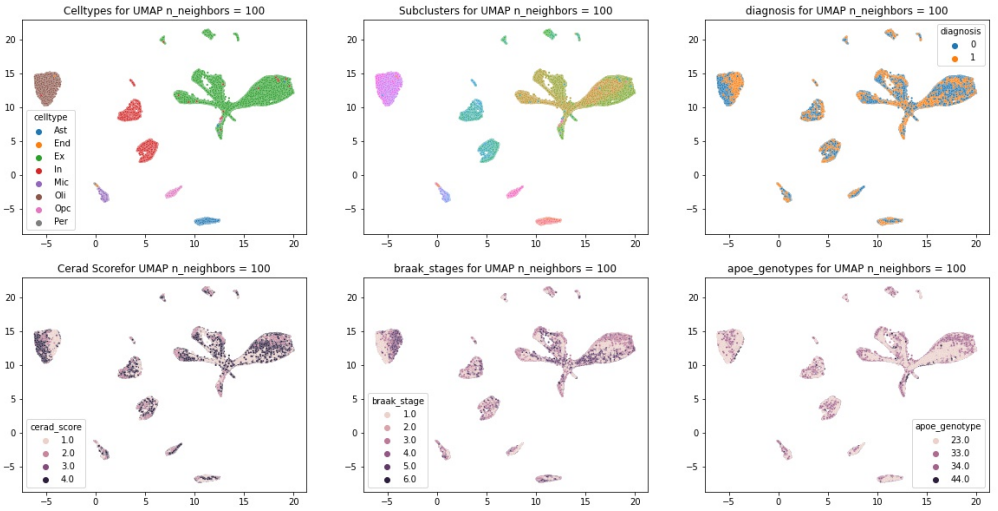


Figure S21: Alzheimer's Disease single-cell data representation; UMAP with 100 nearest neighbours

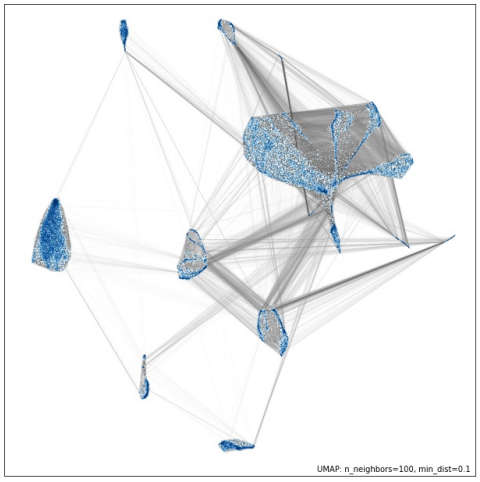


Figure S22: Alzheimer's Disease single cell data representation; UMAP with 100 nearest neighbor connection graph

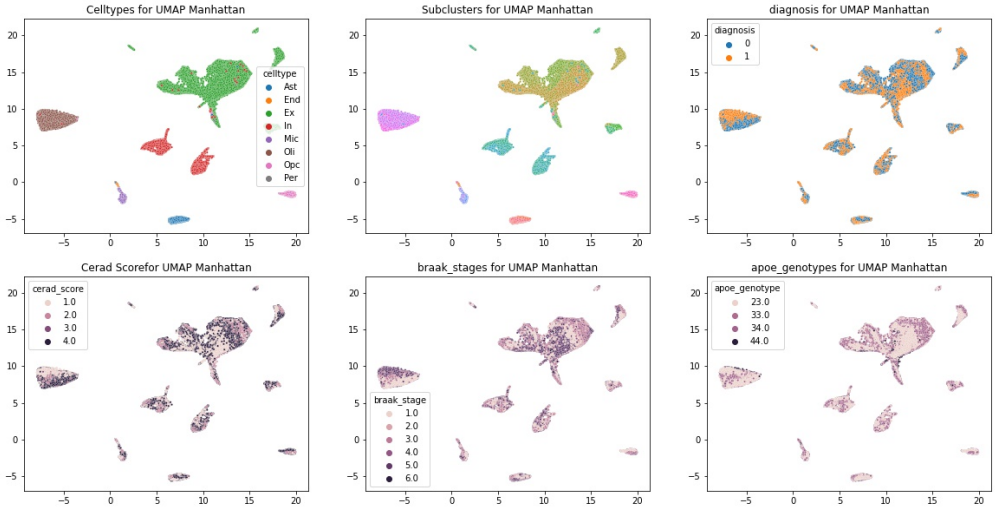


Figure S23: Alzheimer's Disease single-cell data representation; UMAP with Manhattan Distance and 50 nearest neighbours

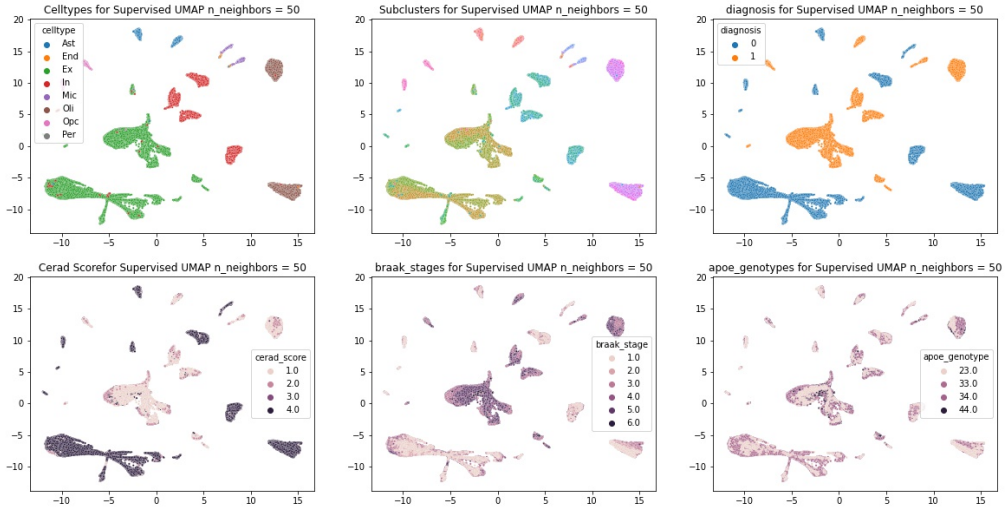


Figure S24: Alzheimer's Disease single-cell data representation; Supervised UMAP with 50 nearest neighbors on the basis of Diagnosis

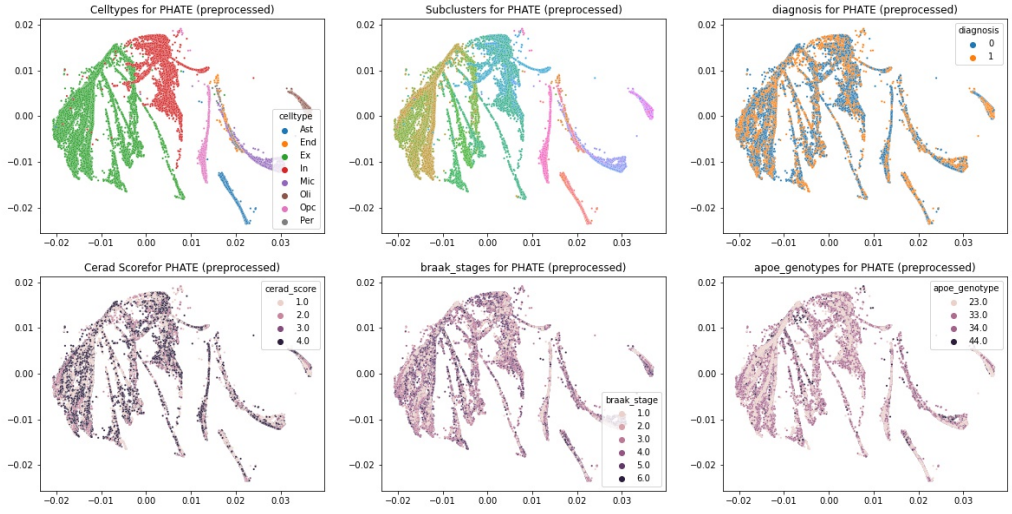
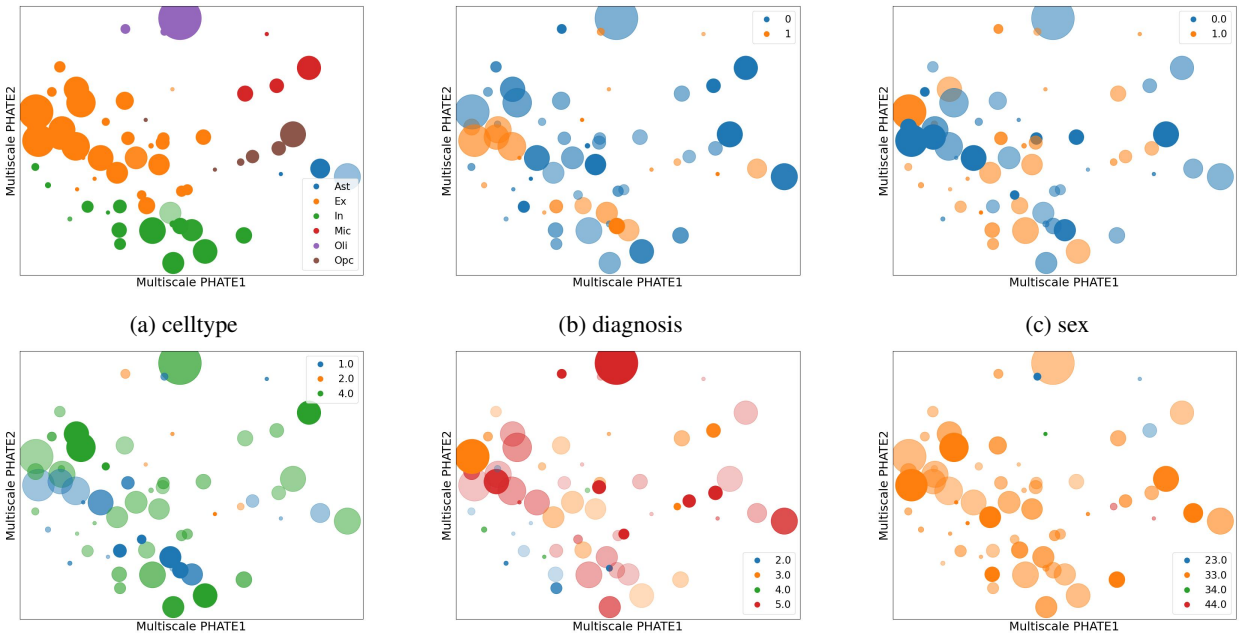


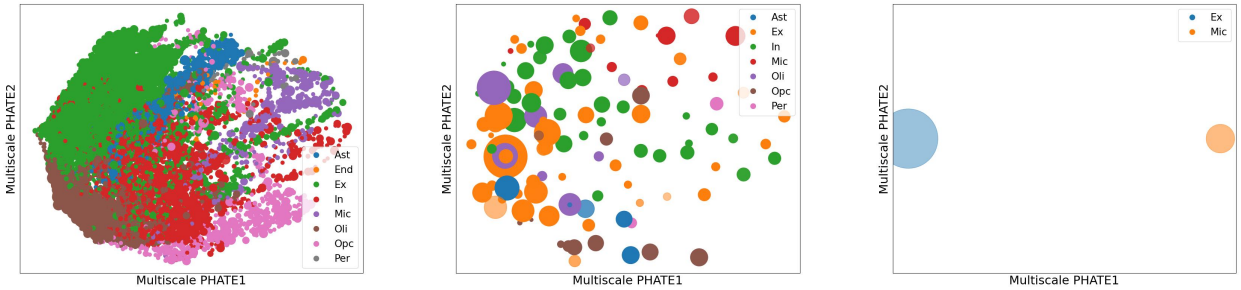
Figure S25: Alzheimer's Disease single cell data representation; PHATE for KNN=4 and t =30



(a) celltype (b) diagnosis (c) sex

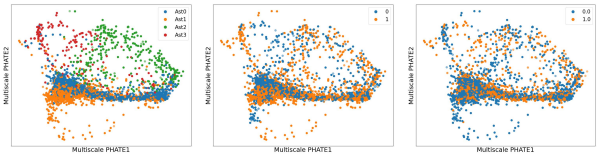
(d) cerad score (e) braak stage (f) APOE genotype

Figure S26: Visualizations with multiscale PHATE labeling in different features (level 90).

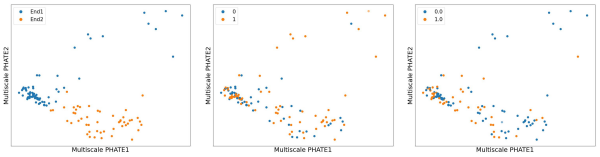


(a) Finest resolution (b) Level = 123 (c) Coarsest resolution

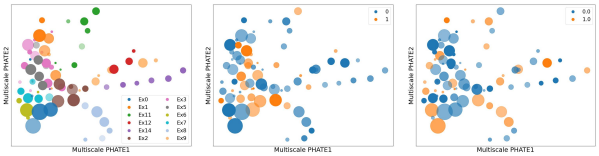
Figure S27: Visualizations after MAGIC imputation in several resolutions.



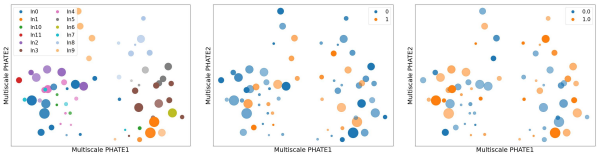
(a) Ast (level = 5)



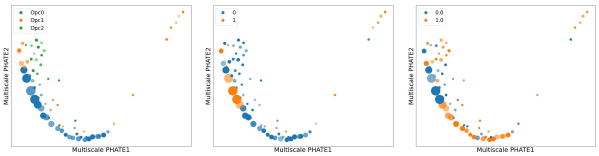
(b) End (Level = 6)



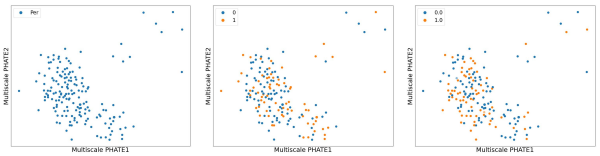
(c) Ex (level = 88)



(d) In (level = 92)

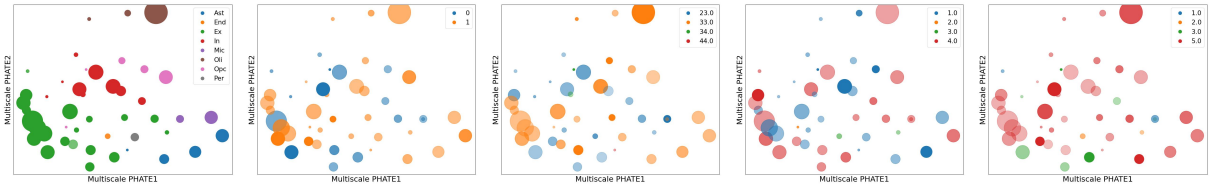


(e) Opc (level = 66)

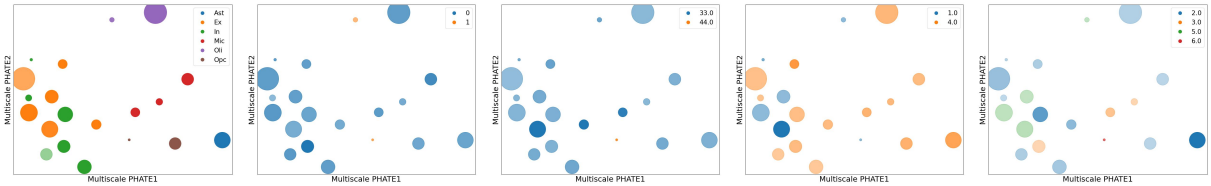


(f) Per (level = 115)

Figure S28: Visualizations of each celltype in best resolution and labeling in sub-cluster, diagnosis and sex. In each sub-figure, the legend is: LEFT: sub-clusters, MIDDLE: Diagnosis, RIGHT: Sex



(a) Female



(b) Male

Figure S29: Visualizations with multiscale PHATE for male and female in level 93, and labeling in celltype, diagnosis, CERAD score and Braak stages (left to right).