This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

FiberNeat: unsupervised streamline clustering and white matter tract filtering in latent space

Bramsh Qamar Chandio [†⋆] *Tamoghna Chattopadhyay* [⋆] *Conor Owens-Walton* [⋆]
Julio E. Villalon Reina [⋆] *Leila Nabulsi* [⋆] *Sophia I. Thomopoulos* [⋆]
Eleftherios Garyfallidis [†] *Paul M. Thompson* [⋆]

[⋆] Imaging Genetics Center, Mark and Mary Stevens Neuroimaging and Informatics Institute, Keck School of Medicine, University of Southern California, Marina del Rey, CA, USA
[†] Department of Intelligent Systems Engineering, School of Informatics, Computing, and Engineering, Indiana University Bloomington, IN, USA

ABSTRACT

Whole-brain tractograms generated from diffusion MRI digitally represent the white matter structure of the brain and are composed of millions of streamlines. Such tractograms can have false positive and anatomically implausible streamlines. To obtain anatomically relevant streamlines and tracts, supervised and unsupervised methods can be used for tractogram clustering and tract extraction. Here we propose FiberNeat, an unsupervised streamline clustering and tract filtering method. FiberNeat takes an input set of streamlines that could either be unlabeled clusters or labeled tracts. Individual clusters/tracts are projected into a latent space using nonlinear dimensionality reduction techniques, such as t-SNE and UMAP, to find spurious and outlier streamlines. In addition, outlier streamline clusters are detected using DBSCAN and then removed from the data in streamline space. Quantitative comparisons with expertly delineated tracts show the promise of the approach. This approach can be deployed as a filtering step after tracts are extracted.

***Index Terms*—** Tractography, Clustering, t-SNE, UMAP

1. INTRODUCTION

The structural architecture of the brain can be computationally reconstructed from a diffusion magnetic resonance imaging (MRI) [1] dataset using tractography algorithms [2]. Tractography algorithms exploit the direction and paths of water diffusion in neural connections of the brain to generate digital neural pathways, otherwise called streamlines. Streamlines are thus used as a computational approximation of the brain's white matter fibers. Tractography algorithms often generate streamlines that are false positives or anatomically implausible, such as streamlines that loop, that have sharp curves and angles, or that terminate prematurely in white matter, or connect anatomically implausible regions of the brain [3], [4].

In the past two decades, researchers have used both supervised and unsupervised white matter tract segmentation methods to reduce the number of false positive streamlines in the data. The unsupervised category focuses on clustering methods [5], [6] that divide whole-brain tractograms into clusters of streamlines that are spatially similar in shape and size. Resultant clusters often suffer from spurious streamlines or poor alignment with neuroanatomical definitions of the tracts. Furthermore, these clusters are unlabeled and can have sub-clusters within one cluster. The supervised category consists of white matter tract segmentation methods that are trained with pre-labeled datasets. Automatic traact segmentation methods include ROI-based [7], atlas-based [8], [9], and deep learning-based methods [10]. Although such supervised methods result in labeled streamlines that match their anatomical tract definitions, they can still produce spurious streamlines due to biases stemming from limitations of the prior anatomical reference, subject variability, and tractography reconstruction issues. Moreover, different tract segmentation methods may rely on different definitions of the same tracts [11].

In this paper, we propose FiberNeat, a method which uses dimensionality reduction techniques t-SNE (t-distributed stochastic neighbor embedding) [12] and UMAP (uniform manifold approximation and projection) [13] to find and remove outlier streamlines in latent space [1]. The input to FiberNeat is a set of streamlines that can either be unlabeled clusters or labeled tracts. It populates an $N \times N$ square distance matrix by calculating pair-wise distances among all N streamlines in the cluster/tract using the streamline based minimum direct-flip distance (MDF) metric [6]. We chose MDF distance metric as a solution to the inconsistent streamline orientation problem. The distance matrix is fed to nonlinear dimensionality reduction methods, i.e., t-SNE or

[1] Mapping high-dimensional data to a latent space refers to transforming complex forms of raw data into a simpler, lower-dimensional representation

UMAP, to project data into 2D space. In 2D space, spatially close streamlines are placed together and spurious streamlines are placed far from others. Hence, it becomes easier to visually and algorithmically filter out outlier clusters in the latent space. FiberNeat uses the density-based clustering method DBSCAN [14] to computationally label clusters in 2D space. It only keeps the streamlines of the largest clusters and removes small outlier clusters of streamlines. We use labels of streamlines to filter out the outlier streamlines in streamline-space. FiberNeat is an unsupervised algorithm that does not require any anatomical reference atlas or labeled training data. It is data-driven and takes the subject's own anatomy into consideration for filtering.

2. METHODS

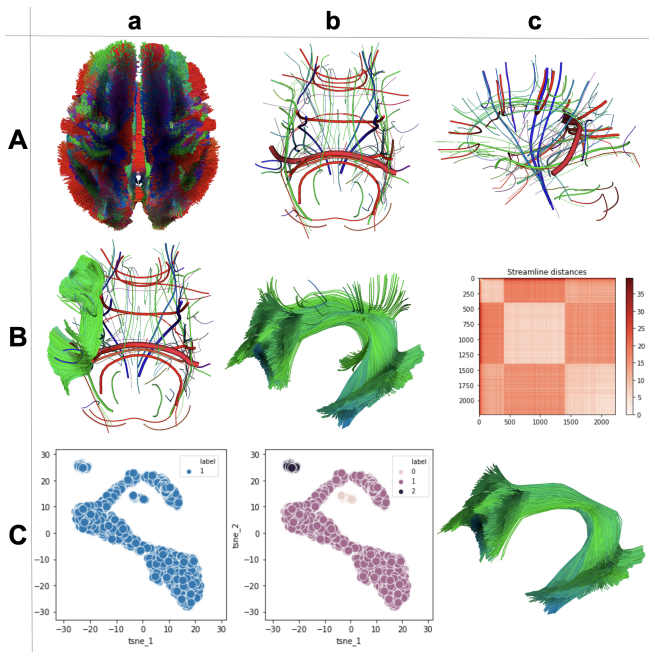


Fig. 1. Overview of the FiberNeat method. Panel A shows the preprocessing step of tractogram clustering. Panel B shows a cluster of streamlines (B.a and B.c) and visualization of their MDF distance matrix (B.c). Each streamline is mapped to a single 2D point using t-SNE (C.a) and clustered over the t-SNE embedding (C.b). Outlier streamlines are filtered out in the streamline (C.c).

Input to FiberNeat can be individual clusters from a whole-brain tractogram or extracted white matter tracts, where cluster/tract C is a set of N streamlines. $C = \{S_1, S_2, ..., S_n\}$, $S_i \in C$, $S_i = \{s_1, s_2, ..., s_n\}$, where s_i is a 3D vector point. The number of points per streamline may vary.

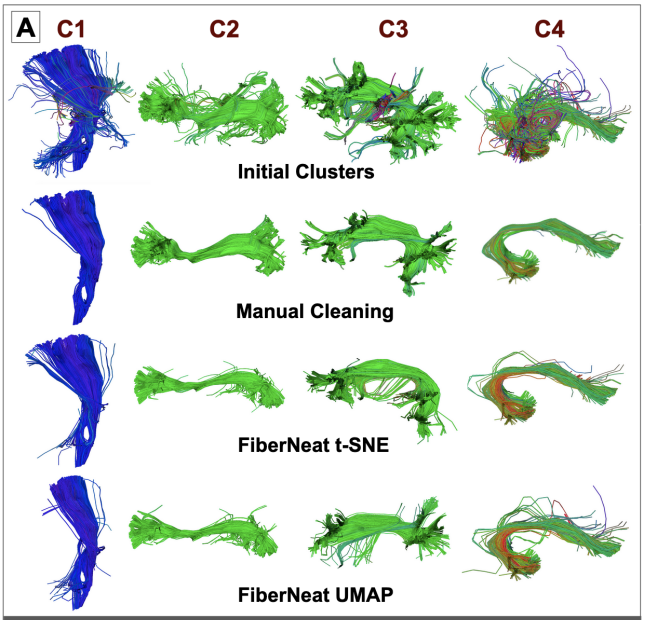
The FiberNeat method consists of the following steps:

1. Set all streamlines to have k number of points
2. Populate $N \times N$ distance matrix D by calculating pairwise MDF distance among all streamlines in the set C.
3. Project streamlines into 2D space using the precomputed streamline distances D.
   - Use either t-SNE or UMAP for the dimensionality reduction.
4. Cluster the streamlines in the 2D latent space using DBSCAN. Smaller clusters of 2D points are considered outliers. Streamlines belonging to the largest cluster in 2D space are kept in streamline space; streamlines belonging to the small clusters are removed.

Figure 1 illustrates steps of the FiberNeat method. A is a preprocessing step, where a whole-brain tractogram (A.a) is clustered using QuickBundles to obtain a sparse representation of the tractogram (A.b, A.c). B and C are the main steps of FiberNeat. We project individual clusters into lower dimensional space using t-SNE (B, C). We take an individual cluster of streamlines (B.a, B.c) and calculate pairwise streamline distances within that cluster (B.c) using the streamline-based MDF distance metric [6]. The MDF distance metric takes into account that streamlines traversing the brain in the same direction can be saved with opposite orientation. This step calculates a direct distance between two streamlines with their default orientation and a distance between a streamline and a streamline with a flipped orientation and selects the minimum of two. We provide t-SNE with this pre-calculated distance matrix as it embeds relevant information on similarities and differences between pairs of streamlines. As both t-SNE and UMAP are manifold learning approaches for non-linear dimensionality reduction, t-SNE could also be replaced by UMAP in this case. While the former captures and preserves local structure, the latter aims to preserve both local and global structure in the data. Streamlines are projected into 2D space by t-SNE (C.a) and the results are then clustered using the density-based clustering method, DBSCAN (C.b). This helps to visually and algorithmically locate outlier streamlines, as those tend to be placed and clustered together (C.b). Class 0 and 2 show outlier streamlines and are filtered out from the initial cluster (B.b) in streamline space (C.c). The entire process is completely unsupervised with no external information provided about anatomy. Visually, (C.c) agrees well with the expected trajectory of the arcuate fasciculus bundle in the left hemisphere of the brain.

3. RESULTS

We show results on data from a 26-30 year-old male HCP (the Human Connectome Project) [15] participant, scanned with 90 diffusion weighting directions and 6 b=0 acquisitions. Diffusion weighting consisted of 3 shells of b=1000, 2000, and 3000 s/mm2. Tractogram was generated using deterministic local tracking. In Figure 2A, we show results on four clusters selected from all clusters of whole-brain tractogram given by



B

Manual Cleaning	FiberNeat t-SNE		FiberNeat UMAP	
	SM	BMD	SM	BMD
C1	**0.942**	**0.537**	0.643	13.160
C2	**0.620**	**15.710**	0.600	17.370
C3	0.637	23.840	**0.719**	**9.485**
C4	**0.942**	**0.900**	0.932	0.915

Fig. 2. Part A, first row shows 4 initial clusters, the second row shows clusters manually cleaned by an expert. The third and fourth rows show clusters cleaned by FiberNeat t-SNE and FiberNeat UMAP, respectively. Part B shows the quantitative comparison of FiberNeat t-SNE and FiberNeat UMAP clusters with expert's cleaned clusters. Shape similarity score (SM) and bundle minimum distance (BMD) are calculated between clusters.

QuickBundles. The first row shows the initial four clusters. The second row shows clusters cleaned manually by a trained neuroanatomist, using visualization tools in DSI Studio [16] and DIPY [17]. We keep them as a ground truth to compare performance of FiberNeat t-SNE and FiberNeat UMAP. The third and fourth rows show clusters filtered using FiberNeat with t-SNE and UMAP embedding, respectively. In Figure 2B, a quantitative comparison of FiberNeat t-SNE and FiberNeat UMAP's filtered clusters with expert cluster cleaning is shown. Here, SM stands for shape similarity score [9] among two clusters and BMD stands for bundle-based minimum distance [18] between clusters. SM scores range from 0 to 1, where 0 implies least shape similarity between two clusters/tracts and 1 means highest shape similarity. BMD calculates streamline-based distance between clusters in mm. A lower value of BMD implies that two clusters are closer and more similar in shape and streamline count. FiberNeat t-SNE's filtered clusters have higher shape similarity and lower BMD distance with expert's cleaned clusters except for cluster C3. FiberNeat UMAP's output for C3 has higher shape similarity and lower BMD distance with an expert's cleaned cluster C3. Overall, qualitatively and quantitatively FiberNeat t-SNE performs better than FiberNeat UMAP.

4. DISCUSSION

Tractography data is unstructured complex data that is often linearly non-separable. It becomes extremely difficult to perfectly separate outliers from clusters corresponding to known anatomical tracts in the streamline space. t-SNE and UMAP are both manifold learning approaches for non-linear dimensionality reduction. Clustering based on t-SNE and UMAP embedding of tractography makes it easier to separate streamline clusters and outliers. Some researchers caution against clustering the t-SNE embedding space, at least for some applications, due to metric distortions. t-SNE a stochastic method and could generate different embeddings in different runs for the same data and parameters. It does not preserve the global metric structure and favors the preservation of the local structure only. t-SNE can sometimes disconnect/split parts of the data by putting them in separate clusters. This repelling effect of t-SNE is advantageous in our application as we want to untangle streamlines that are otherwise very closely knitted together in the original space, as seen in Figure 1C.a. The stochastic nature of t-SNE does not affect our approach as we do not use embedding again, for further data analytics. It is used once per input dataset and the method is invariant to where clusters are placed or what the global distance among clusters is. t-SNE does extreme dimensionality reduction by going directly to 2D space as opposed to other dimensionality reduction methods that provide options to project data into n>2 dimensions. But in our case, a streamline is $k \times 3$ D, and going to 2D is not an extreme dimensionality reduction. We also provide an option to use UMAP embedding instead of t-SNE. Theoretically, UMAP should give superior performance relative to t-SNE. UMAP tries to preserve both local and most of the global structure in the data. UMAP can map data to latent spaces with any number of dimensions and does not need the pre-dimensionality reduction step such as PCA or an autoencoder. Hence, UMAP can project data on n components and is not limited to 3D or 2D embeddings (as required by t-SNE). UMAP is computationally faster than t-SNE. However, in our experiments, we find t-SNE to outperform UMAP. This could be because the nature of the problem we are solving takes advantage of the data splitting/repelling property of t-SNE to find outlier streamlines in streamline sets that hard to distinguish in the original brain's 3D space. Further work is needed to evaluate the method on more tracts, and diverse datasets (in terms of age, diagnosis, and scanning protocol).

5. CONCLUSION

In this paper, we introduce FiberNeat, a method to clean streamline clusters and tracts. It takes clusters/tracts of streamlines as input and projects them into a latent space using dimensionality reduction techniques. A single streamline has a $k \times 3$ shape where k is the number of points on the streamline and k could vary per streamline. A streamline cluster or a tract contains N streamlines. It becomes difficult to find clusters of spurious streamlines in the streamline space. FiberNeat calculates the MDF distance among all streamlines of the cluster/tract and populates an $N \times N$ matrix. This precomputed streamline distance matrix is given to either t-SNE and UMAP to project streamline data into 2D space. In the 2D space, it becomes easier to detect outlier streamlines. FiberNeat applies DBSCAN clustering in 2D space. Smaller clusters are removed from the original data in the streamline space. We tried FiberNeat with t-SNE and UMAP on several same clusters of streamlines and found FiberNeat t-SNE to perform better than FiberNeat UMAP both qualitatively and quantitatively. FiberNeat can be used as a post-processing step in tract segmentation methods to aid better tractometry [9].

6. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by the Human Connectome Project [15]. Additional ethical approval was not required as confirmed by the license attached with the open access data.

7. ACKNOWLEDGMENTS

We would like to acknowledge that research reported in this publication was supported by the NIH (U.S. National Institutes of Health) under the AI4AD project grant U01 AG068057, grant number P41 EB015922, and the National Institute Of Biomedical Imaging And Bioengineering of the National Institutes of Health under Award Number R01EB027585.
Paul M Thompson received a research grant from Biogen, Inc. for research unrelated to this manuscript.

References

[1] P. J. Basser, J. Mattiello, and D. LeBihan, "MR diffusion tensor spectroscopy and imaging," *Biophysical journal*, vol. 66, no. 1, pp. 259–267, 1994.

[2] M. Catani and M. T. De Schotten, "A diffusion tensor imaging tractography atlas for virtual in vivo dissections," *Cortex*, vol. 44, no. 8, pp. 1105–1132, 2008.

[3] M.-A. Côté, G. Girard, A. Boré, E. Garyfallidis, J.-C. Houde, and M. Descoteaux, "Tractometer: Towards validation of tractography pipelines," *Medical image analysis*, vol. 17, no. 7, pp. 844–857, 2013.

[4] D. K. Jones, T. R. Knösche, and R. Turner, "White matter integrity, fiber count, and other fallacies: The do's and don'ts of diffusion MRI," *NeuroImage*, vol. 73, pp. 239–254, 2013.

[5] A. Brun, H. Knutsson, H.-J. Park, M. E. Shenton, and C.-F. Westin, "Clustering fiber traces using normalized cuts," in *MICCAI*, Springer, 2004, pp. 368–375.

[6] E. Garyfallidis, M. Brett, M. M. Correia, G. B. Williams, and I. Nimmo-Smith, "Quickbundles, a method for tractography simplification," *Frontiers in neuroscience*, vol. 6, p. 175, 2012.

[7] K. Oishi, K. Zilles, K. Amunts, A. Faria, H. Jiang, X. Li, K. Akhter, K. Hua, R. Woods, A. W. Toga, *et al.*, "Human brain white matter atlas: Identification and assignment of common anatomical structures in superficial white matter," *NeuroImage*, vol. 43, no. 3, pp. 447–457, 2008.

[8] E. Garyfallidis, M.-A. Côté, F. Rheault, J. Sidhu, J. Hau, L. Petit, D. Fortin, S. Cunanne, and M. Descoteaux, "Recognition of white matter bundles using local and global streamline-based registration and clustering," *NeuroImage*, vol. 170, pp. 283–295, 2018.

[9] B. Q. Chandio, S. L. Risacher, F. Pestilli, D. Bullock, F.-C. Yeh, S. Koudoro, A. Rokem, J. Harezlak, and E. Garyfallidis, "Bundle analytics, a computational framework for investigating the shapes and profiles of brain pathways across populations," *Scientific reports*, vol. 10, no. 1, pp. 1–18, 2020.

[10] V. Gupta, S. I. Thomopoulos, F. M. Rashid, and P. M. Thompson, "Fibernet: An ensemble deep learning framework for clustering white matter fibers," in *MICCAI*, Springer, 2017, pp. 548–555.

[11] K. G. Schilling, F. Rheault, L. Petit, C. B. Hansen, V. Nath, F.-C. Yeh, G. Girard, M. Barakovic, J. Rafael-Patino, T. Yu, *et al.*, "Tractography dissection variability: What happens when 42 groups dissect 14 white matter bundles on the same dataset?" *NeuroImage*, vol. 243, p. 118 502, 2021.

[12] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[13] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[14] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *kdd*, vol. 96, 1996, pp. 226–231.

[15] D. C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. E. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. W. Curtiss, *et al.*, "The Human Connectome Project: A data acquisition perspective," *NeuroImage*, vol. 62, no. 4, pp. 2222–2231, 2012.

[16] F.-C. Yeh, *DSI Studio*, version 2021 June, Jun. 2021. DOI: 10.5281/zenodo.4978980. [Online]. Available: https://doi.org/10.5281/zenodo.4978980.

[17] E. Garyfallidis, M. Brett, B. Amirbekian, A. Rokem, S. Van Der Walt, M. Descoteaux, and I. Nimmo-Smith, "Dipy, a library for the analysis of diffusion MRI data," *Frontiers in neuroinformatics*, vol. 8, p. 8, 2014.

[18] E. Garyfallidis, O. Ocegueda, D. Wassermann, and M. Descoteaux, "Robust and efficient linear registration of white-matter fascicles in the space of streamlines," *NeuroImage*, vol. 117, pp. 124–140, 2015.