Supplementary Materials to: Deep white matter analysis (DeepWMA): fast and consistent tractography segmentation

Fan Zhang[a], Suheyla Cetin Karayumak[a], Nico Hoffmann[a], Yogesh Rathi[a], Alexandra J. Golby[a], Lauren J. O'Donnell[a]

[a]*Brigham and Women's Hospital, Harvard Medical School, Boston, USA*

1. Supplementary Table 1: List of the 54 anatomical tracts in the training data

Supplementary Table 1: A total of 54 anatomical fiber tracts are included in our training tractography dataset, including 24 association tracts, 7 commissural tracts, 18 projection tracts, and 5 cerebellar tracts. For each fiber tract, the number of fiber samples and the proportion over all fibers in the 54 anatomical fiber tracts are provided.

Tract category (number of tracts)	Tract name (L - left; R - right; C - commissural)	number of fiber samples (proportion)
Association tracts (24)	arcuate fasciculus (AF) – LR	L - 8300 (3.04%); R - 3543 (1.30%)
	cingulum bundle (CB) – LR	L - 8530 (3.12%); R - 8233 (3.01%)
	external capsule (EC) – LR	L - 1225 (0.45%); R - 1648 (0.60%)
	extreme capsule (EmC) – LR	L - 2182 (0.80%); R - 1334 (0.49%)
	inferior longitudinal fasciculus (ILF) – LR	L - 8854 (3.24%); R - 9391 (3.44%)
	inferior occipito-frontal fasciculus (IoFF) – LR	L - 6479 (2.37%); R - 6053 (2.21%)
	middle longitudinal fasciculus (MdLF) – LR	L - 7741 (2.83%); R - 9590 (3.51%)
	posterior limb of internal capsule (PLIC) – LR	L - 1122 (0.41%); R - 1164 (0.43%)
	superior longitudinal fasciculus I (SLF I) – LR	L - 7572 (2.77%); R - 7375 (2.70%)
	superior longitudinal fasciculus II (SLF II) – LR	L - 5639 (2.06%); R - 9318 (3.41%)
	superior longitudinal fasciculus II (SLF III) – LR	L - 3246 (1.19%); R - 3536 (1.29%)
	uncinate fasciculus (UF) – LR	L - 2924 (1.07%); R - 3549 (1.30%)
Cerebellar tracts (5)	cortico-ponto-cerebellar (CPC) – LR	L - 859 (0.31%); R - 1019 (0.37%)
	inferior cerebellar peduncle (ICP) – LR	L - 1174 (0.43%); R - 1143 (0.42%)
	middle cerebellar peduncle (MCP) – C	C - 5329 (1.95%)
Commissural tracts (7)	corpus callosum 1 (CC 1) – C	C - 1604 (0.59%)
	corpus callosum 2 (CC 2) – C	C - 14431 (5.28%)
	corpus callosum 3 (CC 3) – C	C - 7938 (2.90%)
	corpus callosum 4 (CC 4) – C	C - 7186 (2.63%)
	corpus callosum 5 (CC 5) – C	C - 5817 (2.13%)
	corpus callosum 6 (CC 6) – C	C - 14502 (5.30%)
	corpus callosum 7 (CC 7) – C	C - 4742 (1.73%)
Projection tracts (18)	corticospinal tract (CST) – LR	L - 5045 (1.85%); R - 5986 (2.19%)
	corona-radiata-frontal (CR-F) – LR	L - 5094 (1.86%); R - 5662 (2.07%)
	corona-radiata-parietal (CR-P) – LR	L - 779 (0.28%); R - 1137 (0.42%)
	striato-frontal (SF) – LR	L - 9362 (3.42%); R - 7750 (2.83%)
	striato-occipital (SO) – LR	L - 791 (0.29%); R - 637 (0.23%)
	striato-parietal (SP) – LR	L - 924 (0.34%); R - 1011 (0.37%)
	thalamo-frontal (TF) – LR	L - 11675 (4.27%); R - 11465 (4.19%)
	thalamo-occipital (TO) – LR	L - 2257 (0.83%); R - 1496 (0.55%)
	thalamo-parietal (TP) – LR	L - 4307 (1.58%); R - 3709 (1.36%)

2. Supplementary Text I: Data preprocessing, whole-brain tractography, and compared tract segmentation methods

A. Data preprocessing

For the HCP dataset, we used the already processed diffusion MRI data (following the processing pipeline in (Glasser et al., 2013)). We extracted the b = 3000 shell of 90 gradient directions and all b = 0 scans for each subject, as applied in our previous studies (O'Donnell et al., 2017; Zhang et al., 2017b,a). Angular resolution is better and more accurate at high b-values such as 3000 (Descoteaux et al., 2007; Ning et al., 2015), and this single shell was chosen for reasonable computation time and memory use when performing tractography.

For the dHCP datasets, we used the already processed diffusion MRI data (following the processing pipeline in (Makropoulos et al., 2017)). Unlike the adult HCP dataset (where only high b = 3000 shell was used), we kept all the gradient directions based on our visual inspection that a high b shell tended to be very noisy in the dHCP dataset.

For the ABIDE, the CNP and the PPMI datasets, we pre-processed the provided raw imaging data using the following steps. DWIConvert[1] was first applied to convert the original data format (DICOM or NIFTI) to NRRD. Eddy current-induced distortion correction and motion correction were conducted using the Functional Magnetic Resonance Imaging of the Brain (FMRIB) Software Library tool (Jenkinson et al., 2012). To further correct for distortions caused by magnetic field inhomogeneity (which leads to intensity loss and voxel shifts), an EPI distortion correction was performed with reference to the T2-weighted image using the Advanced Normalization Tools (ANTS) (Avants et al., 2009). Because T2-weighted images were not available in all of these datasets (no T2 images were provided in the ABIDE and the CNP datasets, and in the PPMI dataset not all subjects had T2 images), we generated a synthetic T2-weighted image from a T1-weighted image for each subject (T1 images were available for in all datasets) using the T1 to T2 conversion toolbox[2]. For each subject, a nonlinear registration (registration was restricted to the phase encoding direction) was computed from the b0 image to the synthetic T2-weighted image to make an EPI corrective warp. Then, the warp was applied to each diffusion image. A semi-automated quality control (using in-house developed Matlab scripts) was conducted on all diffusion images. Individuals that had diffusion images with any apparent signal drops were excluded from the analyses. For the remaining subjects, all gradient directions were retained for analysis.

For the BTP dataset, we followed processing steps tuned specifically for brain tumor patient data, as reported in (O'Donnell et al., 2017). Diffusion images were corrected for motion and eddy current distortions using DTIPrep (Oguz et al., 2014). Images from all gradient directions were retained based on visual inspection of several patient datasets with an in-house tool indicating no gradients should be removed. Thus, all 30 gradient directions were retained for analysis.

[1] https://github.com/BRAINSia/BRAINSTools
[2] https://github.com/pnlbwh/T1toT2conversion

B. Whole-brain tractography

After obtaining the preprocessed DWI data, we applied the same UKF parameters for all subjects under study, as follows. Tractography was seeded in all voxels within the brain mask where fractional anisotropy (FA) was greater than 0.1. Tracking stopped where the FA value fell below 0.08 or the normalized mean signal (the sum of the normalized signal across all gradient directions) fell below 0.06. The normalized average signal measure was employed to robustly distinguish between white/gray matter and cerebrospinal fluid (CSF) regions. These seeding and stopping thresholds were set slightly below the default values to enable higher sensitivity for fiber tracking, in particular for the subjects (such as neonates, children and brain tumor patients) that might have low white matter anisotropy. Fibers that were longer than 40 mm were retained to avoid any bias towards implausible short fibers (Guevara et al., 2012; Jin et al., 2014; Lefranc et al., 2016).

The proposed DeepWMA method was also applied to the tractography data generated using another two fiber tracking methods different from the UKF method that was used for the atlas generation. The two additional tractography methods included a traditional second-order Runge-Kutta single-tensor DTI streamline method (Basser et al., 2000) (implemented in 3D Slicer[3] via SlicerDMRI[4] (Norton et al., 2017; Zhang et al., 2020)) and an advanced second-order integration over fiber orientation distributions (iFOD2) CSD method (Tournier et al., 2010) (implemented in MRtrix3[5]). Parameters for the two methods were set as follows. For the DTI method, tractography was seeded within the brain mask in all voxels where FA was greater than 0.1, and stopped when FA fell below 0.08. This generated approximately 100 thousand fibers per subject. For the CSD method, fiber orientation distribution (FOD) was computed using the default *tournier* algorithm for response function estimation (Tournier et al., 2007), and fiber tracking was performed using an iFOD2 method (Tournier et al., 2010) with the default parameters as suggested by the software (without any anatomical constraints). On average, approximately 1 million fibers were generated for each subject (similar to the average number of fibers in the UKF method).

C. Brief algorithmic details of compared tract segmentation methods

We evaluated the proposed DeepWMA method with comparison to WMA (a fiber-based tract segmentation method) and TractSeg (a voxel-based tract segmentation method). Parameters of each of these two comparison methods were set to their default values as suggested by the software. Additional details of each method follow.

The WMA method performed tract segmentation by leveraging the ORG fiber clustering atlas (Zhang et al., 2018). First, a tractography-based registration (O'Donnell et al., 2012) was performed to align subject-specific tractography data into the ORG atlas space. Then, a fiber spectral embedding (O'Donnell and Westin, 2007) was performed to identify fibers corresponding to the curated fibers in the ORG atlas for tract segmentation.

The TractSeg method performed tract segmentation by applying a CNN model, which was trained based on fiber orientation distribution function (fODF) peaks of 72 semi-automatically segmented reference tracts

[3] `https://www.slicer.org`
[4] `http://dmri.slicer.org`
[5] `http://www.mrtrix.org`

(Wasserthal et al., 2018, 2019). First, an FA-image-based rigid registration was conducted to align subject-specific DWI data into the reference space (MNI space). Then, the CNN model was applied to predict tract orientation maps (TOMs) that encoded the presence and orientations of the 72 tracts. Finally, fiber tracking was performed within each TOM following the predicted fiber orientations for tract segmentation.

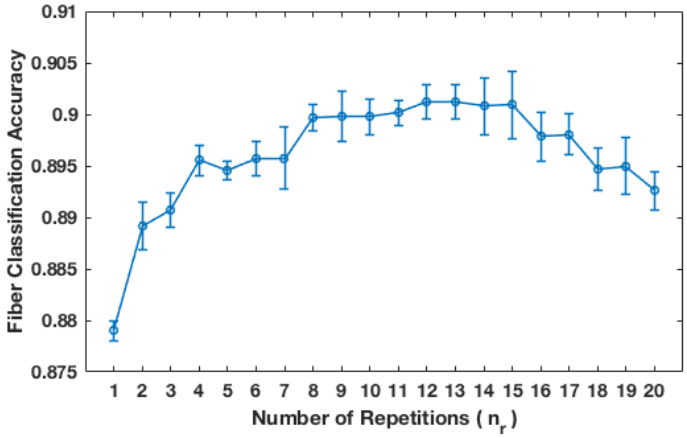
3. Supplementary Experiments

In this supplementary material, we provide three additional experiments. First, we performed two experiments to investigate the effects of two hyperparameters involved in our method, including the number of repetitions parameter (n_r) (Section 3.1) and the number of sampled points per fiber (n_p) (Section 3.2). These two experiments were performed on the training tractography data (the 1 million fibers with tract labels), where 80% of the fibers were used for model training and 20% for validation. The fiber classification accuracy on the validation data is reported. Then, we performed an experiment to investigate the effects of fiber length, i.e, if short or long fibers tended to be misclassified. To do so, we calculated the length distribution of the misclassified fibers in the validation data under the best-performing CNN model that generated the highest fiber classification accuracy of 90.99% (Section 3.3).

3.1. ***Effect of number of repetitions parameter (n_r)***

We tested the performance of fiber classification given different numbers of repetitions $n_r \in [1, 20]$ using the training tractography data. For each n_r value, we trained a CNN fiber classification model and computed the overall fiber classification accuracy (the percentage of fibers that were correctly classified into their ground truth tract category). This process was performed 10 times, and the mean $\pm$ the standard deviation of the accuracy were reported. We note that for higher n_r values from 8 to 20, the same CNN architecture was used (as described in the main paper content), while for the lower n_r values from 1 to 7, the number of convolutional layers and the corresponding ReLU, max pooling, and dropout layers was reduced to the maximum number of layers that was feasible. Padding was performed when $n_r = 1$ so that a 3×3 convolution kernel could be applied.

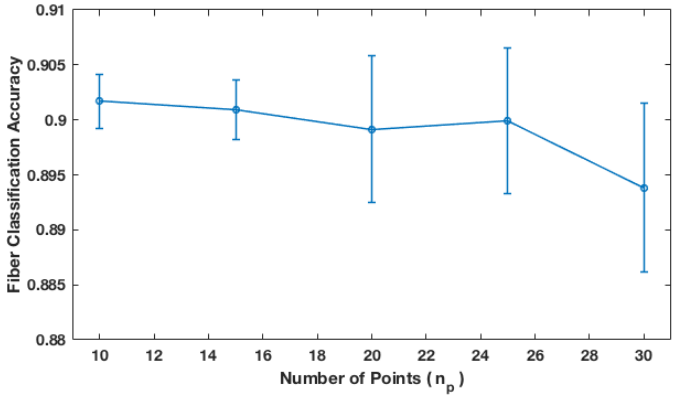
As shown in Supplementary Figure 1, the mean fiber classification accuracy increases as n_r increases from 1 to 7, and it is generally stable from $n_r = 8$ to $n_r = 15$. Then, the mean accuracy decreases as n_r keeps increasing to 20. The lower fiber classification accuracies when n_r is from 1 to 7 was likely due to a fewer number of convolution operations could be applied to the FiberMap descriptor. In general, n_r from 8 to 15 obtained the best mean fiber classification accuracy. We chose $n_r = 15$ in our study because it achieved the maximum accuracy across all testing runs in Figure 1 and it generated a square feature map for easy processing by CNNs. Additional CNN training runs were performed using $n_r = 15$ and the most successful model (with fiber classification accuracy 90.99%) was selected to be used in the following experiments.



Supplementary Figure 1: Effect of number of repetitions n_r. For each n_r value (from 1 to 20), 10 fiber classification CNN models were trained, and the mean and standard deviation of the fiber classification accuracy were reported.

3.2. ***Effect of sampled fiber points parameter*** *(n_p)*

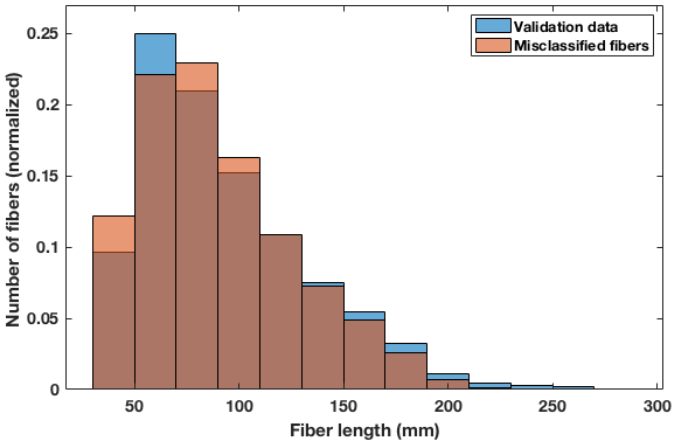
We tested the performance of fiber classification given different numbers of sampled points per fiber $n_p \in [10, 30]$. For each n_p, we set n_r to be the same as n_p, and we trained a CNN fiber classification model and computed the overall fiber classification accuracy (the percentage of fibers that were correctly classified into their ground truth tract category). This process was performed 10 times, and the mean $\pm$ the standard deviation of the accuracy were reported. As shown in Supplementary Figure 2, the mean fiber classification accuracies across the different tested n_p values are similar. Interestingly, sampling a higher number of points did not improve the fiber classification accuracy, potentially because a CNN with more layers would be needed to handle a larger amount of data. In the present study, we extract $n_p = 15$ points per fiber, a reasonable number for fiber representation that has been shown to be successful in our previous studies (Zhang et al., 2018; O'Donnell and Westin, 2007).



Supplementary Figure 2: Effect of number of sampled points per fiber n_p. For each n_p value (from 10 to 30), 10 fiber classification CNN models were trained, and the mean and standard deviation of the fiber classification accuracy are reported.

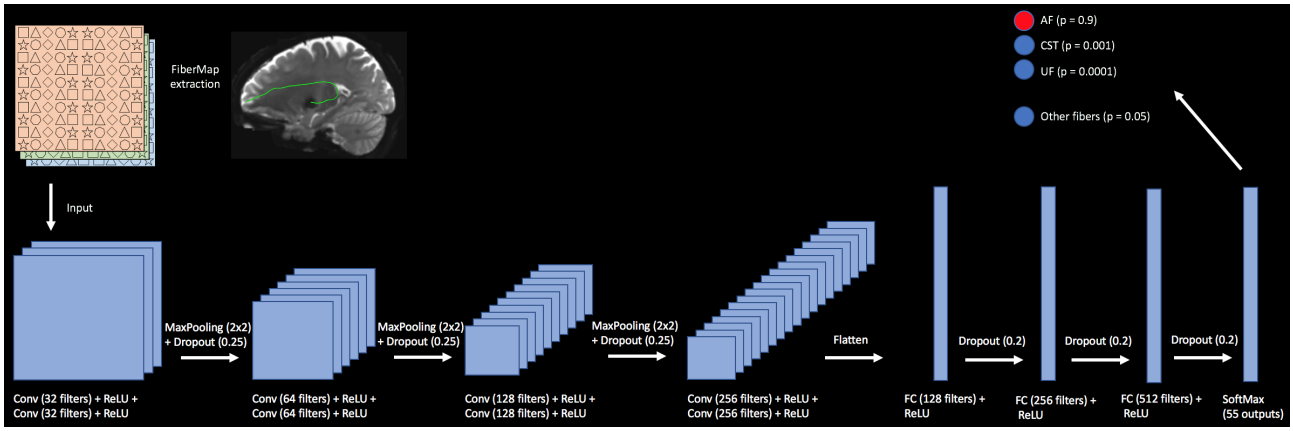
3.3. ***Fiber length distribution of the misclassified fibers in the training tractography data***

Supplementary Figure 3 gives the fiber length distribution of all fibers in the validation data and that of the misclassified fibers from DeepWMA. Overall, we can observe that the fiber length distributions are highly similar, suggesting that our DeepWMA method is robust to fibers with different lengths.



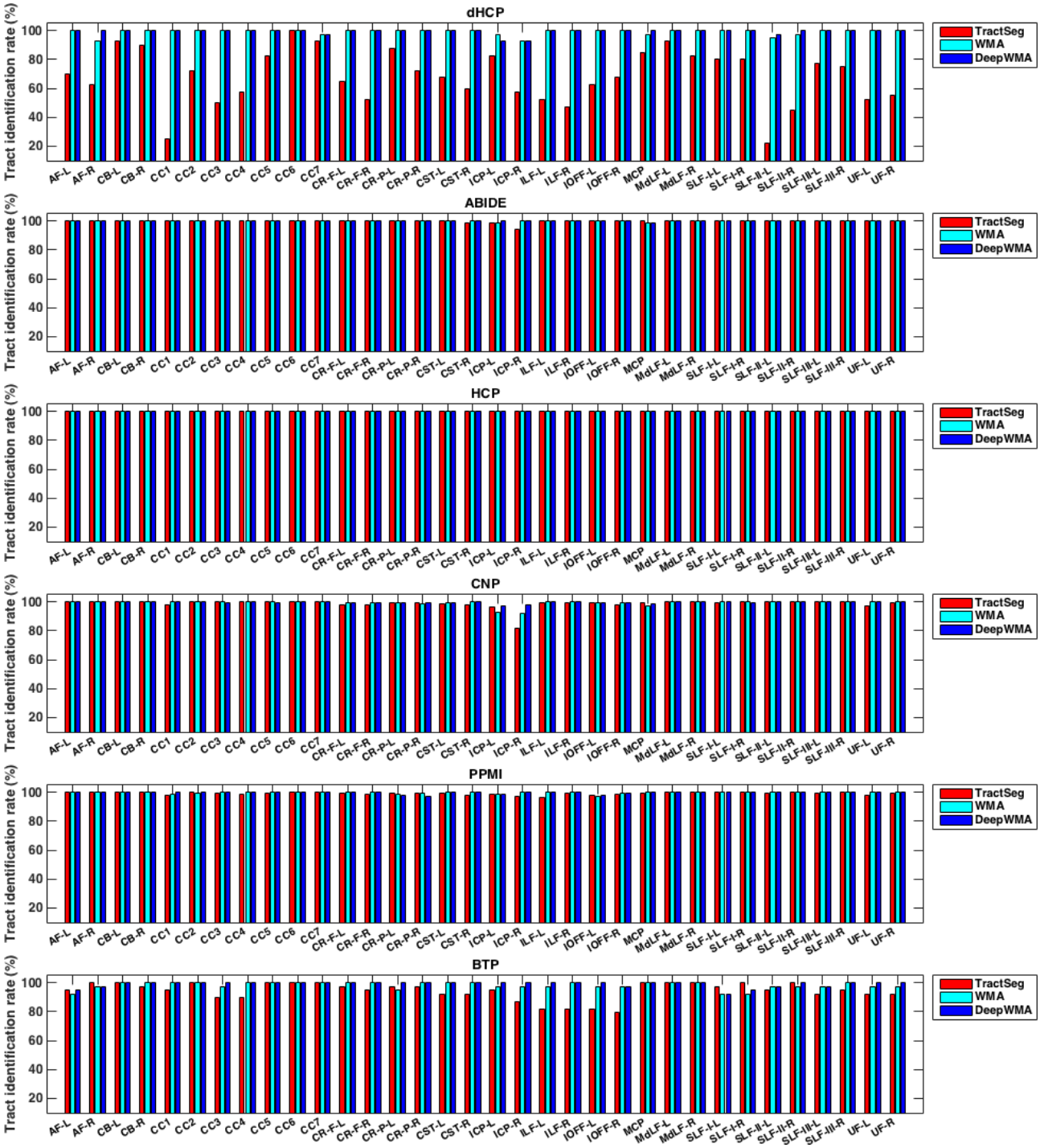
Supplementary Figure 3: Comparison of fiber length distribution of misclassified fibers and that of all fibers in the validation data. For each histogram, the values are normalized so that the sum of all bins is equal to 1.

4. Supplementary Figure 4: A detailed network architecture diagram



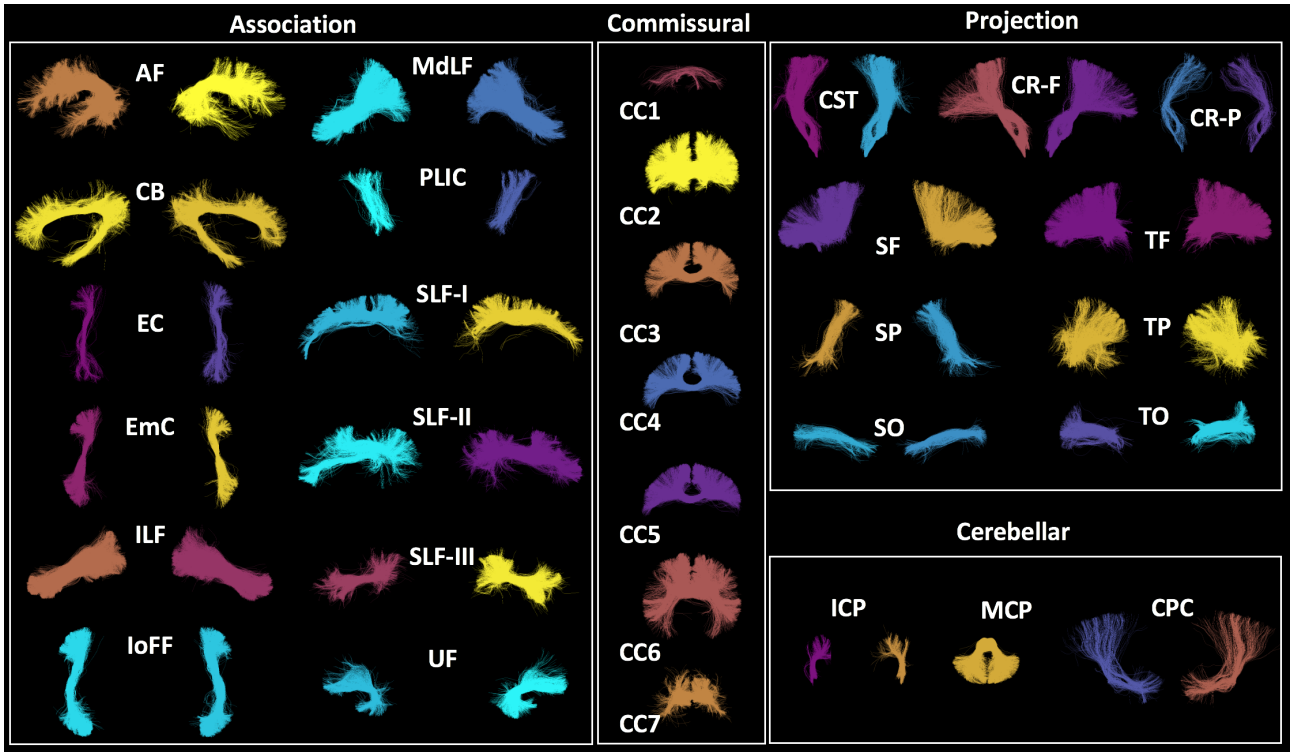
Supplementary Figure 4: Diagram of the network architecture. This network contains eight convolutional layers (32, 32, 64, 64, 128, 128, 256 and 256 filters, respectively) with kernel size 3, where each convolutional layer is followed by a rectified linear units (ReLU) activation layer. A max pooling layer of size 2 and a dropout layer of 0.25 are used after every two ReLU layers to prevent overfitting. The last convolutional layer is followed by three fully connected layers of size 128, 256 and 512, where each fully connected layer is followed by a dropout layer of 0.2. Then, a softmax layer is used with 55 outputs for the 54 anatomical tracts and the category of "other fibers." The convolution layers with 64, 128, and 256 filters and the corresponding ReLU, max pooling, and Dropout layers are important for constructing a deeper network, while increasing random sampling of the output after the first max pooling layer, where each row is the same.

5. Supplementary Figure 5: Tract identification rate of each individual tract for each tract segmentation method.



Supplementary Figure 5: Comparison of the tract identification rate for each individual tract across the three tract segmentation methods. A total of 34 tracts that are commonly defined across the three compared methods are included.

6. Supplementary Figure 6: Visualization of all 54 anatomical fiber tracts for one HCP subject

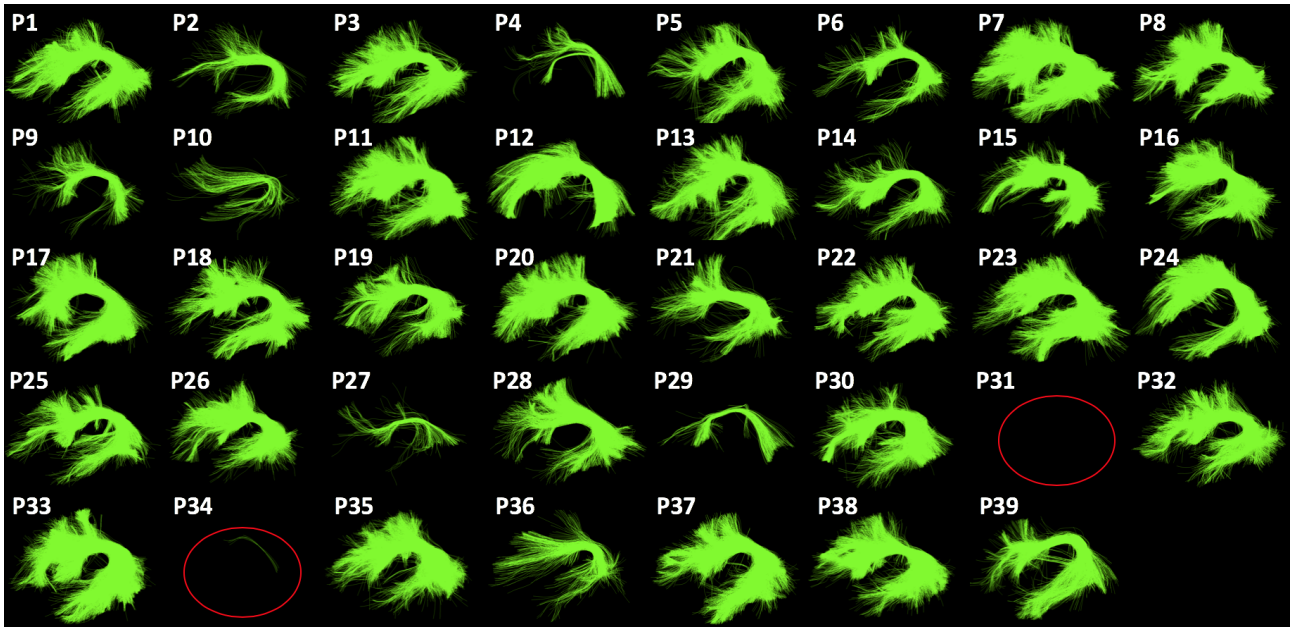


Supplementary Figure 6: Visualization of all 54 anatomical fiber tracts computed using DeepWMA from tractography data of one example HCP subject.

7. Supplementary Figure 7: Segmented left AF tracts of all subjects in dHCP.

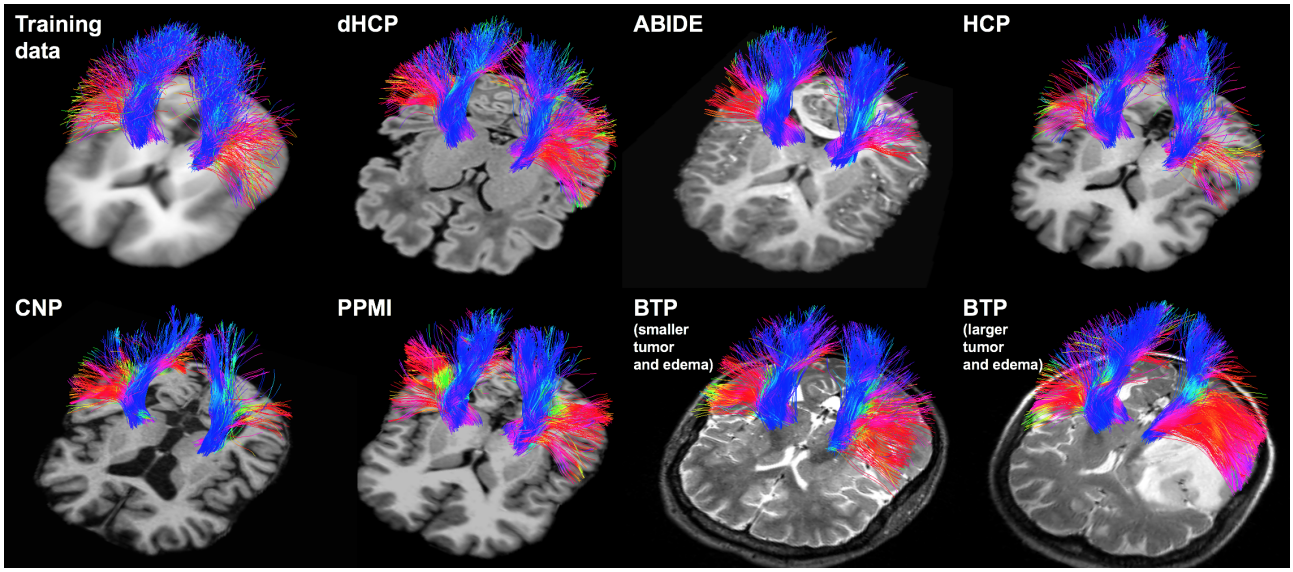
Supplementary Figure 7: Visualization of the segmented left AF tracts of all 40 neonates in the dHCP dataset computed using DeepWMA. The left AF tracts were successfully detected in all subjects.

8. Supplementary Figure 8: Segmented left AF tracts of all subjects in BTP.



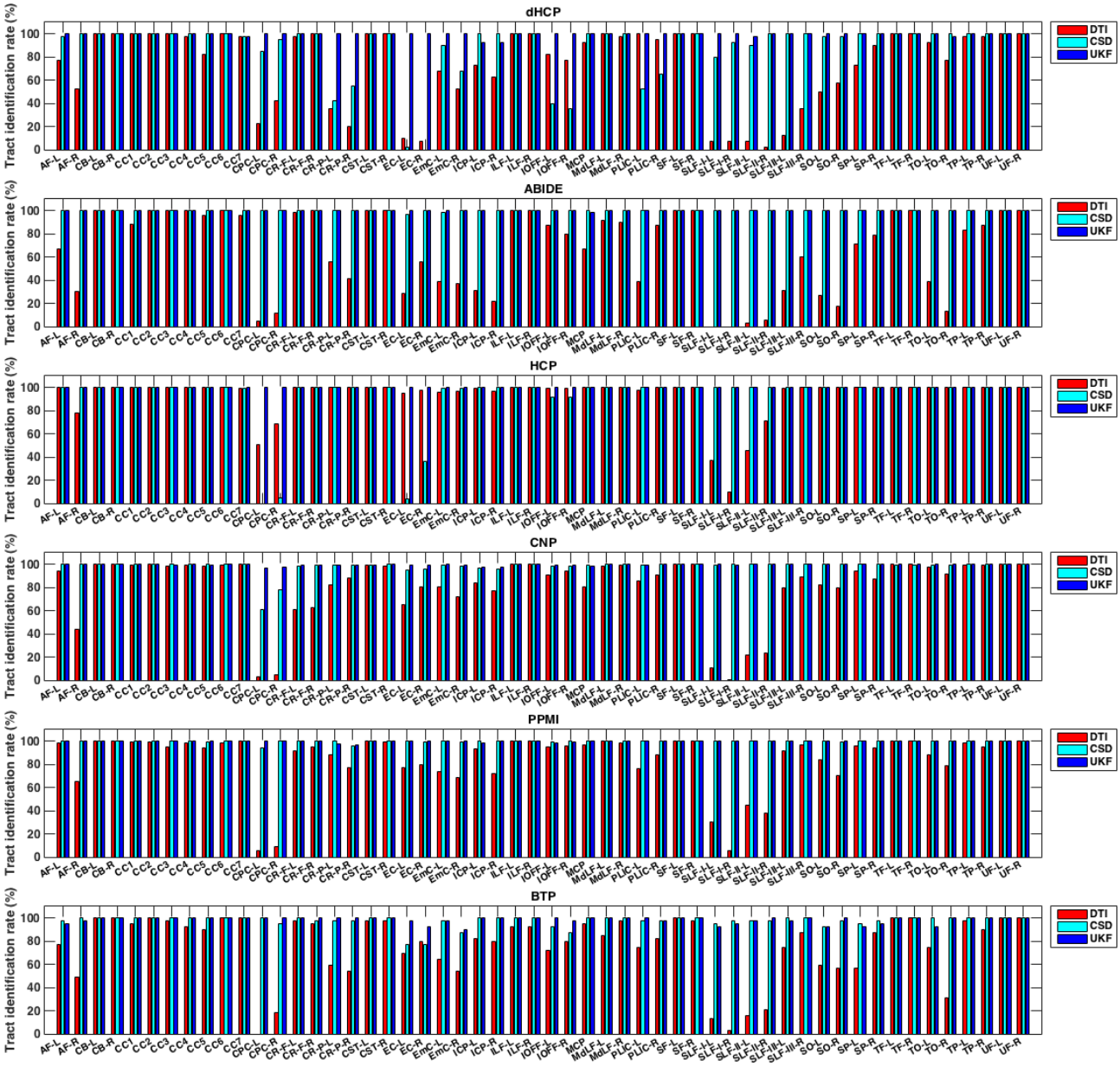
Supplementary Figure 8: Visualization of the segmented left AF tracts of all 39 neurosurgical patients in the BTP dataset computed using DeepWMA. The left AF tracts were successfully detected in all patients except for patients P31 and P34. We note that neither of the two compared methods, i.e. WMA and TractSeg, could detect the left AF tracts from the data of these two patients.

9. Supplementary Figure 9: Visualization of CST segmented using DeepWMA overlaid on anatomical MRI images



Supplementary Figure 9: Visualization of the CST fibers in the training tractography data, overlaid on the atlas population mean T1w image, and visualization of the subject-specific CSTs segmented using the proposed DeepWMA method, overlaid on the T1w or T2w images of the subject. The same subjects as shown in Figures 5 and 6 in the main paper content are selected for visualization. Overall, the subject-specific CSTs pass through the same brain regions as the CST in the training tractography data – starting from the midbrain, passing through the posterior limb of the internal capsule (as shown in the figure), and then connecting to primary motor, primary somatosensory, and dorsal premotor cortices, plus the supplementary motor area. This pathway corresponds to the known anatomy of the CST (Seo and Jang, 2013).

10. Supplementary Figure 10: DeepWMA tract identification rate of each individual tract for each fiber tracking method.



Supplementary Figure 10: DeepWMA tract identification rate for each individual tract in each of three fiber tracking methods. A total of 54 tracts provided in the training data are included.

References

Avants, B. B., Tustison, N., Song, G., 2009. Advanced normalization tools (ANTS). Insight j 2, 1–35.

Basser, P. J., Pajevic, S., Pierpaoli, C., Duda, J., Aldroubi, A., 2000. In vivo fiber tractography using DT-MRI data. Mag. Res. Med. 44 (4), 625–632.

Descoteaux, M., Angelino, E., et al., 2007. Regularized, fast, and robust analytical Q-ball imaging. Mag. Res. Med. 58 (3), 497–510.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Essen, D. C. V., Jenkinson, M., 2013. The minimal preprocessing pipelines for the Human Connectome Project. NeuroImage 80, 105–124.

Guevara, P., Duclap, D., Poupon, C., Marrakchi-Kacem, L., Fillard, P., Le Bihan, D., Leboyer, M., Houenou, J., Mangin, J.-F., 2012. Automatic fiber bundle segmentation in massive tractography datasets using a multi-subject bundle atlas. NeuroImage 61 (4), 1083–1099.

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., Smith, S. M., 2012. FSL. NeuroImage 62 (2), 782–790.

Jin, Y., Shi, Y., Zhan, L., Gutman, B. A., de Zubicaray, G. I., McMahon, K. L., Wright, M. J., Toga, A. W., Thompson, P. M., 2014. Automatic clustering of white matter fibers in brain diffusion MRI with an application to genetics. NeuroImage 100, 75–90.

Lefranc, S., Roca, P., Perrot, M., Poupon, C., Le Bihan, D., Mangin, J.-F., Rivière, D., 2016. Groupwise connectivity-based parcellation of the whole human cortical surface using watershed-driven dimension reduction. Medical Image Analysis 30, 11–29.

Makropoulos, A., Robinson, E. C., Schuh, A., Wright, R., Fitzgibbon, S., Bozek, J., Counsell, S. J., Steinweg, J., Passerat-Palmbach, J., Lenz, G., Mortari, F., Tenev, T., Duff, E. P., Bastiani, M., Cordero-Grande, L., Hughes, E., Tusor, N., Tournier, J.-D., Hutter, J., Price, A. N., Murgasova, M., Kelly, C., Rutherford, M. A., Smith, S. M., Edwards, A. D., Hajnal, J. V., Jenkinson, M., Rueckert, D., 2017. The Developing Human Connectome Project: a minimal processing pipeline for neonatal cortical surface reconstruction. bioRxiv, 125526.

Ning, L., Laun, F., Gur, Y., DiBella, E. V., Deslauriers-Gauthier, S., Megherbi, T., Ghosh, A., Zucchelli, M., Menegaz, G., Fick, R., St-Jean, S., Paquette, M., Aranda, R., Descoteaux, M., Deriche, R., O'Donnell, L. J., Rathi, Y., 2015. Sparse reconstruction challenge for diffusion MRI: Validation on a physical phantom to determine which acquisition scheme and analysis method to use? Medical Image Analysis 26 (1), 316–331.

Norton, I., Essayed, W. I., Zhang, F., Pujol, S., Yarmarkovich, A., Golby, A. J., Kindlmann, G., Wasserman, D., Estepar, R. S. J., Rathi, Y., Pieper, S., Kikinis, R., Johnson, H. J., Westin, C.-F., O'Donnell, L. J., 2017. SlicerDMRI: Open source diffusion MRI software for brain cancer research. Cancer Research 77 (21), e101–e103.

O'Donnell, L. J., Suter, Y., Rigolo, L., Kahali, P., Zhang, F., Norton, I., Albi, A., Olubiyi, O., Meola, A., Essayed, W. I., Unadkat, P., Ciris, P. A., Wells III, W. M., Rathi, Y., Westin, C.-F., Golby, A. J., 2017. Automated white matter fiber tract identification in patients with brain tumors. NeuroImage: Clinical 13, 138–153.

O'Donnell, L. J., Wells III, W. M., Golby, A. J., Westin, C.-F., 2012. Unbiased groupwise registration of white matter tractography. In: MICCAI. pp. 123–130.

O'Donnell, L. J., Westin, C.-F., 2007. Automatic tractography segmentation using a high-dimensional white matter atlas. IEEE Trans. Med. Img. 26 (11), 1562–1575.

Oguz, I., Farzinfar, M., Matsui, J., Budin, F., Liu, Z., Gerig, G., Johnson, H., Styner, M., 2014. DTIPrep: quality control of diffusion-weighted images. Neuroinformatics 8, 4.

Seo, J., Jang, S., 2013. Different characteristics of the corticospinal tract according to the cerebral origin: DTI study. American Journal of Neuroradiology 34 (7), 1359–1363.

Tournier, J.-D., Calamante, F., Connelly, A., 2007. Robust determination of the fibre orientation distribution in diffusion MRI: non-negativity constrained super-resolved spherical deconvolution. NeuroImage 35 (4), 1459–1472.

Tournier, J. D., Calamante, F., Connelly, A., 2010. Improved probabilistic streamlines tractography by 2nd order integration over fibre orientation distributions. In: ISMRM. p. 1670.

Wasserthal, J., Neher, P., Hirjak, D., Maier-Hein, K. H., 2019. Combined tract segmentation and orientation mapping for bundle-specific tractography. arXiv:1901.10271.

Wasserthal, J., Neher, P., Maier-Hein, K. H., 2018. TractSeg - fast and accurate white matter tract segmentation. NeuroImage 183, 239–253.

Zhang, F., Kahali, P., Suter, Y., Norton, I., Rigolo, L., Savadjiev, P., Song, Y., Rathi, Y., Cai, W., Wells III, W. M., Golby, A. J., O'Donnell, L. J., 2017a. Automated connectivity-based groupwise cortical atlas generation: Application to data of neurosurgical patients with brain tumors for cortical parcellation prediction. In: IEEE International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 774–777.

Zhang, F., Noh, T., Juvekar, P., Frisken, S. F., Rigolo, L., Norton, I., Kapur, T., Pujol, S., Wells, W., Yarmarkovich, A., Kindlmann, G., Wassermann, D., San Jose Estepar, R., Rathi, Y., Kikinis, R., Johnson, H. J., Westin, C.-F., Pieper, S., Golby, A. J., O'Donnell, L. J., 2020. SlicerDMRI: Diffusion MRI and Tractography Research Software for Brain Cancer Surgery Planning and Visualization. JCO Clinical Cancer Informatics 4, 299–309.

Zhang, F., Norton, I., Cai, W., Song, Y., Wells III, W. M., O'Donnell, L. J., 2017b. Comparison between two white matter segmentation strategies: an investigation into white matter segmentation consistency. In: IEEE International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 796–799.

Zhang, F., Wu, Y., Norton, I., Rigolo, L., Rathi, Y., Makris, N., O'Donnell, L. J., 2018. An anatomically curated fiber clustering white matter atlas for consistent white matter tract parcellation across the lifespan. NeuroImage 179, 429–447.