# Topological Analysis for Single-Cell RNA-seq data Identified in the ALM region of a Mouse Brain

Zhaoyi Guo

## 1. Introduction

### 1.1 The Motivation

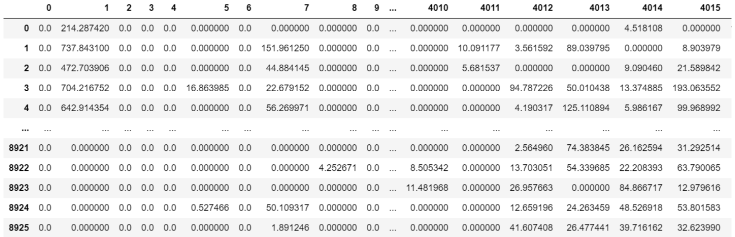
Can we analyze the topological features for single-cell RNA-seq data identified in a mouse brain? This is a question concerned by both genealogists and people in the medical field. Genealogists would want to know the topological structures for different types of genes to lead to new discoveries. People in the medical field would want the topological features for RNA-seq data to provide new medical targets in the future.

Because of this, I'm interested in computing the topological persistence profile and investigating the topological structure for the RNA-seq data. Hopefully, in this project, we can understand this RNA-seq data through a topological scope and gain more topological insights for both genealogists and people in the medical field.

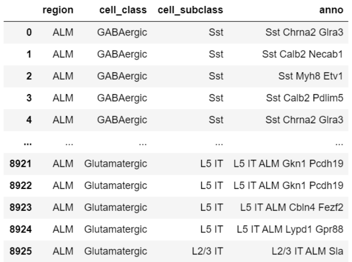
### 1.2 Dataset

The datasets I used contain single-cell RNA-seq data identified in the ALM (anterior lateral motor cortex) region in a mouse brain. There are two data files. The first one is *features.txt*. It contains vectors, which can be interpreted as high-dimensional point clouds. There are 8925 rows and 4020 columns. Each row is a single cell, while each column is a type of gene. The value in each cell is a gene expression quantification for a certain gene (column) and a certain cell (row), and this dataset is the main dataset for topological analysis.

The second dataset is anno.txt. It is the interpretation for every single cell in the first dataset. The first column is the brain region for each cell. In this case, all the cells' brain regions are ALM. The second column is the cell class or the chemical agents for each cell. There are two types of chemical agents in this column: GABAergic and Glutamatergic. GABA is considered an inhibitory neurotransmitter that decreases activity in our nervous system [1]. Glutamate is a powerful excitatory neurotransmitter that is released by nerve cells in the brain, and I will use these two types of neurotransmitters for my topological analysis [2]. Then, the third column is the cell's subclass, and the fourth column is the cell's annotation.



| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | ... | 4010 | 4011 | 4012 | 4013 | 4014 | 4015 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0.0 | 214.287420 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.000000 | 0.000000 | 0.0 | ... | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 4.518108 | 0.000000 |
| **1** | 0.0 | 737.843100 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | 151.961250 | 0.000000 | 0.0 | ... | 0.000000 | 10.091177 | 3.561592 | 89.039795 | 0.000000 | 8.903979 |
| **2** | 0.0 | 472.703906 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | 44.884145 | 0.000000 | 0.0 | ... | 0.000000 | 5.681537 | 0.000000 | 0.000000 | 9.090460 | 21.589842 |
| **3** | 0.0 | 704.216752 | 0.0 | 0.0 | 0.0 | 16.863985 | 0.0 | 22.679152 | 0.000000 | 0.0 | ... | 0.000000 | 0.000000 | 94.787226 | 50.010438 | 13.374885 | 193.063552 |
| **4** | 0.0 | 642.914354 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | 56.269971 | 0.000000 | 0.0 | ... | 0.000000 | 0.000000 | 4.190317 | 125.110894 | 5.986167 | 99.968992 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **8921** | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.000000 | 0.000000 | 0.0 | ... | 0.000000 | 0.000000 | 2.564960 | 74.383845 | 26.162594 | 31.292514 |
| **8922** | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.000000 | 4.252671 | 0.0 | ... | 8.505342 | 0.000000 | 13.703051 | 54.339685 | 22.208393 | 63.790065 |
| **8923** | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.000000 | 0.000000 | 0.0 | ... | 11.481968 | 0.000000 | 26.957663 | 0.000000 | 84.866717 | 12.979616 |
| **8924** | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.527466 | 0.0 | 50.109317 | 0.000000 | 0.0 | ... | 0.000000 | 0.000000 | 12.659196 | 24.263459 | 48.526918 | 53.801583 |
| **8925** | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | 1.891246 | 0.000000 | 0.0 | ... | 0.000000 | 0.000000 | 41.607408 | 26.477441 | 39.716162 | 32.623990 |



| | region | cell_class | cell_subclass | anno |
|---|---|---|---|---|
| **0** | ALM | GABAergic | Sst | Sst Chrna2 Glra3 |
| **1** | ALM | GABAergic | Sst | Sst Calb2 Necab1 |
| **2** | ALM | GABAergic | Sst | Sst Myh8 Etv1 |
| **3** | ALM | GABAergic | Sst | Sst Calb2 Pdlim5 |
| **4** | ALM | GABAergic | Sst | Sst Chrna2 Glra3 |
| **...** | ... | ... | ... | ... |
| **8921** | ALM | Glutamatergic | L5 IT | L5 IT ALM Gkn1 Pcdh19 |
| **8922** | ALM | Glutamatergic | L5 IT | L5 IT ALM Gkn1 Pcdh19 |
| **8923** | ALM | Glutamatergic | L5 IT | L5 IT ALM Cbln4 Fezf2 |
| **8924** | ALM | Glutamatergic | L5 IT | L5 IT ALM Lypd1 Gpr88 |
| **8925** | ALM | Glutamatergic | L2/3 IT | L2/3 IT ALM Sla |

**Figure 1.** *Left*: features.txt; *Right*: anno.txt.

# 2. Methods.

## 2.1 Persistence Diagram with Rips Filtration

I computed the RNA-seq data's topological persistence profile using Rips filtration. The Rips complex generalizes the point cloud. Each vertex is a point or a single cell in this case, and a complex is formed only when the distance between two vertices is less than the threshold. In this project, the threshold is the maximum value from the distance matrix computed from the original dataset. Then, the persistence diagram induced by Rips filtration computes the topological features of the data. I implemented this method using the python version of GUDHI.

## 2.2 The Mapper Methodology: UMAP

I used UMAP to visualize the manifold structure of my data. UMAP is a Mapper method for dimension reduction. The Mapper method creates the mapper structure with response to certain filter functions, where connected components are computed by some clustering algorithm. Then, it visualizes the graph skeleton of the Mapper structure using a graph layout algorithm.

# 3. Results

## 3.1. Persistence Diagram with Rips Filtration

I first computed the persistence diagram using the Gudhi software in Jupyter Notebook. To start, I computed the distance matrix. I tried 3 kinds of metrics: Euclidean, Manhattan, and Cosine. Then, I computed the persistence diagram induced by the Rips filtration. I accessed the persistence intervals per dimension. After that, I plotted the persistence diagram for H0 and H1 in two separate plots and kept track of the largest persistent point for each persistence interval.

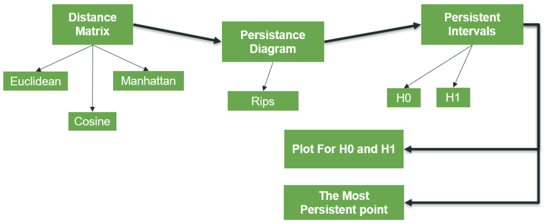


**Figure 2.** Pipeline for computing the persistence diagram.

I plotted six persistence diagrams for each metric and for each dimension. The cluster locations for each metric are different. The persistence diagram (PD) of Euclidean has more points gathering from the birth of 6000 to 14000. The PD of Manhattan has more points gathered from the birth of 170000 to 220000. The PD of Cosine has more points gathered from the birth of 0.03 to 0.13. Thus, persistence diagrams with different metrics have different point clusters.

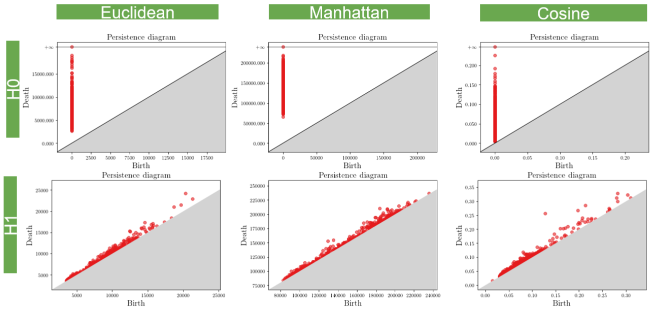
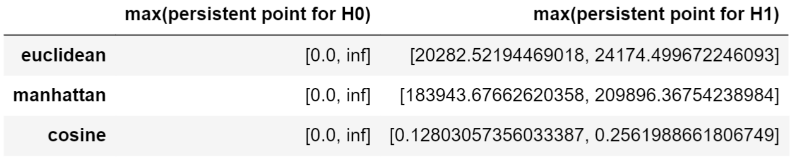


**Figure 3.** *First column:* Persistence Diagrams using Euclidean distance; *Second column:* Persistence Diagrams using Manhattan distance; *Third column:* Persistence Diagrams using Cosine distance; *First row:* Persistence Diagrams for the first dimension; *Second row:* Persistence Diagrams for the second dimension.

In addition to the locations of the point clusters, there are distinct differences in scale. All three of the most persistent points for H1 are drastically different. The most persistent point for Manhattan is the largest, and the most persistent point for Cosine is the smallest. It happens because the cosine distance between two vertices is the degree of angle between the two vectors. Cosine similarity is used to determine how similar two vectors are, so the cosine distance is small. On the other hand, the Manhattan distance between two points in n-dimensional space is the sum of the distances in each dimension, so its persistent point is greater than the Cosine distance and the Euclidean distance [3].



| | max(persistent point for H0) | max(persistent point for H1) |
|---|---|---|
| **euclidean** | [0.0, inf] | [20282.52194469018, 24174.499672246093] |
| **manhattan** | [0.0, inf] | [183943.67662620358, 209896.36754238984] |
| **cosine** | [0.0, inf] | [0.12803057356033387, 0.2561988661806749] |

**Figure 4.** *First row:* The most persistence points using Euclidean distance; *Second row:* The most persistence points using Manhattan distance; *Third row:* The most persistence points using Cosine distance; *First column:* The most persistence points for the first dimension; *Second column:* The most persistence points for the second dimension.

## 3.2 The Mapper Methodology: UMAP

Next, I downloaded the basic libraries for UMAP. Using UMAP, I found the low dimensional representation of the data by fitting and transforming the data. Then, I visualized the results using matplotlib to draw scatter plots of the transformed data. Since UMAP has a few parameters with great influence on the results, I tried a few values for these parameters: *metric*, *n_components*, *n_neighbors*, and *min_dist*.

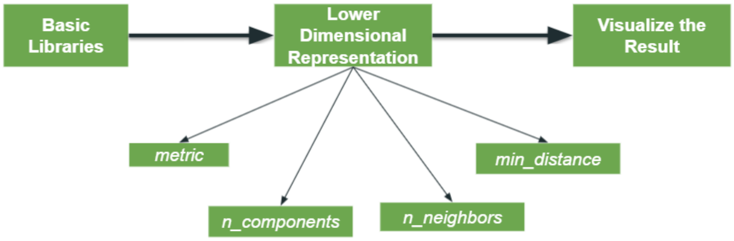


**Figure 5**. The pipeline for implementing UMAP.

The first parameter I looked at is *metric*. The *metric* parameter controls how distance is computed, and I tried Euclidean, Manhattan, and Cosine, which are the same three metrics I used for the persistence diagrams. As shown in **Figure 6**, for all metrics, they were all able to separate chemical agents GABAergic and Glutamatergic, but they have different shapes, and the locations for each agent are different.

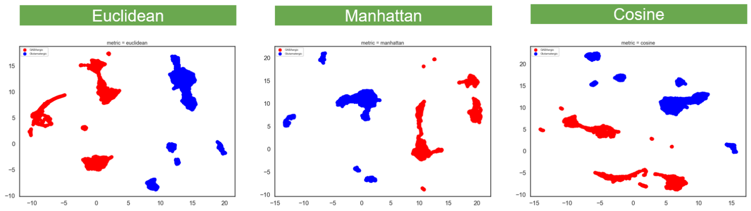


**Figure 6.** *Left*: scatter plot of the UMAP transformed data with the *metric* of Euclidean; *Middle*: scatter plot of the UMAP transformed data with the *metric* of Manhattan; *Right*: scatter plot of the UMAP transformed data with the *metric* of Cosine.

The second parameter I looked at is *n_neighbors*. This parameter controls how UMAP balances local versus global structure in the data. That is, it limits the size of the local neighborhood UMAP so that when *n_neighbors* is small, UMAP focuses on local structure more. Just like what we see in **Figure 7**, both chemical agents formed one big cluster in the middle when *n_neighbors* = 2. Also, as *n_neighbors* increase, it forces UMAP to watch for a broader region of each point when estimating the manifold structure of the data. Thus, a larger *n_neighbors* parameter leads to fewer details of the structure [4].

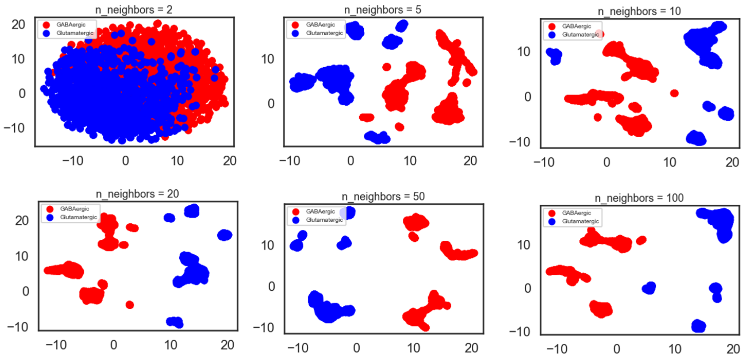


**Figure 7.** *Left to Right and Top to Bottom:* the scatter plots of the UMAP transformed data with the parameter of *n_neighbors* equal to 2, 5, 10, 20, 50, and 100.

In addition to *n_neighbors*, I also considered *min_dist*. The *min_dist* parameter controls how closely UMAP is allowed to pack points together. It provided the minimum distance apart that points are allowed to be in the low dimensional representation, so that smaller *min_dist* creates more clusters. Because of this, when *min_dist* is small, there are more clusters that are small. When *min_dist* is large in value,  there are fewer clusters, and these clusters are big in size [4].

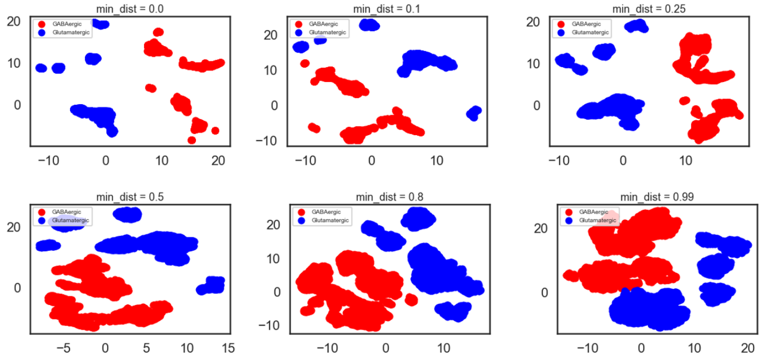


**Figure 8.** *Left to Right and Top to Bottom:* the scatter plots of the UMAP transformed data with the parameters of *min_dist* equal to 0, 0.1, 0.25, 0.5, 0.8, 0.99.

Finally, I looked at the parameter *n_components*. This parameter decides the number of dimensions to be reduced to. I first set *n_components* to 1, which forces UMAP to embed the data in a line. Although there are separations between GABAergic and Glutamatergic, there are a few red dots in the blue clusters, and a few blue dots in the red clusters, so the separation is not clean cut. However, when I changed *n_components* to 2 and 3, the separation between GABAergic and Glutamatergic became clear. Therefore, the more dimensions for UMAP to work with, the easier separation out of the cell class [4].

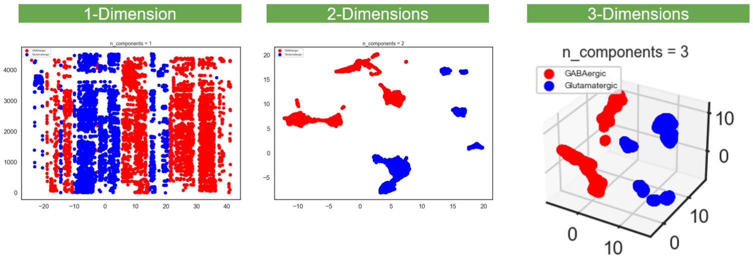


**Figure 9.** *Left to Right and Top to Bottom:* the scatter plots of the UMAP transformed data with the parameters of *n_components* equal to 1, 2, and 3.

# 4. Conclusion

In this project, I did topological analysis for single-cell RNA-seq data in the ALM region of a mouse. By doing this, I found the topological features by computing the persistence diagram with different distance metrics. Also, I did dimension reduction using UMAP and found the global structure of the data with different hyperparameters.

There are a few findings in this study. First, persistence diagrams with different distance metrics can be drastically different in scale. Manhattan has the largest scale, while Cosine has the smallest scale among all three. Also, persistence diagrams with different distance metrics have different topological features because the locations of the point clusters are different.

For UMAP, there are findings for each hyperparameter. For parameter *metric*, scatter plots with different metrics have a clear separation between different cell classes but the locations for each cell class are different. For parameter *n_neighbors*, the smaller the *n_neighbors* are, the more local neighborhood UMAP needs to focus on and vice versa. For parameter *min_dist*, the smaller the *min_dist* is, the clumpier the embeddings will become. The larger the *min_dist*, the bigger and less the clusters are. For parameter *n_components*, the higher the dimension, the easier for UMAP to separate the cell class. The lower the dimension, the less topological structure would be preserved.

In this project, I computed the topological persistence profile of the single-cell RNA-seq data in a mouse brain using Rips filtration and identified a few topological features through the persistence diagrams. Also, I studied the space of the data using UMAP and found the topological structure for both GABAergic and Glutamatergic chemical agents in the ALM region of the mouse brain. Hopefully, the results of this study could help both genealogists and people who work in the medical field to explore the RNA-seq data through the topological scope and to use topological analysis for gene expression analysis [5].

## 5. References

1. Westphalen, Pharm.D., Dena. “What Does Gamma Aminobutyric Acid (GABA) Do?” *Healthline*, https://www.healthline.com/health/gamma-aminobutyric-acid.

2. Liou, Stephanie. “About Glutamate Toxicity.” *HOPES Huntington's Disease Information*, 18 Nov. 2014, hopes.stanford.edu/about-glutamate-toxicity/.

3. Foundation, OpenGenus. “Manhattan Distance (L1 Norm).” *OpenGenus IQ: Computing Expertise*, OpenGenus IQ: Computing Expertise, 24 Dec. 2018, iq.opengenus.org/manhattan-distance/#:~:text=Manhattan%20 distance%20is%20a%20 distance,all%20 dimensions%20of%20two%20 points.

4. “Basic UMAP Parameters.” *Basic UMAP Parameters* – Umap 0.5 Documentation, umap-learn.readthedocs.io/en/latest/parameters.html.

5. Covert, Derek . “*Topological methods for gene correlation analysis of RNA-seq data sets* “,1 Dec. 2017, knot.math.usf.edu/multimedia/Topological_Methods_for_Gene_Correlation_Analysis_of_RNA-seq_Data_Sets.pdf