# Investigating different cell types through an analysis of persistent images of axonal trees

Kerem Kurban
and
Leonardo Pollina

Contact emails: kerem.kurban@epfl.ch and leonardo.pollina@epfl.ch

## *Introduction*

The field of neuroscience has seen significant advancements in dendrite analysis through the application of persistent image analysis [1]. However, there remains a noticeable gap in the study of axons [2]. The advent of new long-range datasets using viral tracing has made it possible to obtain high-quality morphology data, offering an opportunity to fill this gap [2,3,5,6].

The complexity of innervation patterns and the lack of readily available projection classification patterns present a challenge, as most neurons are simply labeled as 'projecting' [2]. This work aims to bridge this gap by trying to identify different sub-classes of projecting axons through an analysis of persistent images obtained via the radial distance-based filtration method or the path distance-based filtration method. We hypothesize that it is possible to identify sub-populations of morphological types for those morphologies that are currently labeled with the same broad type. We aim at comparing different clustering algorithms to identify these sub-populations and at investigating the effect that the choice of filtration method has when using axonal persistent images.

We applied dimensional reduction and clustering techniques on persistent images computed over the axonal trees of morphologies downloaded from the MouseLight project of Janelia. We also used region and cell type labels from Neuromorpho to determine if clusters have significant meaning [2]. We show here that very specific cell types, such as Granule Cells and Purkinje cells, present very well-defined clusters after the projection performed via dimensionality reduction of the persistent images. We then focused on those cells labeled as principal projecting cell or principal projecting pyramidal cells and identify morphological subclusters. We finalized our analysis with a statistical validation of the differences between the clusters identified using several morphometrics computed.

This report identifies gaps in the current literature and proposes future work to move towards the use of the automated morphology classification tools as the true labels for cell classes [4].

# Methods

## *Data Acquisition*

We fetched the morphologies files from Janelia and converted the JSON format to SWC. The morphologies analyzed in the frame of this project all come from mice (the MouseLight project). We then fetched metadata from Neuromorpho and extracted region information from soma position using the ccfv3 25um atlas. Regarding the cell type information, it was extracted from Neuromorpho metadata since this was not available in Janelia. We also combined under a single region label the morphologies belonging to similar regions or sub-layers of the same region (especially for the motor area or for the dentate gyrus). We filtered out all those cells belonging to brain regions with less than 30 occurrences (samples).

## *Persistent Images Generation*

The Topological Morphology Descriptor (TMD) was utilized to generate persistent diagrams and images. To normalize each persistent image (PI), the image's pixel values were divided by the sum of all pixel values, ensuring that the total sum equaled 1. This normalization process effectively transformed the PI into a probabilistic map. Additionally, the PI was scaled by multiplying it with the number of components found in its corresponding persistent diagram, accounting for variations in the number of branches. This scaling can be interpreted as a representation of the probabilistic map illustrating the number of branches present in each neuron's axonal tree. This approach enabled us to distinguish between morphologies exhibiting similar structures but differing in branch count. The resulting scaled-normalized images were subsequently utilized for further analysis. Lastly, all persistent images were adjusted to align with the same x- and y-axes, facilitating comparative evaluations. Note that the persistent images were obtained both by using the ‘radial distance’ filtration method and the ‘path distance’ filtration method.

## *Feature Extraction*

One of the primary objectives in our analysis was to extract meaningful features that could be utilized in subsequent analysis steps. As previously mentioned, we focused on the persistent images, each having a dimension of 100x100. To condense the information contained within these images into a more manageable form, we employed t-distributed Stochastic Neighboring Embedding (t-SNE) as an unsupervised dimensionality reduction method [7]. This allowed us to represent the information residing in the persistent images using just two components, effectively reducing the dimensionality to two features. The choice of two features was based on convenience, considering both computational resources and illustrative purposes.

In brief, t-SNE is a technique that aims to visualize high-dimensional data by mapping it to a lower-dimensional space. It preserves both local and global structure by modeling pairwise similarities between data points. By reducing the dimensionality to two components, t-SNE facilitates the identification of patterns, clusters, and relationships that may be obscured in the original high-dimensional space.

To determine the perplexity value, a key hyperparameter in t-SNE, we followed a common rule of thumb. The perplexity value was set to the square root of the number of samples present in our dataset. This rule of thumb is often employed not only in t-SNE but also in determining the optimal value of 'k' when using the k-Nearest Neighbors algorithm [8].

### *Clustering Algorithms*

To identify potential clusters in the data projected on 2 dimensions after dimensionality reduction via t-SNE, we employed two different unsupervised clustering algorithms: GMM (Gaussian Mixture Model) and DBSCAN (Density-Based Spatial Clustering of Applications with Noise). These algorithms require the setting of important hyperparameters to control their behavior and performance.

For GMM, one crucial hyperparameter is the number of clusters, which needs to be specified in advance. This parameter determines the shape and number of clusters that GMM will attempt to fit to the data. Selecting an appropriate number of clusters can be challenging, and various techniques such as information criteria or cross-validation can be employed to determine the optimal value.

In the case of DBSCAN, two key hyperparameters are epsilon (ε) and minimum points (Min_Pts). ε defines the maximum distance between neighboring points for them to be considered part of the same cluster. It influences the density requirement for points to be connected in a cluster. Min_Pts determines the minimum number of points required to form a dense region or core point. It helps in identifying dense areas and filtering out noise points. Tuning these hyperparameters is crucial to obtain meaningful clusters, and it often requires some experimentation and understanding of the dataset's characteristics.

By utilizing both GMM and DBSCAN, we aimed to comprehensively explore the data, taking advantage of the complementary capabilities of the two algorithms to gain deeper insights into the underlying clustering structures.

### *Optimization*

In order to determine the optimal number of clusters k in GMM, we utilized the silhouette method. The silhouette method calculates the average silhouette score for each value of k, which measures the cohesion within clusters and separation between clusters. Higher silhouette scores indicate better-defined clusters. However, it is important to remember that there is a component of randomness in GMM clusters that arises from the different random initializations used in the algorithm. For this reason and to mitigate this randomness in the identification of optimal k, GMM was run multiple times (N=20) and the average silhouette score was computed in order to select the ideal number of clusters.

For parameter selection in DBSCAN, we performed a grid search, exploring different combinations of epsilon ($\varepsilon$) and minimum points (Min_Pts), while measuring the resulting silhouette score. This allowed us to identify the best parameters that yield meaningful clusters with high inter-cluster separation and low intra-cluster dissimilarity.

### *Morphometrics*

As part of our analysis, we aimed at comparing the clusters identified by the different clustering algorithms at the level of several morphometrics extracted for each axonal tree of each neuron of interest. In particular, 21 features were extracted: number of bifurcations, number of sections per neurite, number of leaves, partition asymmetry, length of partition asymmetry, remote bifurcation angles, section areas, section bifurcations branch orders, section bifurcation lengths, section bifurcation radial distances, section branch orders, average section length, section path distances, section radial distances, section Strahler orders, section terminal branch orders, section terminal lengths, section terminal radial distances, section tortuosity, section volumes, and total length per neurite. Note that in the case of metrics involving multiple values, the average was computed in order to have a single value to be compared across clusters. The different morphometrics were normalized across all the neurons of the dataset for illustration purposes while ensuring that the differences between cells (and clusters) were maintained.

### *Statistical Analysis*

Statistical tests were performed to assess effective differences between clusters when focusing on a specific metric. In the case of only 2 clusters identified, an independent t-test was performed for each metric to test the difference between the clusters (using as significant threshold $\alpha$= 0.01). When multiple clusters were found, an ANOVA followed by post-hoc Tukey tests was performed by keeping an overall family-wise error of 0.01.

# Results

The focus of our analysis was on the information contained on the persistent images when looking at the axonal trees. In **Figure 1**, we report a few examples of persistent images (after normalization and scaling). Note how the two neurons have the same label, but the persistent images look dramatically different.

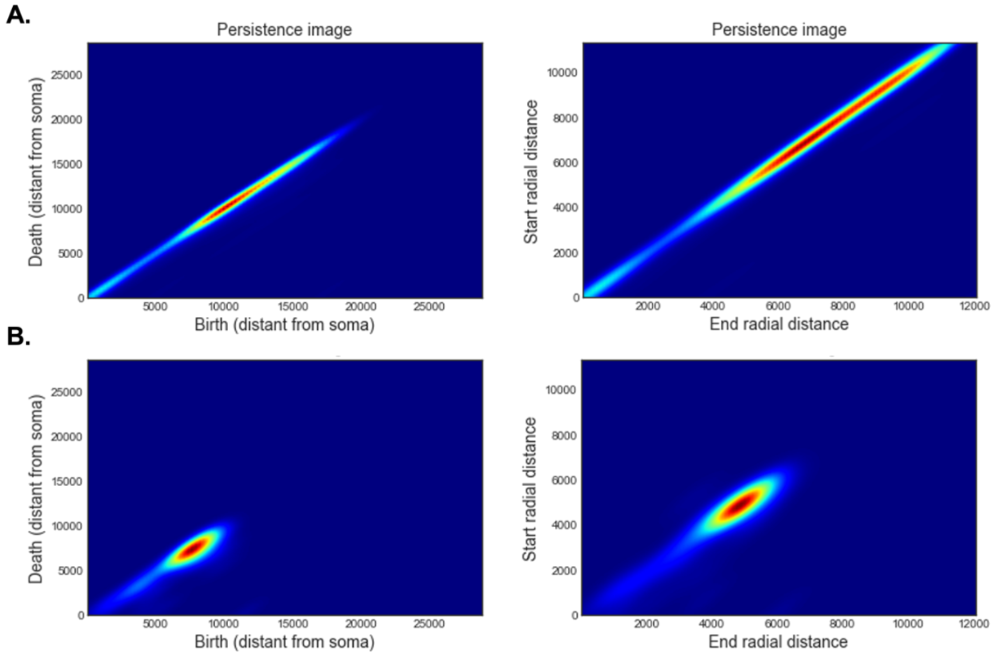

**Figure 1:** Examples of persistent images for two different cells, both labeled as 'principal cell'-'projection'-'pyramidal'. Images are normalized and scaled to the same axes. Each image has been multiplied by the number of components in the corresponding persistent diagram. On the left the persistent images obtained via the path distance filtration method are shown, while on the right the radial distance has been used.

As a first result, we could observe that specific regions (**Figure 2**) corresponding to very specific cell types (**Figure 3**), such as granule cells for the Dentate Gyrus or the Purkinje cells for the Cerebellum, were very well clustered when looking at the scatter plot after dimensionality reduction via t-SNE. This indicates that these types of cells can be very well

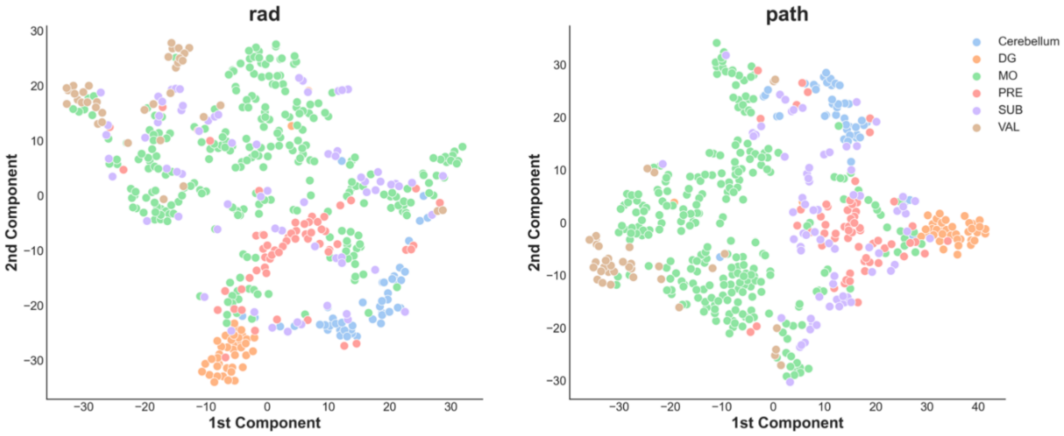

**Figure 2:** t-SNE projection on two dimensions of the persistent images obtained by radial distance (on the left) and path distance (on the right) filtration over the axonal trees. Each dot represents a neuron and colors indicate different brain regions. Very peculiar regions such as the Dental Gyrus (DG) or the Cerebellum show very clustered points.

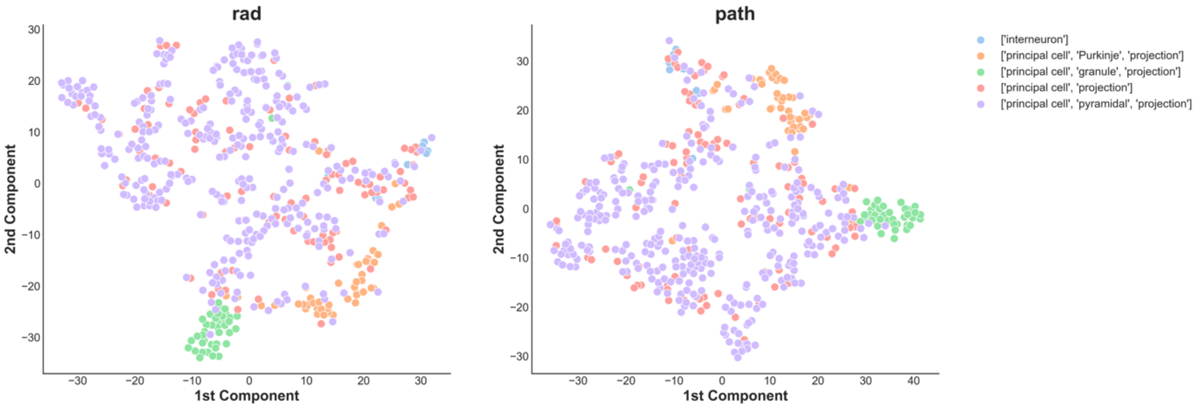

**Figure 3:** t-SNE projection on two dimensions of the persistent images obtained by radial distance (on the left) and path distance (on the right) filtration over the axonal trees. Each dot represents a neuron and colors indicate different morphology types. Very specific types such as Purkinje cells and granule cells show very clustered points.

classified and clustered when using the persistent images of the axonal trees and that the dimensionality reduction method is useful to capture these differences.

On the other side, we can observe that other types of cells, especially the ones referred to as pyramidal projection cells or simply projection cells are very spread (even when considering their large number).

Following this observation and for all subsequent analysis steps, we decided to focus on only the cells being labeled as 'principal-projection cells' or 'principal-pyramidal-projection cells'. We hypothesize that within this subpopulation sub-types can be identified being the labeling very broad in meaning and description.

With the goal of identifying possible sub-clusters, we employed GMM as an unsupervised clustering method. To optimize and find the number of clusters in a data-driven way, we used the Silhouette method, which consists in computing the Silhouette score for different numbers of cluster to identify the k yielding the highest score. In our case, we found k = 2 when using the radial distance and k = 4 when using the path distance.

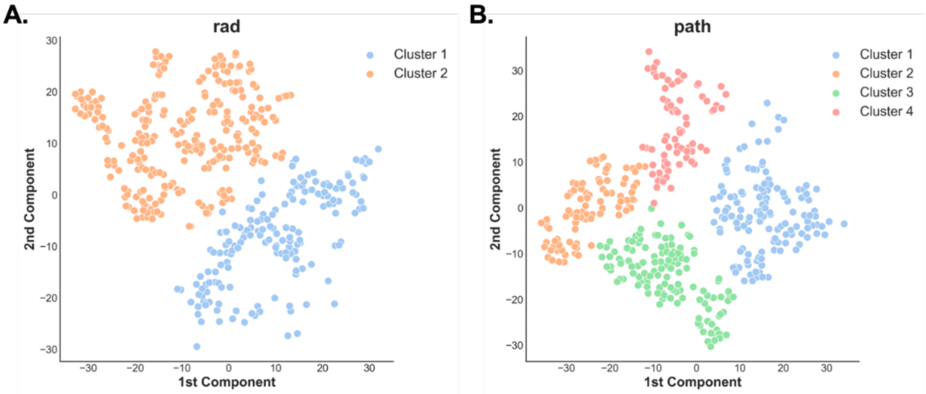

**Figure 4:** The clusters identified by the GMM clustering algorithm: radial distance in **A**, path distance in **B**. Each dot represents a different neuron, and the points are color coded depending on the cluster they have been assigned to.

The neurons labeled with the labels computed by the GMM algorithm fitted with the optimal k identified are shown in **Figure 4**.

We decided to also explore a different clustering approach by using a more flexible algorithm like DBSCAN, which for example has the advantage of being able to handle noisy points and not assign them to any cluster. In the case of DBSCAN, the identification of optimal hyperparameters, such as ε and Min_Pts, is crucial. To find also in this scenario the optimal values in a data-driven fashion, we employed again the Silhouette score, and we ran a grid search over the two hyperparameters of interest.

The combination of hyperparameters yielding the highest Silhouette score was kept. $\varepsilon$ was chosen among 100 values ranging in [0.1, 5], while Min_Pts could vary between 5 and 100, with a increment step of 5. We found that for radial distance, the best combination was $\varepsilon$ = 4.11, Min_Pts = 10, while for path distance it was $\varepsilon$ = 4.65, Min_Pts = 15.

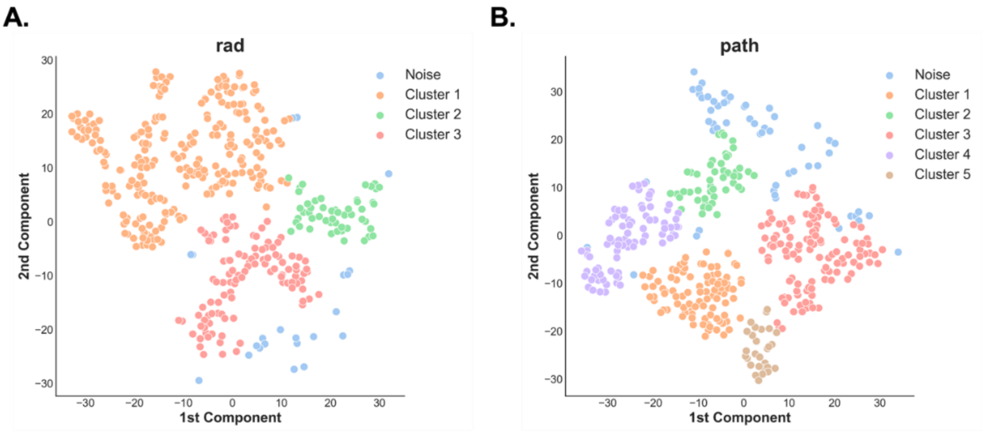

**Figure 5:** The clusters identified by the DBSCAN clustering algorithm: radial distance in **A**, path distance in **B**. Each dot represents a different neuron, and the points are color coded depending on the cluster they have been assigned to. Note that DBSCAN identifies also points being considered as noise (colored in blue).

In **Figure 5**, we show that the clusters identified from the DBSCAN algorithm fitted with the data coming from the radial or path distance and using the corresponding optimal combination of hyperparameters. For the radial distance, the algorithm identified 3 clusters, while for the path distance the number of clusters found was equal to 5.

As shown throughout the whole text, our analysis was always performed by using the radial and the path distances in parallel. This was done in order to assess their similarity and investigate whether these two filtrations methods give converging results or not when it comes to explore persistent images of axonal trees. To quantify the level of similarity of the clusters identified through the two different filtration methods, we computed the amount of overlapping between the different clustering in percentage. This showed us if the clusters identified by the two methods were similar or not and, if no, what kind of division occurred. These overlapping values are presented in **Figure 6**.

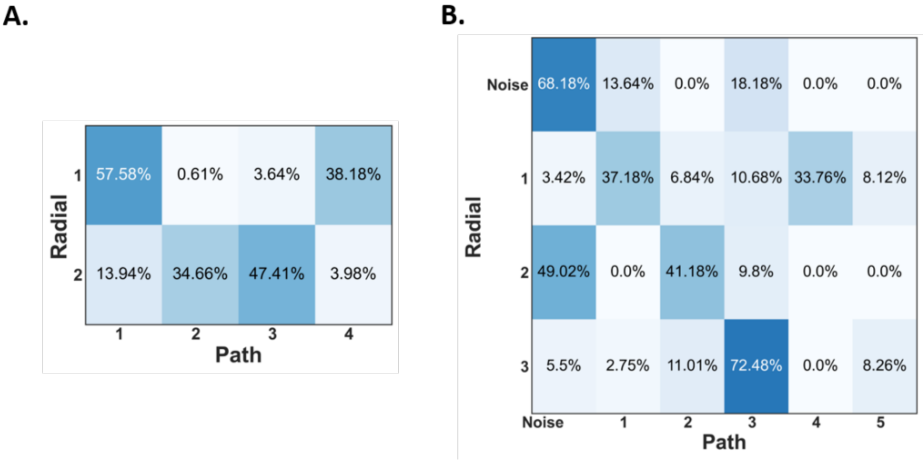

**Figure 6:** The amount of overlap between the clusters identified through radial or path distance by GMM (**A**) and DBSCAN (**B**). High values (closer to 100%) man that two clusters are highly overlapping and thus the same. For DBSCAN, also the noise "clusters" is presented in order to account for the totality of the points.

As we know from the previous analysis, the radial distance gave us a lower number of clusters no matter the clustering algorithm used. From **Figure 6A** we can see that the first radial distance cluster corresponds largely to the first and fourth path distance clusters, while the second radial distance cluster has been divided in the second and third path distance clusters.

Regarding DBSCAN in **Figure 6B**, we can observe immediately that the noise neurons are largely consistent, which means that the clustering algorithm identifies the same cells as being noise or outliers, independently on the filtration method used to obtain the persistent images.The third radial distance cluster overlaps largely with the third path distance cluster. The first radial distance cluster seems to be divided into the first and the fourth path distance clusters. Finally, the second radial distance cluster overlaps largely with the second path distance cluster, but interestingly, a lot of its points are labeled as noise when it comes to the path distance analysis. Also, note how the smallest cluster, that is the fifth cluster in path distance, contains neurons assigned either to the first or the third clusters in radial distance.

Finally, as a last step to our analysis we sought to investigate and characterize the clusters identified by the different clustering algorithms by comparing them on the basis of several morphometrics we extracted. For each combination of filtration method and clustering algorithm, we compared statistically the clusters on each morphometric independently. The results are presented in **Figures 7-10**. Whenever only two clusters were present, a simple independent t-test was performed between the two clusters to assess any significant differences. On the other hand, when multiple clusters were identified by either GMM or DBSCAN, we perform a one-way ANOVA followed by post-hoc Tukey tests to control for multiple comparisons. The overall significant threshold was always kept at 0.01. Overall, what appears evident from the analysis of the different morphometrics is that indeed the clusters appears to be consistently different for a number of morphometrics and in all sort

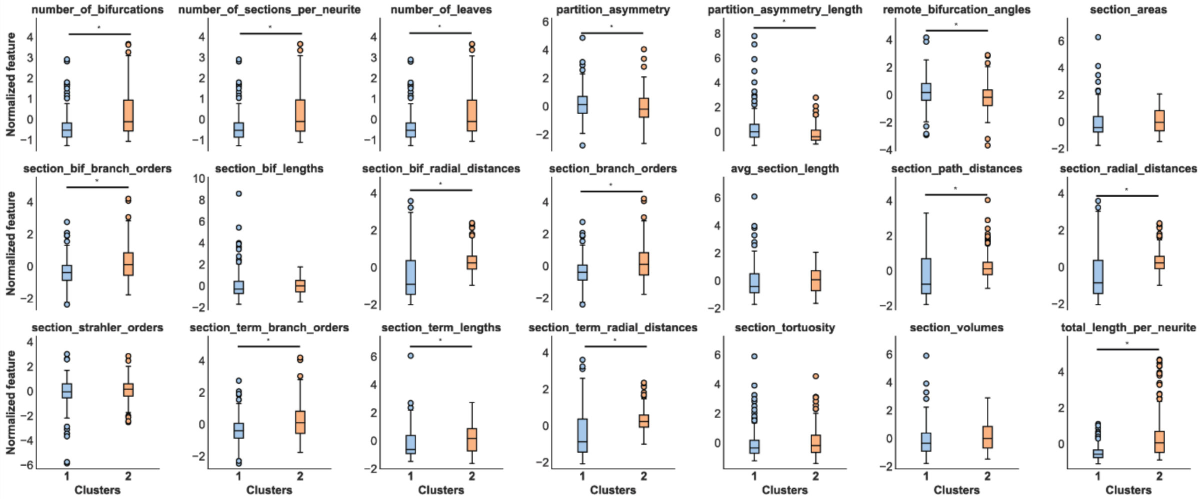

**Figure 7:** Comparison between the two clusters identified by the GMM algorithm when using the radial distance filtration method for the computation of the persistent images. For the metrics consisting of multiple values, the mean is computed and used to have a single value per neuron. * represents statistical significance, independent t-test, $p < 0.01$.

of combinations. This points to the idea that the clusters identified by GMM and/or DBSCAN indeed presents morphologically different characteristics.

Interestingly, a lot of significant differences are also found when using DBSCAN as clustering algorithm and path distance as a filtration method. This is the combination identifying the highest number of clusters (i.e. 5, without considering the neurons labeled as noise). Note how cluster 5 is found to be significantly different from other clusters, especially from clusters 1 and 3, which appear to be the ones closest to it when using the 2 projected components from t-SNE. This highlights how this small clusters seems to still be relevant and indeed morphologically different from the two other bigger ones.

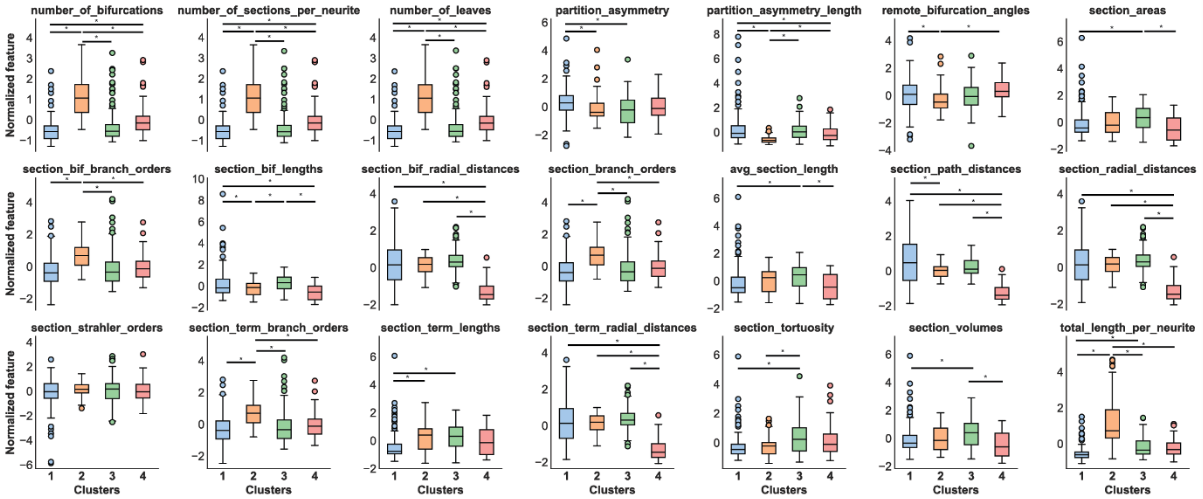

**Figure 8:** Comparison between the two clusters identified by the GMM algorithm when using the path distance filtration method for the computation of the persistent images. For the metrics consisting of multiple values, the mean is computed and used to have a single value per neuron. * represents statistical significance, independent t-test, $p < 0.01$.

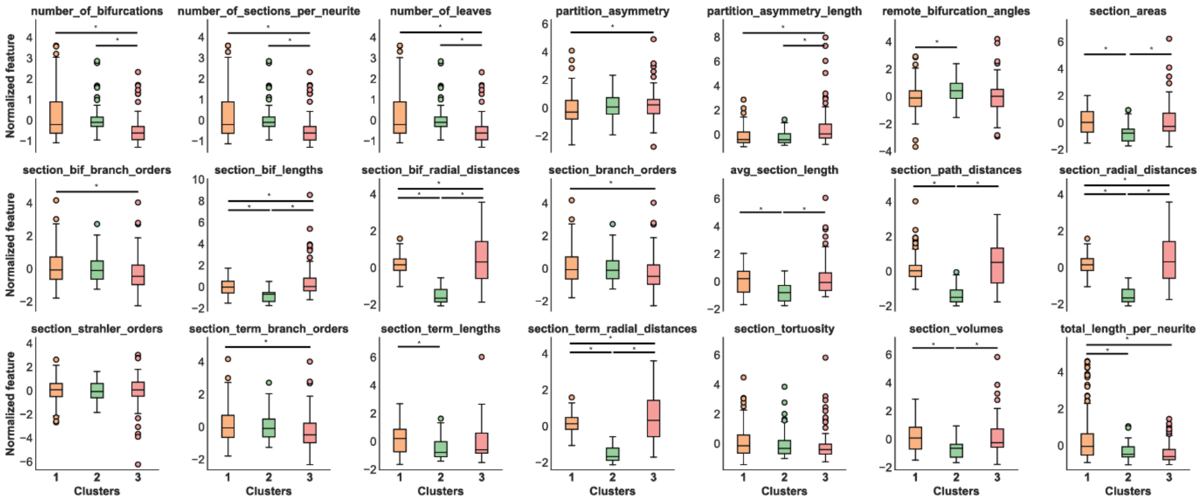

**Figure 9:** Comparison between the two clusters identified by the DBSCAN algorithm when using the radial distance filtration method for the computation of the persistent images. For the metrics consisting of multiple values, the mean is computed and used to have a single value per neuron. * represents statistical significance, independent t-test, $p < 0.01$.

It is also worth mentioning how the 'section_strahler_orders' is the only metric resulting as non-significant for all comparisons of clusters in all combinations of clustering algorithm and filtration method.

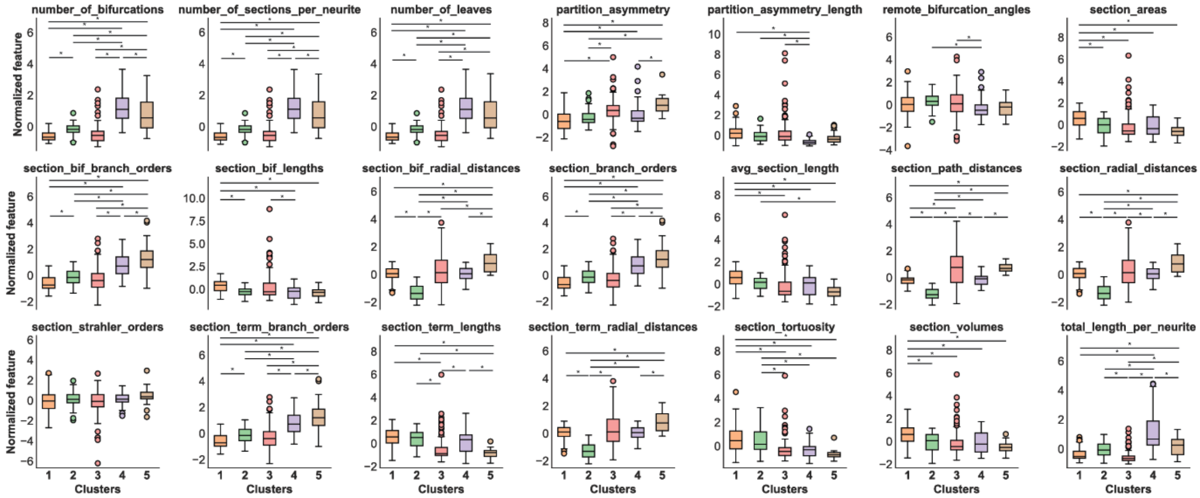

**Figure 10:** Comparison between the two clusters identified by the DBSCAN algorithm when using the path distance filtration method for the computation of the persistent images. For the metrics consisting of multiple values, the mean is computed and used to have a single value per neuron. * represents statistical significance, independent t-test, $p < 0.01$.

## Discussion and Future Directions

Our investigation offers promising insights into a potential classification of projecting axons, demonstrating the potential of our approach to identify sub-populations of neurons. The rationale for such a research questions stems from the need for a more accurate labeling of neurons, since, as mentioned above, multiple cells have the same broad label even when they belong to different morphological types. The identification of distinct clusters based on cell type suggests that our methodology can yield meaningful insights into the morphological diversity of axons.

One of the main goals of the present project was to investigate the differences observed when employing different filtration methods for the obtention of the persistent images. It appears evident from the present results that the two metrics do not give convergent results in terms of clustering, with the exception of a few cases. We believe that this could be explained before of the nature itself of our data, that focuses on the long-range axons of neurons. With such long and probably curved structures (differently from what happens with dendritic trees), the two filtration methods can indeed yield very different results. We think that one option could be to actually combine the two different persistent images and perform the clustering on this combination of features to understand if new and more consistent clustering patterns arise.

The second main focus was the comparison of different clustering algorithms to identify morphological different sub-populations. We focused on GMM and DBSCAN, two popular algorithms using distinct approaches. We also considered k-means, but similar results to GMM were found and we considered GMM to be a better and more robust algorithm than k-means. Both our clustering algorithms managed to identify important clusters, even though we could observe the general pattern where the radial distance-based dataset was divided in a lower number of clusters with respect to its path distance-based counterpart. This again points towards the idea that radial distance might not be the best option when investigating long-range neural trees such as axonal trees. Again, a combination of both could also be explored. It is worth mentioning that the main hyperparameters of both methods were optimized in a data-driven way. However, while the choice of k = 4 for GMM path distance seemed obvious, the same does not hold for the choice of k = 2 for GMM radial distance. Indeed, k = 2 represents the lower number of clusters we tried, but it could also as well be that no evident clusters at all can be identified in this combination of algorithm and filtration method. On the other hand, the optimization of the hyperparameters of DBSCAN seemed successful considering that the identified optimal values were well situated in the range of values we proposed. Moreover, the ability of DBSCAN to detect neurons presenting noisy values helped us keep the identification of interesting clusters cleaner.

The last step of our investigation consisted in the validation of the identified clusters via their comparison through several morphometrics computed. The presence of a lot of significant differences is reassuring and points towards the conclusion that these clusters could represent real morphologically different populations. It is interesting to notice that a lot of significant differences were found also in the combination DBCSAN path distance, which is the one combination yielding 5 different clusters. While the characterization of the identified sub-population is out of the scope of the current investigation, it is worth mentioning that, from the results obtained while comparing clusters for the 21 morphometrics, it seems that the amount of branching of the neurite plays a role in the distinction of the clusters. However, the overall complexity of the branching seems to be the same across the neurons analyzed, as it is shown from the consistent lack of significance when comparing the Strahler orders. A deeper analysis of these morphometrics would be required and we could for example summarize which metrics result significant for which cluster, in order to morphologically characterize better the different clusters.

Finally, it is important to remember that, for sake of conciseness and to keep the final step relatively simple, we averaged those morphometrics giving not a single value per neuron but a whole vector. This could be okay for those metrics presenting normal distributions for example, but it is probably limited for many others. A more thorough and separate analysis could concern only these “multi-values” metrics.

A potentially complementary approach and intuitive continuation of this project could be to use the morphometrics computed to perform classification over the identified clusters. This could give additional insights with respect to the statistical analysis where the morphometrics are treated independently. Indeed, if a classifier such as Random Forest was to be used, one could learn more intuitively which features help the classification the most, hence acquiring knowledge on which feature is the most different across the sub-populations studied.

An interesting alternative to our clustering step of the data could be the idea of employing semi-supervised clustering methods. In this way we could leverage the benefits of both supervised and unsupervised methods. As shown at the beginning of our investigation, very peculiar morphological types such as granule cells and Purkinje cells seem to be well clustered when projected on the 2 first components with t-SNE, no matter if using the radial or the path distance filtration method. We could use these cells as labeled information to enhance the identification of new clusters.

Finally, it is worth remembering that in this project we focused on the use of the persistent images as a method to identify different morphological types when investigating axonal trees, but other features could be used indeed. One could also think to perform the reverse pipeline with respect to what we did here, that is performing a dimensionality reduction (via t-SNE) on the 21 morphometrics extracted and then look at the average persistent images for each cluster.

Still speculating about possible future directions, we could think of incorporate more long-projecting neuron reconstructions into our dataset. By combining data from different labs, such as the Allen Institute [4, 5], we can improve the generalizability of our findings and contribute to a more comprehensive understanding of axon morphology.

In conclusion, our work represents a step towards a more detailed understanding of axon morphology and classification. We believe that our approach, combined with the use of new tools and datasets, could significantly contribute to the advancement of neuroscience research. Future work will focus on refining our methodology, expanding our dataset, and exploring new ways to classify neurons based on their axonal morphology.

## References

1. Lida Kanari, Srikanth Ramaswamy, Ying Shi, Sebastien Morand, Julie Meystre, Rodrigo Perin, Marwan Abdellah, Yun Wang, Kathryn Hess, Henry Markram, Objective Morphological Classification of Neocortical Pyramidal Cells, Cerebral Cortex, Volume 29, Issue 4, April 2019, Pages 1719-1735

2. Ascoli, G. A., Donohue, D. E., & Halavi, M. (2007). NeuroMorpho. Org: a central resource for neuronal morphologies. Journal of Neuroscience, 27(35), 9247-9251.Chicago

3. Gerfen, C. R., Economo, M. N., & Chandrashekar, J. (2016). Long distance projections of cortical pyramidal neurons. In Journal of Neuroscience Research (Vol. 96, Issue 9, pp. 1467–1475). Wiley. https://doi.org/10.1002/jnr.23978

4.[https://github.com/BlueBrain/morphoclass](https://github.com/BlueBrain/morphoclass)

5. Muñoz-Castañeda, R., Zingg, B., Matho, K. S., Chen, X., Wang, Q., Foster, N. N., & Dong, H. W. (2021). Cellular anatomy of the mouse primary motor cortex. Nature, 598(7879), 159-166.

6. Wang, Y., Xie, P., Gong, H., Zhou, Z., Kuang, X., Wang, Y., ... & Veldman, M. B. (2019). Complete single neuron reconstruction reveals morphological diversity in molecularly defined claustral and cortical neuron types. BioRxiv, 675280.

7. Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(11).Chicago

8. Prakash Nadkarni, Clinical Research Computing, 2016