Medical Image Analysis 58 (2019) 101559

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Combined tract segmentation and orientation mapping for bundle-specific tractography

Jakob Wasserthal[a,b], Peter F. Neher[a], Dusan Hirjak[d], Klaus H. Maier-Hein[a,c,*]

[a] Division of Medical Image Computing (MIC), German Cancer Research Center (DKFZ), Heidelberg, Germany
[b] Medical Faculty Heidelberg, University of Heidelberg, Heidelberg, Germany
[c] Section of Automated Image Analysis, Heidelberg University Hospital, Heidelberg, Germany
[d] Department of Psychiatry and Psychotherapy, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

ARTICLE INFO

Article history:
Received 28 January 2019
Revised 9 September 2019
Accepted 11 September 2019
Available online 12 September 2019

Keywords:
Machine learning
Diffusion-weighted imaging
Fiber tractography
Deep learning

ABSTRACT

While the major white matter tracts are of great interest to numerous studies in neuroscience and medicine, their manual dissection in larger cohorts from diffusion MRI tractograms is time-consuming, requires expert knowledge and is hard to reproduce. In previous work we presented tract orientation mapping (TOM) as a novel concept for bundle-specific tractography. It is based on a learned mapping from the original fiber orientation distribution function (FOD) peaks to tract specific peaks, called tract orientation maps. Each tract orientation map represents the voxel-wise principal orientation of one tract. Here, we present an extension of this approach that combines TOM with accurate segmentations of the tract outline and its start and end region. We also introduce a custom probabilistic tracking algorithm that samples from a Gaussian distribution with fixed standard deviation centered on each peak thus enabling more complete trackings on the tract orientation maps than deterministic tracking. These extensions enable the automatic creation of bundle-specific tractograms with previously unseen accuracy.

We show for 72 different bundles on high quality, low quality and phantom data that our approach runs faster and produces more accurate bundle-specific tractograms than 7 state of the art benchmark methods while avoiding cumbersome processing steps like whole brain tractography, non-linear registration, clustering or manual dissection. Moreover, we show on 17 datasets that our approach generalizes well to datasets acquired with different scanners and settings as well as with pathologies. The code of our method is openly available at https://github.com/MIC-DKFZ/TractSeg.

1. Introduction

The white matter of the human brain is made up of a large number of individual fiber tracts. Those tracts overlap, resulting in multiple fiber orientation distribution function (FOD) peaks per voxel and larger bottleneck situations with tracts per voxel outnumbering the peaks per voxel. In consequence, tractography is highly susceptible to false positives (Maier-Hein et al., 2017; Knösche et al., 2015). The only safe solution around false positives so far is the explicit dissection of anatomically well-known tracts. While manual dissection protocols (Stieltjes et al., 2013) can be considered the current gold standard, a variety of approaches has already been developed for automating the process: *Region-of-interest-based* approaches filter streamlines based on their spatial relation to cortical or other anatomically defined regions, which are typically transferred to subject space via atlas registration and segmentation techniques (Wassermann et al., 2016; Yendiki et al., 2011). *Clustering-based* approaches group and select streamlines by measuring intra- and inter-subject streamline similarities, referring to existing reference tracts in atlas space (Garyfallidis et al., 2017; O'Donnell and Westin, 2007; O'Donnell et al., 2016). Concept-wise, many previous approaches have opted for performing a rather blind whole brain tractography and then investing the effort in streamline space, clearing the tractograms from spurious streamlines and grouping the remaining ones. These approaches often have long runtimes, need several processing steps which are tedious to set up, depend on registration which is error prone and have accuracy which is decent but still leaves room for improvement.

* Corresponding author.
E-mail addresses: j.wasserthal@dkfz.de (J. Wasserthal), k.maier-hein@dkfz.de (K.H. Maier-Hein).

https://doi.org/10.1016/j.media.2019.101559

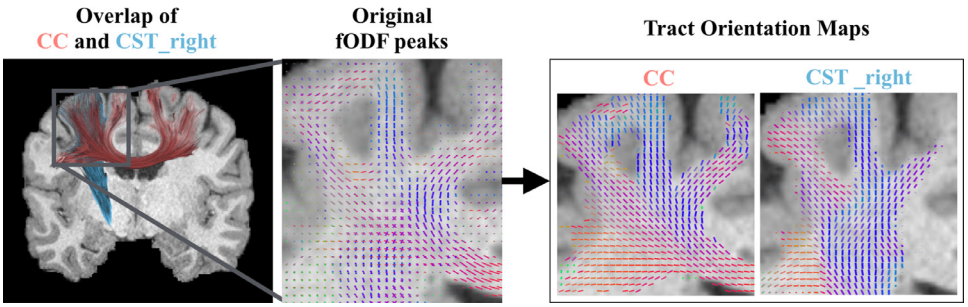

Fig. 1. Exemplary depiction of a slice through two of the reference tracts, the original FOD peak image and the corresponding reference TOMs (CST_right: corticospinal tract; CC: corpus callosum).

In Wasserthal et al. (2018a) we presented a novel concept called tract orientation mapping (TOM) that approaches the problem before doing tractography by learning tract-specific peak images (tract orientation maps, also abbreviated TOM). Each TOM represents one tract, and each voxel contains one orientation vector representing the local tract orientation, i.e. the local mean streamline orientation of the tract (see Fig. 1). These tract orientation maps can then be used as a prior – similar to Rheault et al. (2018), who employed registered atlas information as a tract-specific prior – or directly as orientation field for tractography. In Wasserthal et al. (2018b) we presented a novel method called TractSeg for fast and accurate tract segmentation. Based on these preliminary works we introduce an comprehensive approach to bundle-specific tractography:

On low resolution data, TOM tends to oversegment the individual tracts. In contrast to the complex task of voxel-wise peak regression with TOM, the simpler binary segmentation with TractSeg yields more accurate tract delineations. Therefore we use the segmentation results from TractSeg to filter the TOM tractograms. After filtering with the TractSeg segmentations the tractograms show good spatial extent and orientation. However, a lot of streamlines are still ending prematurely. Filtering the streamlines by a gray matter segmentation is not sufficient, as tracts tend to touch gray matter regions but are not supposed to end there. To obtain proper segmentations of the regions where each tract starts and ends, we trained another convolutional neural network using the same approach as TractSeg (Wasserthal et al., 2018b). Now all streamlines not ending in the start/end regions can easily be removed. Using the steps described so far, highly accurate bundle-specific tractograms can be obtained in most situations. However, in some cases a simple deterministic tracking of the TOM peaks yields suboptimal results, for example due to low image resolution. Therefore we propose a probabilistic approach to TOM tractography which maximizes the sensitivity of the proposed bundle specific tractography pipeline, even on low resolution data or strongly bent tracts. An overview of the entire pipeline can be seen in Fig. 2.

For training and the first part of the evaluation we use the dataset provided by Wasserthal et al. (2018b) containing 105 subjects from the Human Connectome Project (HCP). For the second part of our evaluation we use 17 differently acquired datasets to evaluate how good our approach generalizes to other datasets. We compare our method to seven other state-of-the-art methods for generating bundle-specific tractograms. We show that our approach is easy to set up, fast to run and does not require affine or elastic registration, parcellation or clustering.

In comparison to our previous works (Wasserthal et al., 2018b; 2018a) this paper adds the following contributions: Segmentation of the tract start/end region, combination of tract segmentation, start/end region segmentation and tract orientation maps, more sensitive custom tracking algorithm which is optimized for this approach and extended evaluation on more bundles, phantom data and 17 non-HCP datasets with and without pathologies.

A short note on the terminology we use: When talking about *fibers* or *streamlines* we are referring to the single streamlines. When talking about *bundles* or *tracts* we are referring to a group of streamlines making up an anatomical structure (e.g. the corticospinal tract).

2. Materials and methods

All three methods (tract segmentation, start/end region segmentation and tract orientation mapping) are based on the same fully convolutional neural network architecture (U-Net Ronneberger et al., 2015) that receives as input the fiber orientation distribution function (FOD) peaks. What differs is the training target the network has to learn. For tract segmentation, the network is performing voxel-wise binary classification to discern tract and non-tract voxels. For tract start/end region segmentation it is also doing binary classification but now the number of classes has doubled because for each tract one start and one end region is learned. For tract orientation mapping the network regresses a single 3D peak vector, i.e. three float values, per voxel and bundle. In this way, the models used for the three methods only differ in the number of output channels, the final activation function and the loss function.

2.1. Preprocessing

While we successfully tested raw diffusion weighted images as input for our method, this would have restricted the method to the MRI acquisition used during training, not allowing for any variation in the acquisition, such as a change of b-value or the number of gradient directions, without a complete retraining of the model. Moreover, for high angular resolution datasets, it would have resulted in an input image with an accordingly large number of channels (one channel for each gradient orientation and each b-value), resulting in unfeasible high memory demand and slow file input/output during training. A more condensed representation of the data was chosen to mitigate this problem: The network expects to receive the three principal fiber directions per voxel as input, thus requiring nine different input channels (three per principal direction). In this study, the principal directions were estimated from the diffusion data using the multi-shell multi-tissue constrained spherical deconvolution (CSD) and peak extraction available in MRtrix (Jeurissen et al., 2014; Tournier et al., 2007) with a maximum number of three peaks per voxel. If a voxel contained only one fiber direction e.g. voxels in the corpus callosum, then the second and third peak are set to zero. Another possible input instead of peaks would have been FA maps. However, this resulted in worse results (see supplementary materials). The HCP images have

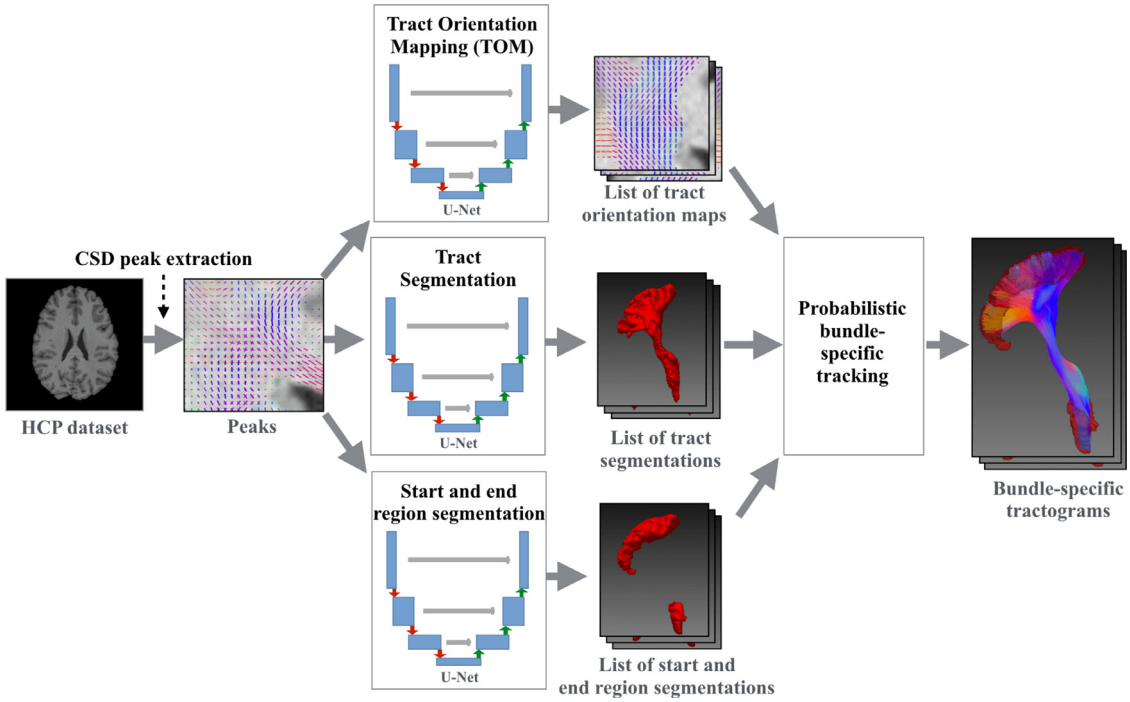

Fig. 2. Pipeline overview: Constrained spherical deconvolution (CSD) is applied to obtain the three principal FOD directions per voxel which is the input to three U-Nets. The three U-Nets are used to create a tract orientation map, a tract mask and a start/end region mask for each tract. Then probabilistic tracking is run on the tract orientation maps. All streamlines leaving the tract mask and not ending in the start/end masks are discarded. The result is one tractogram for each tract.

a spatial resolution of $145 \times 174 \times 145$ voxels. We cropped them to $144 \times 144 \times 144$ without removing any brain tissue to make them fit to our network input size.

2.2. Model

2.2.1. Architecture

The proposed 2D encoder-decoder architecture was inspired by the U-Net architecture previously proposed by Ronneberger et al. (2015). To enable better flow of the error gradients during backpropagation we added deep supervision (Isensee et al., 2018). This reduced the training time and slightly improved the results. A figure of the network architecture can be found in the supplementary materials.

The input for the proposed network is a 2D image at 144×144 voxels and 9 channels corresponding to the 3 peaks per voxel (each 3D peak is represented by three float values). The output is a multi-channel image with spatial dimensions of 144×144 voxels, where each channel contains the voxel-wise results for one tract. For tract segmentation this leads to 72 channels and for start/end region segmentation to $72 \cdot 2 = 144$ channels. Following the same approach $72 \cdot 3 = 216$ channels would have been needed for tract orientation mapping. However, given such a high number of classes the training did not converge anymore. To deal with this issue we chose to only train for 18 tracts (=54 channels) at the same time. So four models had to be trained to cover all 72 tracts.

For the segmentation tasks the networks output a probability between 0 and 1. These probabilities are converted to binary segmentations by thresholding at 0.5. For the tract orientation mapping the networks output one peak per voxel and tract. Peaks shorter than 0.3 are discarded.

To avoid a downsized output in comparison to the input we padded with half the filter size (rounded down). Given a filter size of 3 the padding was set to 1. This is also referred to as SAME padding (Dumoulin and Visin, 2016).

2.2.2. Handling of 3D data

While in principle the U-Net architecture allows extensions to image segmentation with 3D convolutions (Cicek et al., 2016), we here propose a 2D architecture. Using a 3D U-Net, we did not achieve the same performance as when using a 2D U-Net (see supplementary materials). To still leverage the additional information provided by the third dimension, we randomly sampled 2D slices in three different orientations during training: axial, coronal and sagittal. This meant that our model learned to work with all three of these orientations. During inference three predictions per voxel per tract were generated, one for each orientation, resulting in an image with dimensions of $144 \times 144 \times 144 \times nr\_classes \times 3$ (after running our model $144 \cdot 3$ times). We use the mean to merge those three prediction to one final prediction. Running the model three times (once for each orientation) is only done for tract segmentation and start/end region segmentation. For tract orientation mapping it slightly worsened the results (+0.18 angular error degrees on the HCP Quality dataset). Therefore we only run the model once for all slices along the y-axis. The y-axis corresponds to the coronal axis and as shown in Wasserthal et al. (2018b), the y-axis gives the best results when only using one axis.

2.3. Training

2.3.1. Loss

For the segmentation models we trained our network using the binary cross-entropy loss. Sigmoid activation functions were used in the last layer. For a given target y, an output of the network $\hat{y}$ and N number of classes, the loss is calculated as follows:

$$\text{loss}(\hat{y}, y) = -\frac{1}{N}\sum_{i=0}^{N}(y[i]log(\hat{y}[i]) + (1 - y[i])\log(1 - \hat{y}[i])] \qquad (1)$$

For the tract orientation mapping model the network was trained using cosine similarity as loss. Linear activation functions were used in the last layer. The loss is defined as follows

$$loss(\hat{y}, y) = -\frac{1}{N}\sum_{i=0}^{N}\frac{|\langle \hat{y}_i, y_i \rangle|}{||\hat{y}_i||_2 * ||y_i||_2} \quad (2)$$

with N being the number of classes, y the training target and $\hat{y}$ the network output. In Wasserthal et al. (2018a) we used the cosine similarity in combination with the peak length as loss, thus allowing the model to also learn the extent of each bundle. However, learning the tract segmentation and the peak angles in two separate models is giving better results (for details see supplementary materials). Therefore we only use the cosine similarity in the loss of the tract orientation mapping model.

2.3.2. Hyperparameters

Leaky rectified linear units (ReLU) were used as nonlinearity (Nair and Hinton, 2010). A learning rate of 0.001 was used and Adamax (Kingma and Ba, 2014) was chosen as an optimizer. When the validation loss did not decrease for at least 20 epochs the learning rate was reduced by one order of magnitude. The batch size was 47. All hyperparameters were optimized on a validation dataset independent of the final test dataset. The network weights of the epoch with the highest Dice score during validation were used for testing.

2.3.3. Data augmentation

To improve the generalizability of our model, we applied heavy data augmentation to the peak images during training.[1] The following transformations were applied to each training sample. The intensity of each transformation was varied randomly by sampling from a uniform distribution U.

- Rotation by angle $\varphi_x \sim U[-\pi/4, \pi/4]$, $\varphi_y \sim U[-\pi/4, \pi/4]$, $\varphi_z \sim U[-\pi/4, \pi/4]$
- Elastic deformation with alpha and sigma $(\alpha, \sigma) \sim (U[90, 120], U[9, 11])$. A displacement vector is sampled for each voxel $d \sim U[-1, 1]$, which is then smoothed by a Gaussian filter with standard deviation σ and finally scaled by α.
- Displacement by $(\Delta x, \Delta y) \sim (U[-10, 10], U[-10, 10])$
- Zooming by a factor $\lambda \sim U[0.9, 1.5]$
- Resampling (to simulate lower image resolution) with factor $\lambda \sim U[0.5, 1]$
- Gaussian noise with mean and variance $(\mu, \sigma) \sim (0, U[0, 0.05])$

The training samples were normalized to zero mean and unit variance before passing them to the network. When training our network on peaks generated by the MRtrix multi-shell multi-tissue CSD method, we found that it did not work well on peaks generated by the standard MRtrix CSD method. In order to ensure our model worked well with all types of MRtrix peaks, we generated three peak images: (1) multi-shell multi-tissue CSD using all gradient directions, (2) standard CSD using only $b = 1000$ s/mm^2 gradient directions, (3) standard CSD using only 12 gradient directions at $b = 1000$ s/mm^2. During training, we randomly sampled from these three peak images, thus ensuring that our network worked well with all of them. We trained for 250 epochs with each epoch corresponding to 193 batches. This means that over the course of the entire training, the network has seen 2,267,750 slices which have been randomly sampled from axial, coronal and sagittal orientations, randomly sampled from three different peak types and randomly permutated by the data augmentation transformations. The results presented in Section 3 were obtained using an implementation of the proposed method in Pytorch.[2]

2.3.4. Super resolution

Our models were trained with images of size 144×144 corresponding to the 1.25 mm resolution of the HCP data. As mentioned in Section 2.3.3 we were using resampling as data augmentation. This means images were downsampled to a resolution of 2.5 mm to simulate lower resolution images. Then they were upsampled back to 1.25 mm to fit the 144×144 input size of the model. So the resolution kept the same but the images got blurred by the down- and upsampling. This down/upsampling was only done for the input images (peaks) not for the labels (training target). This way the models were able to learn a higher resolution output than was actually provided as input. This is commonly referred to as *super resolution* (Alexander et al., 2017). When our approach receives a low resolution image as input, it is first upsampled to resolution 1.25 mm and then fed to the model which returns a output also in 1.25 mm resolution. This higher resolution especially helps on very thin bundles like the anterior commissure (CA).

2.4. Data

For training our models the dataset published by Wasserthal et al. (2018b) was used. It contains reference delineations of 72 major white matter tracts (see supplementary materials for a list of all tracts) in 105 subjects from the Human Connectome Project. The details of how this dataset was curated are described in Wasserthal et al. (2018b). The reference delineations are provided in form of streamlines. In this paper we refer to this dataset as *reference data* or *reference tracts*.

2.4.1. Preprocessing of reference data for different tasks

To be able to use the dataset for our three tasks (tract segmentation, start/end region segmentation and tract orientation mapping) some preprocessing was necessary:

For the tract segmentation we convert the reference streamlines to binary masks by setting each voxel to *True* where at least one streamline runs through.

For the start/end region segmentation we create binary masks from the streamlines start and end points. However, streamlines have no defined direction. So for example for the corticospinal tract some streamlines start at the cortex whereas other streamlines start at the brain stem. Therefore the start point of one streamline might be in the same region as the end point of the next streamline. The resulting binary mask is the union of the start and end region of a tract. Splitting the union into two binary masks, one for the start and one for the end region is not trivial as for some tracts like the uncinate fasciculus those region can be very close together. To avoid manual separation (which is time consuming and less objective) we took the following approach for splitting the regions: First we used a clustering algorithm (DBSCAN Ester et al., 1996) to create two clusters from the combined region. The clustering was only done on a subset of the data points to avoid long runtimes. When the start and end region were close together the clustering sometimes misassigned points. Therefore we used the results from the clustering to train a random forest. This led to a correct separation of the two region for all subjects and ensured fast runtime when running for all data points. From the points in those two regions binary masks were created. Finally we did binary closing and a small amount of binary dilation using scipy (Jones et al., 2001) to create a consistent region from the single points.

[1] https://github.com/MIC-DKFZ/batchgenerators.

[2] www.pytorch.org.

For the tract orientation mapping the main streamline orientation in each voxel had to be determined for each tract. Using the mean of all streamlines running through a voxel led to rather noisy results. Therefore we used Mean Shift clustering to group the orientations of all streamlines in one voxel. Then the mean of the orientations in the biggest cluster was taken as final orientation for that voxel. This substantially reduced the noise.

2.4.2. Clinical quality dataset

The reference dataset is provided in high HCP data quality (*HCP Quality*). However, in clinical routine, faster MRI protocols are used which result in lower quality data. To test how the proposed method performs on clinical quality data, we downsampled the HCP data to 2.5 mm isotropic resolution and removed all but 32 weighted volumes at $b = 1000\text{s/mm}^2$. We call this dataset *Clinical Quality*. The reference tracts from the *HCP Quality* dataset were reused as our reference tracts here. This provides high quality reference tracts for the low quality data, thus allowing proper evaluation.

2.4.3. Phantom dataset

The *Clinical Quality* dataset has lower resolution and less directions than the *HCP Quality* dataset but it was still acquired by the same scanner. To evaluate how the proposed method generalizes to images from other scanners and other acquisition settings we would need a dataset with reference tract delineations from another scanner. Unfortunately such a dataset is not available and using the same approach as was used for the Wasserthal et al. (2018b) dataset is not feasible: For lower quality datasets it becomes very difficult and ambiguous for an expert to accurately determine where tracts run. The expert delineations would rather be approximations not suitable for detailed evaluation. One solution, however, is to simulate low quality data from a different scanner. Thereby we have perfect ground truth and still low image quality. We used the toolkit FiberFox (Neher et al., 2014) to create such software phantoms. We selected 21 subjects (not used for training) from the reference data and for each simulated the diffusion weighted image of a brain containing only the 72 reference tracts. The simulated images have an isotropic resolution of 2.5 mm, 32 gradient directions at $b = 1000$ mm/s and several artefacts which were randomly chosen from the following list: head motion, ghosts, spikes, eddy currents, ringing, distortions, signal drift and complex Gaussian noise. We call this dataset *Phantom*. As the *Clinical quality* dataset and the *Phantom* dataset only have one b-value shell, we cannot use multi-shell CSD as we did for the *HCP Quality* data. Instead MRtrix standard CSD was used to generate the peaks of the FOD.

2.5. Bundle-specific tractography

2.5.1. Flavors of TOM

There are three different ways how tract orientation maps can be used to create bundle-specific tractograms:

- Directly track on the tract orientation maps
- For each voxel select the peak from the original input peaks which is closest to the orientation predicted by TOM. Then track on these peaks. This has the advantage of staying closest to the original signal, but if the original peaks are quite noisy the chosen peak will also be noisy. This is a problem especially on low quality data.
- Use the tract orientation map as a prior by taking the weighted mean between the predicted orientation from the TOM and the original orientation normally used for tracking.

Fig. 3 shows exemplary results for the different tracking options on one subject from a low resolution dataset. In all four cases the tract masks as well as the start/end region masks were used to filter the tractograms. Tracking on the original signal is insufficient: Deterministic tracking lacks sensitivity whereas probabilistic trackings lacks specificity (many false positives). Tracking on the tract orientation maps gives the best of both: high sensitivity (tract is complete) and high specificity (few false positives). Tracking on the best original peaks also shows good results but is missing small parts of the lateral projections of the CST. Therefore for our experiments we chose the first option: Directly track on the tract orientation maps. This gave the best results, especially on low quality data where the original peaks can be quite noisy.

2.5.2. Probabilistic tracking on peaks

The output of the tract orientation mapping is one tract orientation map for each tract. A tract orientation map contains one 3D vector (one peak) at each voxel telling the main orientation of the respective tract at that voxel. Creating streamline from these maps could easily be done by using deterministic tracking (e.g. Mori et al., 1999; Basser et al., 2000). This works well on high resolution data. However, on low resolution data just following the main orientation in each voxel sometimes leads to small branchings being missed as they cannot be represented on the low resolution. Probabilistic tracking enables more sensitive tracking. By not just following the main orientation in each voxel but sampling from the orientation distribution smaller branchings can be reconstructed that otherwise would be missed. In our case, however, only one orientation per voxel is provided by the tract orientation map, but no orientation distribution. To be able to sample from

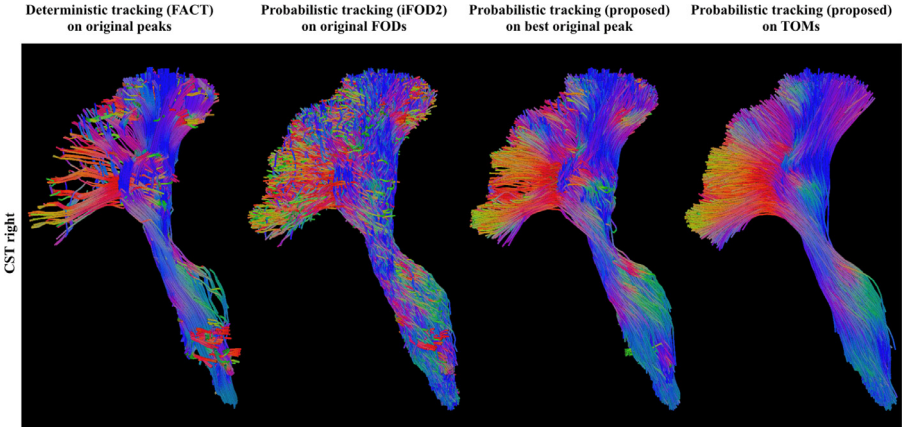

Fig. 3. Right corticospinal tract (CST) in one subject from the BrainGluSchi (Bustillo et al., 2017) dataset (2 mm isotropic resolution, 30x $b = 800$ mm/s^2) reconstructed by the different tracking variants. Probabilistic tracking on TOMs shows the best results in terms of sensitivity and specificity.

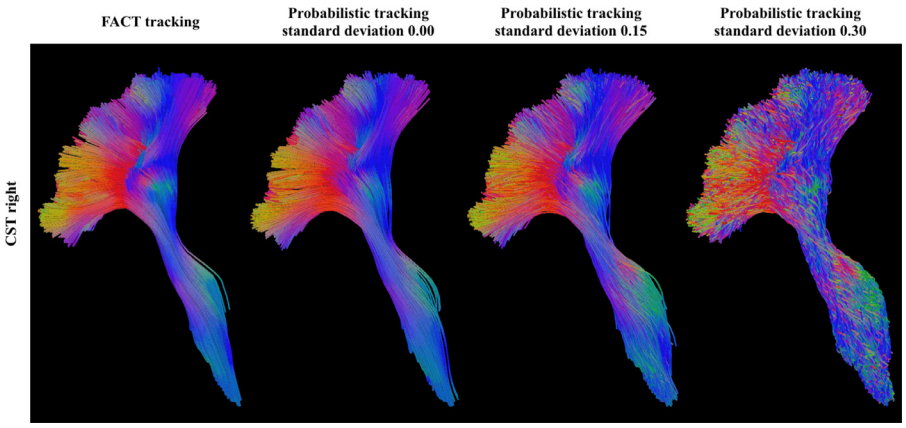

Fig. 4. Reconstructions of the right corticospinal tract (CST) in one subject from the BrainGluSchi (Bustillo et al., 2017) dataset (same subject as in Fig. 3). Comparison of deterministic FACT tracking on tract orientation maps (TOMs) with proposed probabilistic tracking with different standard deviations for the Gaussian distribution used for sampling. Up to a standard deviation of 0.15 the probabilistic tracking results are very consistent with the FACT tracking. Using higher values leads to clearly more spurious streamlines.

orientations around the main orientation, we use a Gaussian distribution centered on the main orientation with a fixed standard deviation. When using a small value for the standard deviation this can be interpreted as modeling a lower bound for the FOD dispersion, as in real data the width of the FOD lobes of all major tracts is non-zero. Even in areas of highly consistent fiber orientation like the corpus callosum there still is dispersion. For the fixed standard deviation we chose a value (0.15) which leads to less dispersion than in the corpus callosum FODs (see supplementary materials for more details). This way the fixed standard deviation is a conservative estimate of the lower bound of the dispersion and helps to increase sensitivity compared to using plain deterministic tracking.

Trying to learn a more complete orientation distribution for each voxel instead of only learning the main orientation and using a global lower bound for the dispersion could be promising. However, experiments in this direction did not lead to better tracking result. Therefore we decided to stick with the easier approach of an empirically determined fixed standard deviation.

Although our approach is probabilistic it is quite different from the algorithms commonly referred to as probabilistic tracking e.g. Behrens et al. (2003). For those probabilistic tracking algorithms the orientation distribution is representing the uncertainty in the underlying signal. In our approach the probabilistic component is only a way for increasing sensitivity to capture small branchings (relating it to Ankele et al., 2017) by assuming a lower bound on the fiber orientation uncertainty.

Our tracking algorithm is based on the deterministic algorithm described by Basser et al. (2000) with the main difference that at each step the next orientation to take is sampled from the given Gaussian distribution. Then the tracking algorithm is taking a step (with fixed step size) along this sampled orientation. At the end the streamlines are interpolated using b-splines. All streamlines have to start and end in the regions segmented by the start/end region segmentation model and are not allowed to leave the mask generated by the tract segmentation model, otherwise they are discarded.

Using probabilistic tracking increases sensitivity but this often comes at the cost of an increased number of false positives. In our case we can keep the number of false positives introduced by the probabilistic tracking quite small as the tracking is highly constrained by the tract mask as well as the start/end region mask. Moreover we use a Gaussian distribution with a quite small standard deviation of 0.15. As can be seen in Fig. 4 using a higher value like 0.3 would lead to many spurious fibers. But using a value of 0.15 results in trackings which are highly consistent with deterministic FACT tracking, showing that our approach is very close to the specificity of deterministic tracking while being more sensitive.

During tracking the following parameters were used: a step size of 0.7 voxels and a minimum streamline length of 50 mm. Seeds were randomly placed inside of the tract mask until a maximum of 2000 streamlines per tract were created.

2.6. Reference methods

We compared our proposed method to 3 methods for automatic tract delineation (comparing segmentation performance in terms of DICE score and orientation quality in terms of voxel-wise angular error): TractQuerier (Wassermann et al., 2016), RecoBundles (Garyfallidis et al., 2017) and streamline atlas. Moreover we compared to 3 methods for tract segmentation (comparing only segmentation performance): custom atlas registration, FSL atlas registration and multiple mask registration. We also compared to 2 methods which give a voxel-wise orientation for each tract (comparing only orientation quality): Peak atlas and the best original peak. These methods include clustering-based as well as ROI-based approaches. We give an outline of how they work (1.) and how we applied them (2.).

2.6.1. TractQuerier

1. TractQuerier (Wassermann et al., 2016) extracts tracts based on the regions the streamlines have to start at, end at and (not) run through. 2. We compared our method to the output from TractQuerier using the same queries as used in Wasserthal et al. (2018b) without any further post-processing. The queries were applied to a whole brain tracking generated with MRtrix (Tournier et al., 2010) and the following settings: Constrained spherical deconvolution (for *HCP Quality* data using the multi-shell multi-tissue option) was used to extract the FOD and probabilistic tractography (iFOD2) (for *HCP Quality* data using the anatomically constrained option) was used to generate a whole brain tractogram. For the *HCP Quality* data 10 million streamlines were generated, for all other datasets 500,000 streamlines. The minimum length was set to 40 mm. The other parameters were kept at their default values. As parcellation the freesurfer Deskian/Killiany atlas was used (Desikan et al., 2006). Freesurfer was applied using the default settings.

2.6.2. RecoBundles

1. Given streamlines of a reference tract in a reference subject, RecoBundles (Garyfallidis et al., 2017) can be used to find the corresponding streamlines in a new subject. 2. We randomly picked 5

reference subjects from the training dataset. Due to the long runtime for RecoBundles, a higher number of reference subjects was not feasible. Then we ran RecoBundles 5 times for the new subject (once for each reference subject) using the default RecoBundles parameters (see supplementary materials) and the same whole brain tractogram used for TractQuerier (see Section 2.6.1). This resulted in 5 extractions of each tract in the new subject. To get a final segmentation, we took the mean of those 5 extractions.

2.6.3. Streamline atlas registration (SLAtlas)

1. Given streamlines of a reference tract in a reference subject, registration can be used to align them with a new subject. 2. The same 5 reference subjects as those selected for RecoBundles were used. To delineate the tracts in a new subject, we registered each of the 5 reference subjects to the new subject. Affine registration of the whole brain tractograms was already done by RecoBundles so we reused these transformations. Finally for each tract we merged the streamlines from the 5 registered reference subjects.

2.6.4. Atlas registration 1 (Atlas Custom)

1. Several subjects can be averaged to an atlas which can then be registered to new subjects to segment structures. 2. We split our dataset into training and testing data, using the same 5-fold cross-validation as used for the evaluation of our proposed method (see Section 3). The training data was used to create a tract atlas. Firstly, we registered all subjects to a random subject using symmetric diffeomorphic registration implemented in DIPY (Avants et al., 2008; Garyfallidis et al., 2014). Registration was performed based on the FA maps of each image. After registration, the FA maps of all images were averaged. Then, in a second iteration all images were registered to this mean FA image. This two-stage approach limits the bias introduced by the initial subject choice in the first iteration. The tract atlas thus contained the tract masks for all 72 reference tracts. For each tract, we took the mean over all subjects, which produced a probability map. We thresholded the probability map at 0.5 to create a final binary atlas. During test time, the atlas was registered to the subjects of interest, yielding a binary mask for each tract in subject space.

2.6.5. Atlas registration 2 (Atlas FSL)

1. We compare to a second atlas method using a different implementation (FSL) and a different atlas template. 2. This method is identical to the previous method (Atlas Custom) except for the following two points: Instead of generating a mean FA template from the dataset we used the FA template provided by FSL (FMRIB58_FA_1mm). Instead of using the diffeomorphic registration of DIPY we used the linear (FLIRT) and nonlinear (FNIRT) registration of FSL (Jenkinson et al., 2012). As configuration we used the predefined configuration file for FA registration provided by FSL (FA_2_FMRIB58_1mm.cnf).

2.6.6. Atlas registration 3 (Atlas MRtrix)

1. We compare to a third atlas method which is not based on registering FA images but on registering the FODs (fiber orientation distributions) thereby make use of the richer information of FODs compared to only using the FA. This method is implemented in MRtrix (Raffelt et al., 2011). 2. This method is identical to the *Atlas Custom* method except for the following two points: Instead of generating a mean FA template from the dataset we used the MRtrix method *population_template* to create a FOD template. Instead of using the diffeomorphic registration of DIPY we used the MRtrix nonlinear FOD registration. This method was not applicable for the *Phantom dataset* as the FODs from the simulated Phantom data were not similar enough to the FODs from the FOD template to allow for meaningful registration.

2.6.7. Multiple mask registration (Multi-Mask)

1. Using an atlas can blur some of the details as it is based on group averages. The blurring can be reduced to some extent by registering the masks of single training subjects to a test subject instead of an averaged atlas. 2. The same 5 reference subjects as those selected for *RecoBundles* were used. To segment the tracts in a new subject, we registered each of the 5 reference subjects to the new subject (symmetric diffeomorphic registration of the FA maps) and averaged the tract masks (from the reference tracts) of all 5 reference subjects. Finally, we thresholded this average at 0.5 to produce a binary mask for each tract in the space of the new subject. This differs from the *Atlas registration* method in that the reference subjects are directly registered to subject space and are merged (1 registration) instead of first being registered to atlas space, then merging and being registered to subject space (2 registrations needed). Moreover, *Atlas Registration* uses 63 subjects while *Multi-Mask* only uses 5.

2.6.8. Peak atlas

This method is identical to *Atlas FSL* with the only difference that instead of using binary masks we use peak images. The transformation calculated from registering the FA images is applied to each of the 3 peaks of the peak image independently using FSL vecreg which makes sure the peaks are reoriented accordingly (Alexander et al., 2001).

2.6.9. Best original peak (BestOrig)

1. Given the peak map of a reference tract, in each voxel we can choose the peak from the original signal that is closest to the peak from the reference tract, resulting in a new peak map. 2. For each subject in the test set we use the reference peaks to extract the best peak from the original signal. As we are using the ground truth in this method, it is not a fair method to directly compare to but it gives a good estimation of how good the original peaks are.

3. Experiments and results

For evaluation 5-fold cross-validation was used, i.e. 63 training subjects, 21 validation subjects (best epoch selection) and 21 test subjects per fold. The Wilcoxon signed-rank test (Wilcoxon, 1945) was used to test for statistical significance when comparing our method and the reference methods in the quantitative evaluation. For multiple testing, we applied the Bonferroni correction.

3.1. Segmentation performance

For evaluating segmentation performance we used the Dice score (Taha and Hanbury, 2015) as our metric. The Dice score measures the overlap between two binary masks. It ranges from 0 to 1 with 1 being a perfect overlap. We calculated the Dice for each subject between each of the 72 reference tracts and the respective prediction of either our proposed method or one of the reference methods (e.g. RecoBundles). Then we averaged the Dice results for all 72 tracts to get one final Dice score per subject per method. Over all three datasets (*HCP Quality, Clinical Quality* and *Phantom*) our proposed method significantly ($p < 0.01$) outperformed the reference methods by a large margin: on the *HCP Quality* dataset it outperformed the reference methods on average by 14 Dice points and on the low quality datasets on average by 18 Dice points (*Clinical Quality*) and 22 Dice points (*Phantom*) (Fig. 5). In general, the proposed method was less affected by the quality loss in the *Clinical Quality* and *Phantom* data than the reference methods.

3.2. Orientation performance

For evaluating orientation performance we use the voxel-wise angular error as metric. We calculate the angular error between

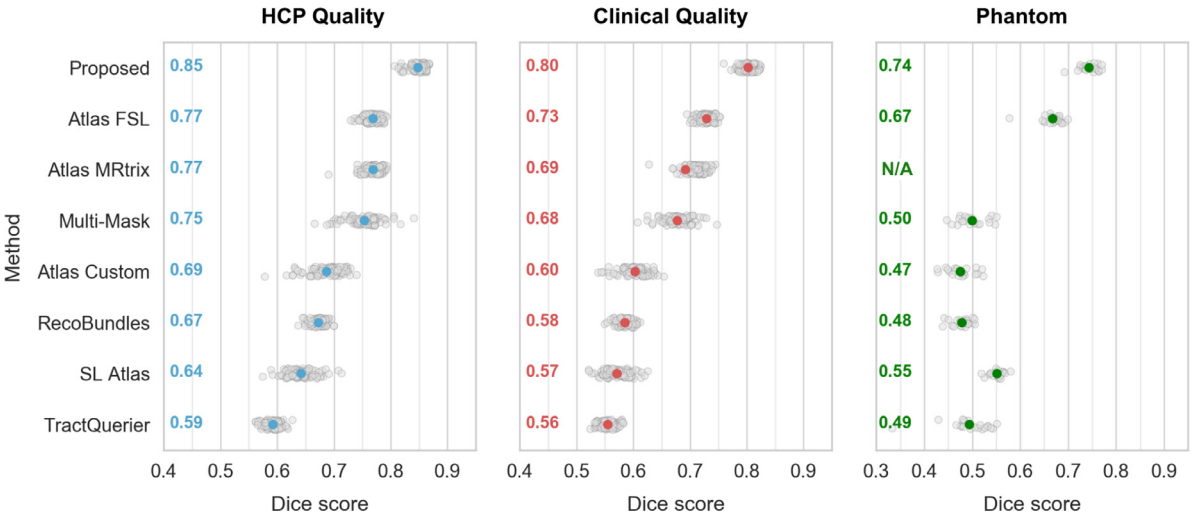

Fig. 5. Segmentation results on the *HCP Quality,Clinical Quality* and *Phantom* dataset with a gray dot per subject (mean over all tracts) and a colored dot for the mean over all subjects. Proposed: Our method; Multi-Mask: Multiple mask registration; Atlas: Atlas registration; SLAtlas: Streamline atlas.

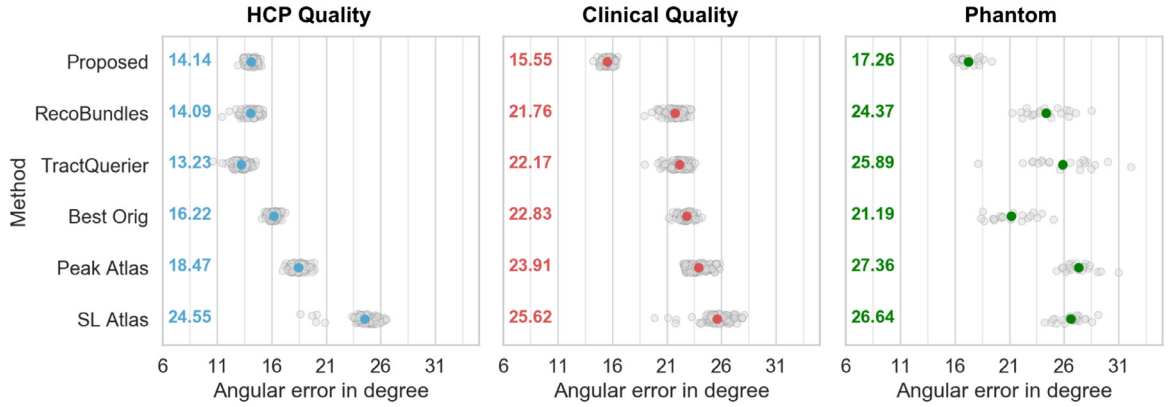

Fig. 6. Orientation performance results on the *HCP Quality,Clinical Quality* and *Phantom* dataset with a gray dot per subject (mean over all tracts) and a colored dot for the mean over all subjects. Proposed: Our method; BestOrig: Best original peak; SLAtlas: Streamline atlas.

the reference orientation and the orientation of the proposed method. We do this for every voxel where the reference peak and the peak of the proposed method have a length greater than zero. Then we average the errors to get one final angular error per subject per method. To calculate the voxel-wise main streamline orientation for the methods which output streamlines (*RecoBundles, TractQuerier* and *SLAtlas*) we used the same technique as used for calculating the main streamline orientation for the reference data (see Section 2.4.1): The streamline orientations in each voxel were first clustered and then the mean of the biggest cluster was chosen. On the *HCP Quality* data *RecoBundles* and *TractQuerier* show slightly better orientation errors than our proposed method. Those methods have difficulties finding the borders of tracts (poor segmentation performance) but they are good at finding the correct streamlines belonging to the core of the tract and therefore show low orientation errors, as long as the underlying whole brain tractogram is of high quality. As soon as the image quality gets lower (*Clinical Quality* and *Phantom* dataset), the whole brain tracking also suffers and therefore the angular error of these methods rises significantly. Our proposed method on the other hand is not dependent on the whole brain tracking and quite robust to lower image quality as it was trained with extensive data augmentation. As a results the angular error only rises by 1 ° when using our proposed method on the *Clinical Quality* data compared to the *HCP Quality* data (Fig. 6). *SLAtlas* and *Peak Atlas* show high angular errors for all three datasets.

Fig. 7 shows the angular error for each tract independently on the *Clinical Quality* dataset.

3.3. Qualitative evaluation

For the qualitative evaluation, one subject (623844) was selected from the test set. We chose a subject whose Dice scores were closest to the mean Dice scores for the entire datasets to make the subject representative for the entire dataset. Since the scope of this manuscript does not allow us to show results for all 72 tracts, we selected three tracts that represent different degrees of reconstruction difficulty according to Maier-Hein et al. (2017): The inferior occipito-frontal fascicle (IFO), corticospinal tract (CST) and anterior commissure (CA). The IFO is a tract which is fairly easy to reconstruct, which is reflected by its consistently good scores for all methods. The CST is more difficult to reconstruct. Its beginning at the brain stem is easy to reconstruct but as the fibers get closer to the cortex, they start to fan out. Finding these lateral projections is more difficult. Finally, the CA is a tract that is difficult to reconstruct. Due to its very thin body, it is hard to find streamlines running the entire way from the right to the left temporal lobe. The CA is one of the tracts with the lowest performance out of all of the methods. We show results for all reference methods that produce streamline output (*RecoBundles, TractQuier* and *SLAtlas*). For each tract one 3D view is shown as well as one 2D slice allowing more in detail evaluation. On the 2D slice the

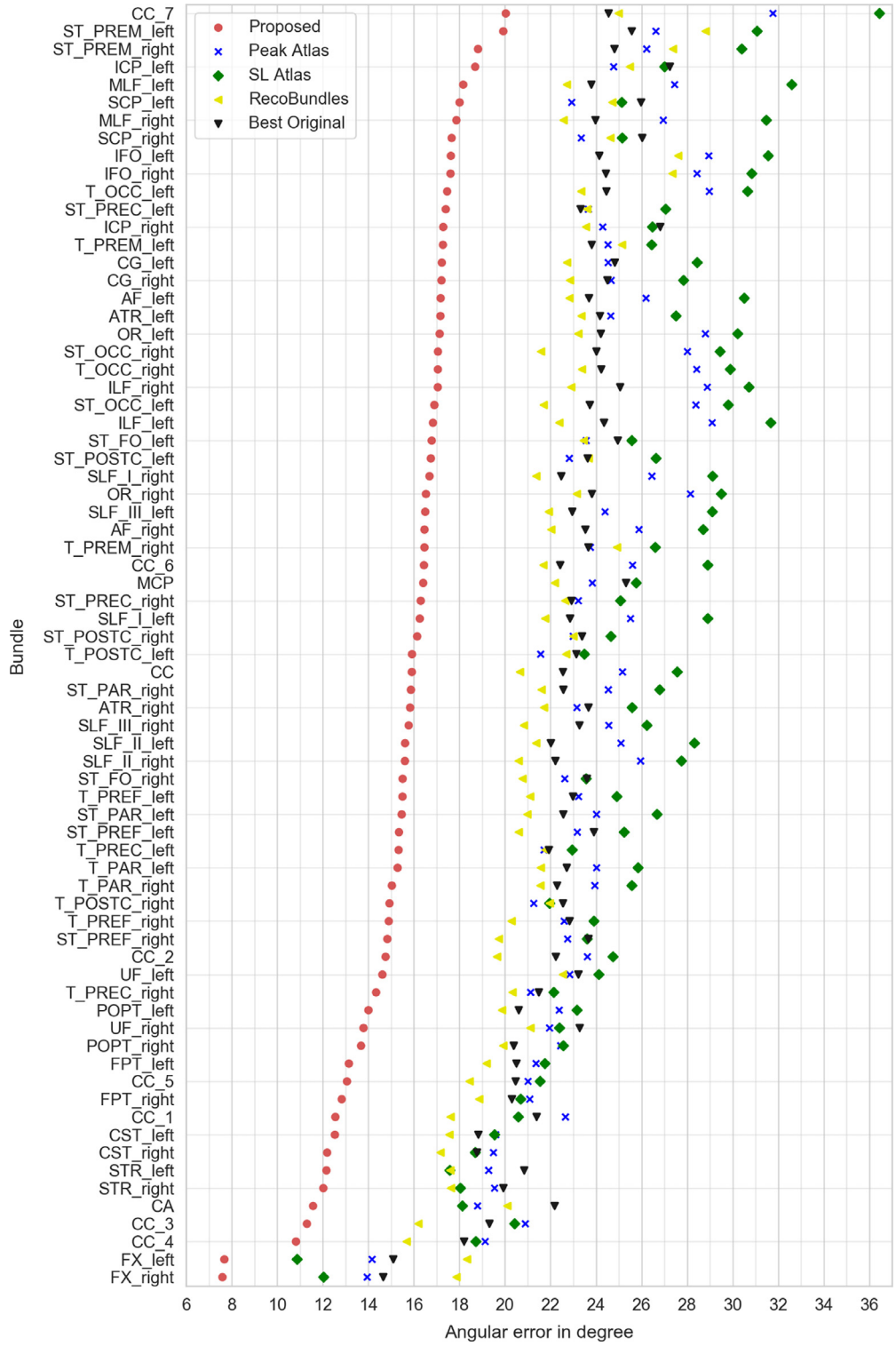

Fig. 7. Angular errors for all 72 tracts on the *Clinical Quality* dataset for our proposed method and all reference methods sorted by error. The following list shows the full names of each tract: Arcuate fascicle (AF), Anterior thalamic radiation (ATR), Anterior commissure (CA), Corpus callosum (Rostrum (CC 1), Genu (CC 2), Rostral body (CC 3), Anterior midbody (CC 4), Posterior midbody (CC 5), Isthmus (CC 6), Splenium (CC 7)), Cingulum (CG), Corticospinal tract (CST), Middle longitudinal fascicle (MLF), Fronto-pontine tract (FPT), Fornix (FX), Inferior cerebellar peduncle (ICP), Inferior occipito-frontal fascicle (IFO), Inferior longitudinal fascicle (ILF), Middle cerebellar peduncle (MCP), Optic radiation (OR), Parieto-occipital pontine (POPT), Superior cerebellar peduncle (SCP), Superior longitudinal fascicle I (SLF I), Superior longitudinal fascicle II (SLF II), Superior longitudinal fascicle III (SLF III), Superior thalamic radiation (STR), Uncinate fascicle (UF), Thalamo-prefrontal (T_PREF), Thalamo-premotor (T_PREM), Thalamo-precentral (T_PREC), Thalamo-postcentral (T_POSTC), Thalamo-parietal (T_PAR), Thalamo-occipital (T_OCC), Striato-fronto-orbital (ST_FO), Striato-prefrontal (ST_PREF), Striato-premotor (ST_PREM), Striato-precentral (ST_PREC), Striato-postcentral (ST_POSTC), Striato-parietal (ST_PAR), Striato-occipital (ST_OCC).

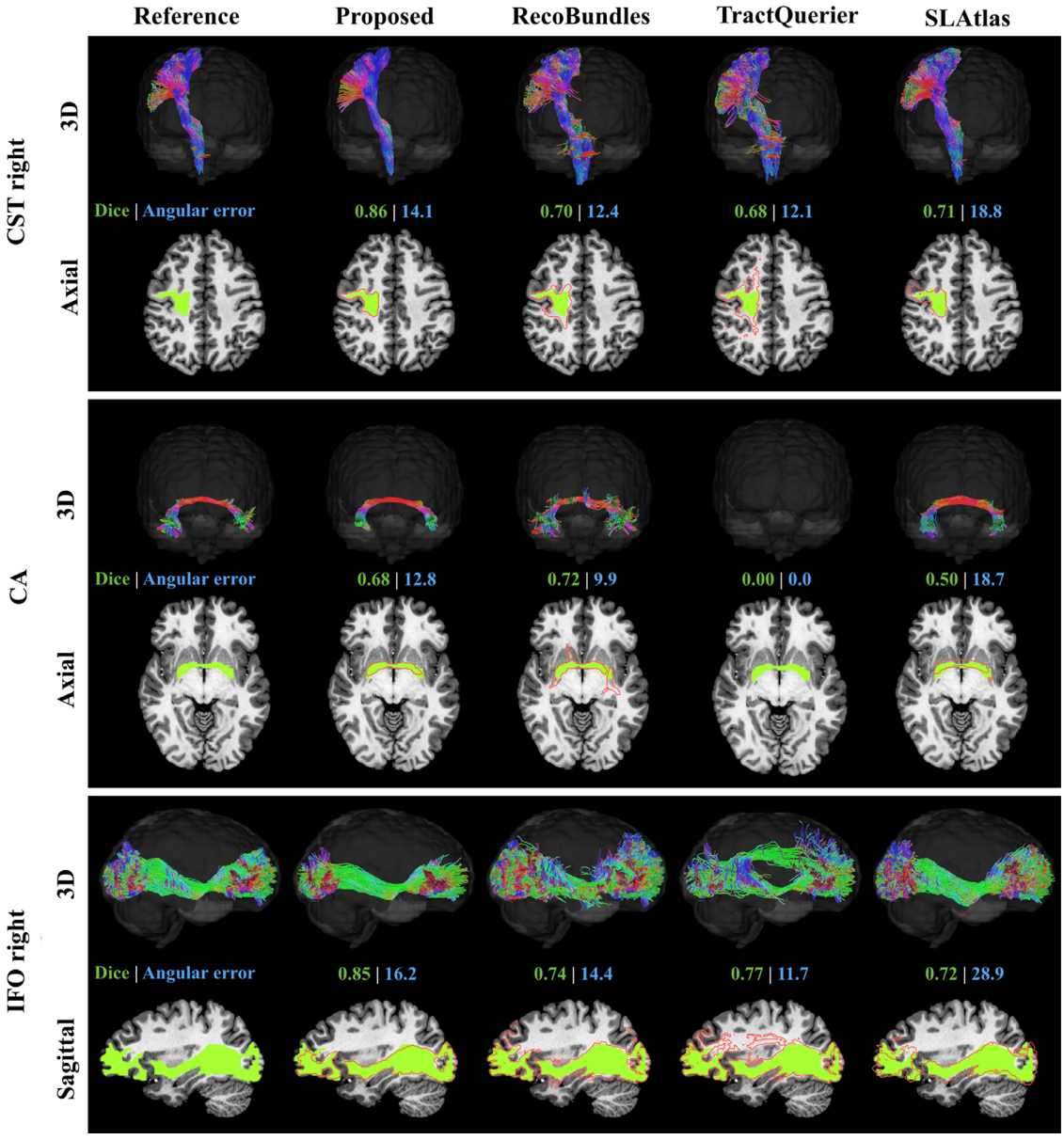

Fig. 8. Qualitative comparison of results on *HCP Quality* test set: reconstruction of right corticospinal tract (CST), anterior commissure (CA) and right inferior occipito-frontal fascicle (IFO) on subject 623844. Green shows the reference tract and red shows the tract mask of the respective method. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

mask of the bundle is shown (red) as well as the mask of the reference bundle (green).

As can be seen in Fig. 8, the *Proposed* method yielded accurate and spatially coherent reconstructions on all three tracts. *RecoBundles* oversegmented the CST to neighbouring gyri and selected many streamlines ending prematurely instead of reaching the correct start and end regions of the tract. *TractQuerier* did not properly segment any of the example tracts. As it defines tracts mainly by their endpoints, it leaves much room for wrong turns between the start and end points. *TractQuerier* extracts a lot of false positives, especially when using probabilistic tracking. The CA cannot be properly reconstructed with *TractQuerier* as the default Freesurfer parcellation is not precise enough for the small parts of the CA. *SLAtlas* produces reconstructions looking convincing on first sight but when looking at them on the 2D view we can see that it involves severe oversegmentation (e.g. segmenting gray matter and non-brain area for the CST) and slightly shifted tracts (e.g. CA). This is most probably owed to the affine registration, which cannot fully resolve the inter-subject variability.

For the *Phantom* dataset the different methods in principle show the same shortcomings as for the *HCP Quality* dataset, but now more severely (see Fig. 16 in the supplementary materials).

3.4. Generalization to other datasets and pathologies

To test the capability of the proposed method to generalize beyond HCP, which it was trained on, we applied it to 17 differently acquired datasets (including many public datasets like the OASIS, IXI, COBRE or Rockland datasets). A full list of all datasets can be found in the supplementary materials. These 17 datasets represent a wide variety of data: Different scanners, different spatial resolutions, different b-values, different number of gradients, healthy and diseased, normal and abnormal brain anatomy.

An expert manually dissected the three tracts already shown in the above qualitative evaluation (CST, CA and IFO) from one

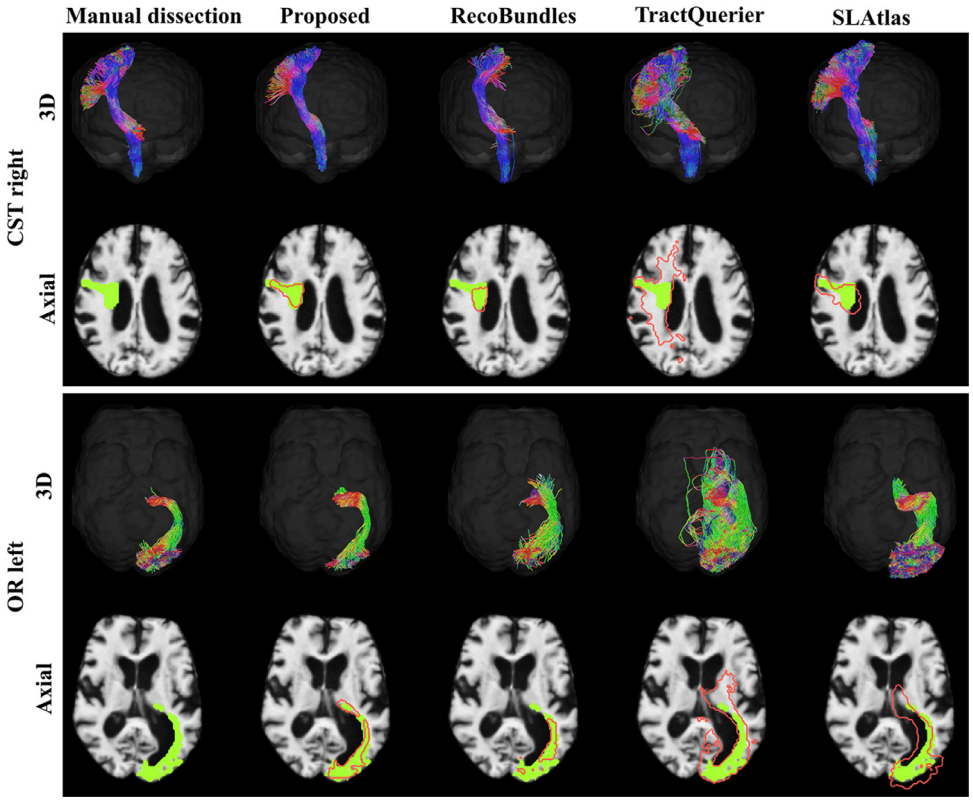

Fig. 9. Qualitative comparison of results on one Alzheimer patient with enlarged ventricles from the OASIS dataset: reconstruction of right corticospinal tract (CST) and left optic radiation (OR). Green shows the manual dissection and red shows the tract mask of the respective method. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

randomly chosen subject from each dataset shown in table 1. Visual comparisons were performed between manual dissections and the results of our proposed method as well as the previously introduced reference methods *RecoBundles, TractQuerier* and *SLAtlas*. Methods depending on reference data (all three methods) were provided with our HCP reference data. All subjects (except for subjects from HCP datasets which did already receive basic preprocessing) were denoised (using MRtrix Veraart et al., 2016), corrected for eddy currents and motion artifacts (using FSL eddy Andersson and Sotiropoulos, 2016) and rigidly registered to MNI space. This rigid registration is not required for our proposed method to work. It only requires that the left/right, front/back and up/down orientation of the images are the same as for the HCP data (i.e. images are not mirrored). Rigid registration to MNI space is an easy way to ensure this.

Our proposed method showed anatomically plausible results for all subjects and most of the tracts. Only the CA was not completely reconstructed in around 20% of the subjects. We observed partly incomplete manual reference dissections in these areas as well, indicating that the size of this very thin tract is reaching the resolution limit of the underlying imaging acquisition. Figs. 9–11 show exemplary results for three subjects with pathologies. For those subjects we show the corticospinal tract (CST), the optic radiation (OR) and the thalamo-postcentral tract (TPOSTC) as those tracts are heavily affected by the respective pathology. The other 19 subjects can be found in the supplementary materials.

Fig. 9 shows the results for an alzheimer patient with abnormally large ventricles from the OASIS dataset. Even though our proposed method has only seen healthy subjects with normally sized ventricles during training it managed to properly reconstruct the CST and the OR which are heavily distorted by the enlarged ventricles. *RecoBundles* also managed to find the distorted tracts. However, it failed to find the lateral projections of the CST and oversegmented the Meyer's loop of the OR. *TractQuerier* showed severe oversegmentation of both the CST and OR. *SLAtlas* did not manage to adapt to the enlarged ventricles. It placed the tracts inside of the ventricles as the affine registration is not able to resolve these distortions.

Fig. 10 shows the results for an multiple sclerosis (MS) patient with severe lesions in the pathways of the CST and OR (marked with arrows in the figure). Inside of MS lesions demyelination takes place, leading to a loss in diffusion-weighted signal. However, the axons themselves are still intact. Therefore fibers are still running through the lesions but they are harder to reconstruct as the signal is weakened by the demyelination. Our proposed method manages to properly reconstruct streamlines running through these lesions. *RecoBundles* and *TractQuerier* were also able to reconstruct streamlines running through the lesions as they use tracking based on constrained spherical deconvolution which shows good results in reconstructing orientation information inside of the lesions (using a simple tensor model would not be sufficient to reconstruct the orientation information inside of the lesions). However, *RecoBundles* and *TractQuerier* are failing in properly reconstructing the entirety of the tracts: *RecoBundles* fails to reconstruct the Meyer's loop of the OR and *TractQuerier* show severe oversegmentation of both tracts. We do not show a reference tract delineation for this subject as the lesions would not be visible anymore then.

Fig. 11 shows the results for a patient with mild brain volume loss and schizoaffective disorder. Around the postcentral gyrus the volume loss is more severe. We show results for the Thalamo-postcentral tract (T_POSTC), containing fibers running from the thalamus to the postcentral gyrus. Despite the reduced brain volume our proposed method managed to correctly reconstruct the

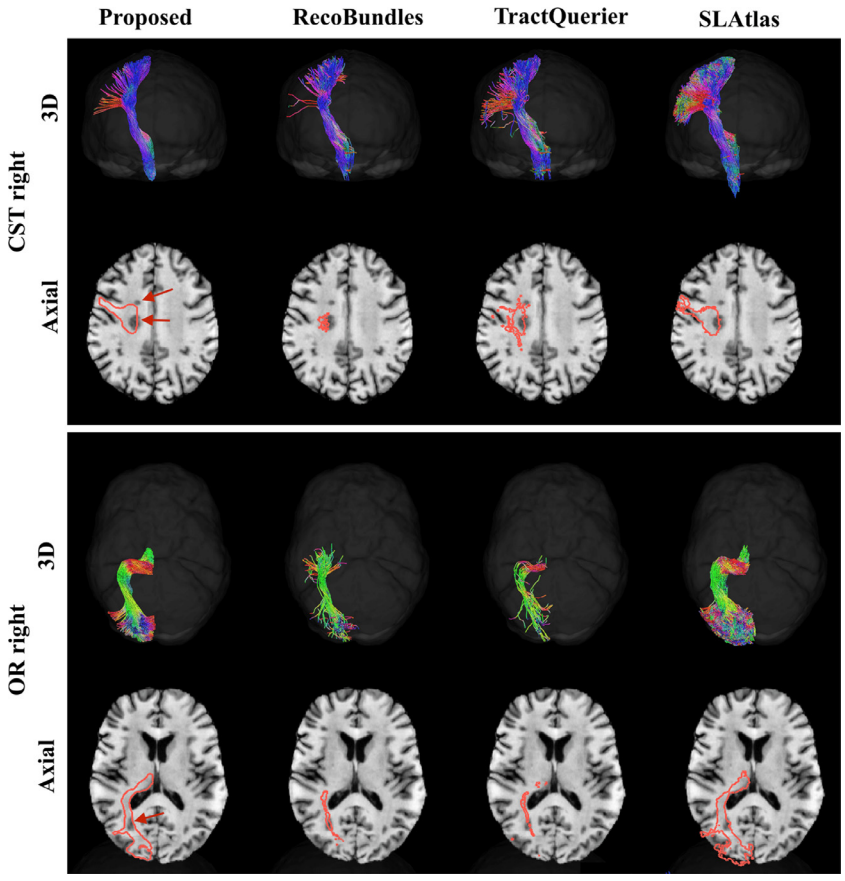

Fig. 10. Qualitative comparison of results on one multiple sclerosis patient with several lesions inside of the tracts: reconstruction of right corticospinal tract (CST) and right optic radiation (OR). Red arrows show lesions close to the tracts.

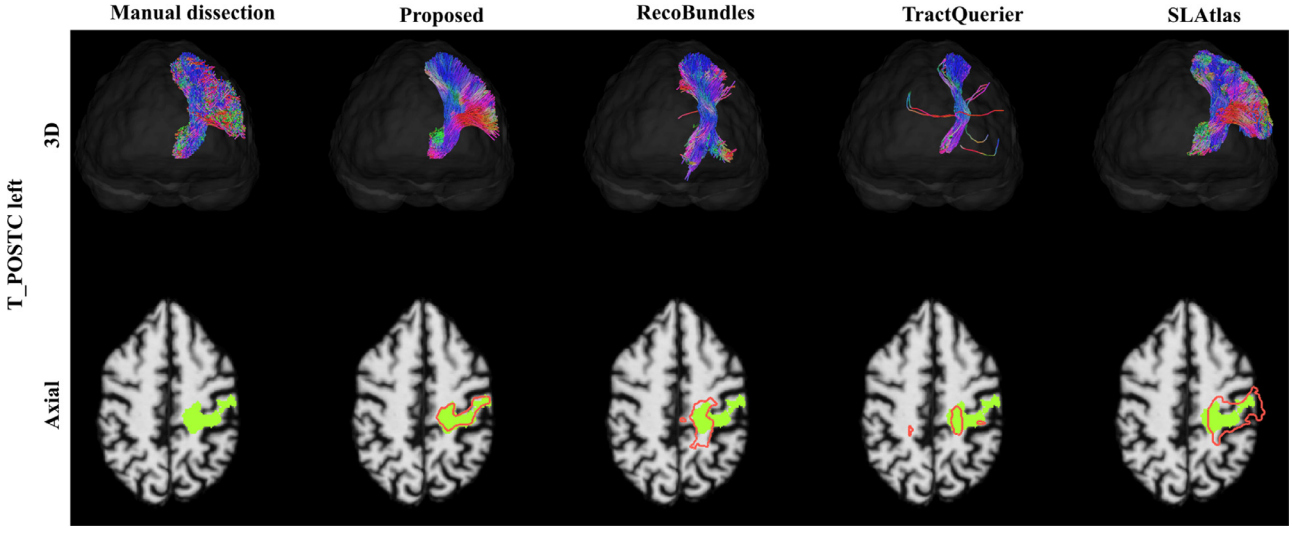

Fig. 11. Qualitative comparison of results on one subject with brain volume loss and schizoaffective disorder: reconstruction of left Thalamo-postcentral tract (T_POSTC). Green shows the manual dissection and red shows the tract mask of the respective method. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

fibers in the postcentral gyrus. *RecoBundles* is missing major parts of the tract and *SLAtlas* is not able to adapt to the reduced brain volume leading to streamlines running outside of the postcentral gyrus.

We also tested our method on subjects with brain tumors. However, given the vast distortions a tumor can produce, it is unclear where exactly certain tracts run. Experts can only assess if a tract could be a plausible reconstruction not containing any obvious errors (e.g. running through the tumor). The tract reconstructions in tumor patients produced by our method were rated as plausible by an expert. However, given the difficulty of evaluating tracts in tumor cases we do not show any results.

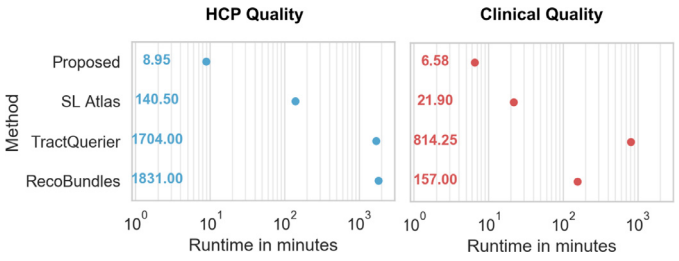

Fig. 12. Runtime of our proposed method and reference methods to generate bundle-specific trackings for all 72 tracts in one subject. On *HCP Quality* 10 million streamlines are used for the whole brain tracking, on *Clinical Quality* only 500,000 streamlines.

3.5. Runtime

Runtime experiments were performed using a server with 16 2 GHz Intel Xeon cores and an NVIDIA Titan X for the GPU-based approaches. We evaluated the runtime of all methods producing streamline output (*Proposed, RecoBundles, TractQuerier* and *SLAtlas*). The runtime does not include the fitting of the constrained spherical deconvolution model as this is identical for all methods. For the *HCP Quality* experiments whole brain tractograms with 10 million streamlines where used. For the *Clinical Quality* experiments whole brain tractograms with 500,000 streamlines were used. This reduced the runtime significantly. For our proposed approach on *HCP Quality* and *Clinical Quality* 2000 streamlines where generated for each tract. Fig. 12 shows the results for each method when reconstructing all 72 tracts in a previously unseen subject. Our method was 137x faster than the reference methods for *HCP Quality* and 50x faster for *Clinical Quality*.

4. Discussion and conclusion

4.1. Overview

Our proposed approach is a novel method for bundle-specific tractography. It was evaluated on 72 tracts in a cohort of 105 HCP subjects in original high quality and also on reduced quality, more clinical-like datasets. Moreover we evaluated the approach on synthetic software phantoms. Seven methods were used as a benchmark. Our experiments demonstrated that our approach achieves yet unprecedented accuracy and runtime while being less affected by the reduction in resolution in the clinical quality data. It also generalizes well to unseen datasets and pathologies.

4.2. Reference data

The tract delineations from the reference dataset used for training and evaluation do not represent a real ground truth. They are approximations based on diffusion weighted images, which has several limitations. However, given the high quality of the HCP data and the manual inspection by an expert, the employed dataset represents one of the best existing in vivo approximations of known white matter anatomy in a cohort of that size. Moreover, by using synthetic software phantoms we were able to evaluate our method on a dataset where the real ground truth is available.

On the *Phantom* data Dice scores were lower than on the *Clinical Quality* data. This had two main reasons: First the phantoms were simulated containing major artifacts whereas *Clinical Quality* contains only little artifacts as it based one the high quality HCP data. Second the domain shift between the training data and the *Phantom* data is significantly greater than the shift between the training data and the *Clinical Quality* data: On the one hand different acquisition settings were used during phantom simulation and on the other hand the simulated brains only contain the 72 reference tracts. Those cover the majority of the brain, but several tracts (like for example all u-fibers) were not included in the phantom. Therefore the training data and the phantoms are less similar and the resulting scores are reduced.

Despite the reduced Dice scores in comparison to the *Clinical Quality* data, our approach still shows complete reconstructions on the *Phantom* dataset with only minor errors (Fig. 16 in supplementary materials).

4.3. Reference methods

Selecting appropriate reference methods for a fair comparison was not easy as all methods have slightly different approaches and requirements. The comparison with our selected reference methods is also subject to some limitations: *TractQuerier* was part of the pipeline used for the creation of the reference dataset which was then used to evaluate *TractQuerier* against, thus inducing a potential positive bias for the method. For *RecoBundles* we were only able to use 5 reference subjects due to the long runtime of *RecoBundles*. Using all 63 subjects from the training set as reference subjects would have been computationally infeasible for 72 tracts and tractograms with 10 million streamlines. Moreover, as suggested by our *Atlas Custom* and *Multi-Mask* experiments, averaging more subjects, does not necessarily increase accuracy as small details become blurred. Using 5 reference subjects therefore provides a good estimation of the performance of *RecoBundles*. We used the default *RecoBundles* settings. Optimizing those might improve the results to some degree. *SLAtlas* is showing high angular errors because it is only based on affine registration making the registered tract not align properly on the new subject. *Peak Atlas* is based on elastic registration which leads to better alignment of tracts. However, it is also showing higher angular errors as elastic registration is still not able to completely resolve the complex inter-subject variability that exists between human brains.

As we have shown, our comparison to the reference methods has some limitations. However, those limitations do not apply to all reference methods and those limitations alone cannot explain the large accuracy gap between our method and all reference methods, indicating the great potential of the proposed method.

4.4. Preprocessing

The preprocessing of the reference data to extract start/end region segmentations and to determine the main streamline orientation in each voxel is made up of several non-trivial steps (see Section 2.4.1) containing a multitude of parameters which have to be chosen. The influence of these parameters on the final results was not systematically evaluated. Therefore changes to the parameters could significantly alter the final results.

4.5. Generalization

Our method is based on supervised learning, bearing the inherent limitation of depending on the availability and quality of training data. This is similar to most of the reference methods which also require reference tracts or atlases. Using scanners and acquisition sequences different from the training data introduces a domain shift and therefore reduces the performance. By using heavy data augmentation during training this domain shift can be reduced. Our experiments on the phantom data have shown that our method generalizes well to unseen acquisition sequences. We have also shown on a wide range of unseen datasets from different scanners with and without pathologies that the proposed method produces anatomically plausible results in most cases.

4.6. Code availability

The proposed method is openly available as an easy-to-use python package with pretrained weights: https://github.com/MIC-DKFZ/TractSeg/.

Declaration of Competing Interest

None.

Acknowledgments

HCP data were provided by the Human Connectome Project, WU- Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. Data used in preparation for this article were obtained from the SchizConnect database (http://schizconnect.org). As such, the investigators within SchizConnect contributed to the design and implementation of SchizConnect and/or provided data but did not participate in analysis or writing of this report. Data collection and sharing for this project was funded by NIMH cooperative agreement 1U01 MH097435. BrainGluSchi data was downloaded from the COllaborative Informatics and Neuroimaging Suite Data Exchange tool (COINS; http://coins.mrn.org/dx) and data collection was funded by NIMH R01MH084898- 01A1. COBRE data was downloaded from the COllaborative Informatics and Neuroimaging Suite Data Exchange tool (COINS; http://coins.mrn.org/dx) and data collection was performed at the Mind Research Network, and funded by a Center of Biomedical Research Excellence (COBRE) grant 5P20RR021938/P20GM103472 from the NIH to Dr. Vince Calhoun. This work was supported by the German Research Foundation (DFG) grant MA 6340/10-1 and grant MA 6340/12-1. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer 's Association; Alzheimer 's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer 's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. Data [in part] was collected at Brain and Mind centre, Sydney University and funded by NMSS and Novartis. Data were provided [in part] by OASIS-3: Principal Investigators: T. Benzinger, D. Marcus, J. Morris; NIH P50AG00561, P30NS09857781, P01AG026276, P01AG003991, R01AG043434, UL1TR000448, R01EB009352. AV-45 doses were provided by Avid Radiopharmaceuticals, a wholly owned subsidiary of Eli Lilly.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2019.101559.

References

Alexander, D.C., Pierpaoli, C., Basser, P.J., Gee, J.C., 2001. Spatial transformations of diffusion tensor magnetic resonance images. IEEE Trans. Med. Imag. 20 (11), 1131–1139. doi:10.1109/42.963816.

Alexander, D.C., Zikic, D., Ghosh, A., Tanno, R., Wottschel, V., Zhang, J., Kaden, E., Dyrby, T.B., Sotiropoulos, S.N., Zhang, H., Criminisi, A., 2017. Image quality transfer and applications in diffusion MRI. NeuroImage 152, 283–298. doi:10.1016/j.neuroimage.2017.02.089.

Andersson, J.L.R., Sotiropoulos, S.N., 2016. An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. NeuroImage 125, 1063–1078. doi:10.1016/j.neuroimage.2015.10.019.

Ankele, M., Lim, L.-H., Groeschel, S., Schultz, T., 2017. Versatile, robust, and efficient tractography with constrained higher-order tensor fodfs. Int. J. Comput. Assist.Radiol. Surgery 12 (8), 1257–1270. doi:10.1007/s11548-017-1593-6.

Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. 12 (1), 26–41. doi:10.1016/j.media.2007.06.004.

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

Behrens, T.E.J., Woolrich, M.W., Jenkinson, M., Johansen-Berg, H., Nunes, R.G., Clare, S., Matthews, P.M., Brady, J.M., Smith, S.M., 2003. Characterization and propagation of uncertainty in diffusion-weighted MR imaging. Magn. Reson. Med. 50 (5), 1077–1088. doi:10.1002/mrm.10609.

Bustillo, J.R., Jones, T., Chen, H., Lemke, N., Abbott, C., Qualls, C., Stromberg, S., Canive, J., Gasparovic, C., 2017. Glutamatergic and neuronal dysfunction in gray and white matter: a spectroscopic imaging study in a large schizophrenia sample. Schizophrenia Bull. 43 (3), 611–619. doi:10.1093/schbul/sbw122.

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage 31 (3), 968–980. doi:10.1016/j.neuroimage.2006.01.021.

Dumoulin, V., Visin, F., 2016. A guide to convolution arithmetic for deep learning. arXiv:1603.07285 [cs, stat].

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In: Kdd, 96, pp. 226–231.

Garyfallidis, E., Brett, M., Amirbekian, B., Rokem, A., Van Der Walt, S., Descoteaux, M., Nimmo-Smith, I., 2014. Dipy, a library for the analysis of diffusion MRI data. Front. Neuroinform. 8. doi:10.3389/fninf.2014.00008.

Garyfallidis, E., Côté, M.-A., Rheault, F., Sidhu, J., Hau, J., Petit, L., Fortin, D., Cunanne, S., Descoteaux, M., 2017. Recognition of white matter bundles using local and global streamline-based registration and clustering. NeuroImage doi:10.1016/j.neuroimage.2017.07.015.

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., Maier-Hein, K. H., 2018. Brain tumor segmentation and radiomics survival prediction: contribution to the BRATS 2017 challenge. arXiv:1802.10508 [cs].

Jenkinson, M., Beckmann, C.F., Behrens, T.E.J., Woolrich, M.W., Smith, S.M., 2012. FSL. NeuroImage 62 (2), 782–790. doi:10.1016/j.neuroimage.2011.09.015.

Jeurissen, B., Tournier, J.-D., Dhollander, T., Connelly, A., Sijbers, J., 2014. Multi-tissue constrained spherical deconvolution for improved analysis of multi-shell diffusion MRI data. NeuroImage 103, 411–426. doi:10.1016/j.neuroimage.2014.07.061.

Jones, E., Oliphant, T., Peterson, P., et al., 2001. SciPy: open source scientific tools for Python. [Online; accessed].

Kingma, D. P., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv:1412.6980 [cs].

Knösche, T.R., Anwander, A., Liptrot, M., Dyrby, T.B., 2015. Validation of tractography: comparison with manganese tracing. Human Brain Map. 36 (10), 4116–4134. doi:10.1002/hbm.22902.

Maier-Hein, K.H., Neher, P.F., Houde, J.-C., Côté, M.-A., Garyfallidis, E., Zhong, J., Chamberland, M., Yeh, F.-C., Lin, Y.-C., Ji, Q., Reddick, W.E., Glass, J.O., Chen, D.Q., Feng, Y., Gao, C., Wu, Y., Ma, J., Renjie, H., Li, Q., Westin, C.-F., Deslauriers-Gauthier, S., González, J.O.O., Paquette, M., St-Jean, S., Girard, G., Rheault, F., Sidhu, J., Tax, C.M.W., Guo, F., Mesri, H.Y., Dávid, S., Froeling, M., Heemskerk, A.M., Leemans, A., Boré, A., Pinsard, B., Bedetti, C., Desrosiers, M., Brambati, S., Doyon, J., Sarica, A., Vasta, R., Cerasa, A., Quattrone, A., Yeatman, J., Khan, A.R., Hodges, W., Alexander, S., Romascano, D., Barakovic, M., Auría, A., Esteban, O., Lemkaddem, A., Thiran, J.-P., Cetingul, H.E., Odry, B.L., Mailhe, B., Nadar, M.S., Pizzagalli, F., Prasad, G., Villalon-Reina, J.E., Galvis, J., Thompson, P.M., Requejo, F.D.S., Laguna, P.L., Lacerda, L.M., Barrett, R., Dell'Acqua, F., Catani, M., Petit, L., Caruyer, E., Daducci, A., Dyrby, T.B., Holland-Letz, T., Hilgetag, C.C., Stieltjes, B., Descoteaux, M., 2017. The challenge of mapping the human connectome based on diffusion tractography. Nature Commun. 8 (1), 1349. doi:10.1038/s41467-017-01285-x.

Mori, S., Crain, B.J., Chacko, V.P., van Zijl, P.C., 1999. Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging. Ann. Neurol. 45 (2), 265–269.

Nair, V., Hinton, G., 2010. Rectified linear units improve restricted Boltzmann machines. ICML.

Neher, P.F., Laun, F.B., Stieltjes, B., Maier-Hein, K.H., 2014. Fiberfox: facilitating the creation of realistic white matter software phantoms. Magn. Reson. Med. 72 (5), 1460–1470.

O'Donnell, L.J., Westin, C.-F., 2007. Automatic tractography segmentation using a high-dimensional white matter atlas. IEEE Trans. Med. Imag. 26 (11), 1562–1575. doi:10.1109/TMI.2007.906785.

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, W.M., Rathi, Y., Westin, C.-F., Golby, A.J., 2016. Automated white matter fiber tract identification in patients with brain tumors. NeuroImage : Clinical 13, 138–153. doi:10.1016/j.nicl.2016.11.023.

Raffelt, D., Tournier, J.-D., Fripp, J., Crozier, S., Connelly, A., Salvado, O., 2011. Symmetric diffeomorphic registration of fibre orientation distributions. NeuroImage 56 (3), 1171–1180. doi:10.1016/j.neuroimage.2011.02.014.

Rheault, F., St-Onge, E., Sidhu, J., Maier-Hein, K., Tzourio-Mazoyer, N., Petit, L., Descoteaux, M., 2018. Bundle-specific tractography with incorporated anatomical and orientational priors. NeuroImage 186, 382–398. doi:10.1016/j.neuroimage.2018.11.018.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: convolutional networks for biomedical image segmentation. arXiv:1505.04597 [cs].

Stieltjes, B., Brunner, R., Maier-Hein, K., Laun, F., 2013. Diffusion tensor imaging: introduction and atlas.

Taha, A.A., Hanbury, A., 2015. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. BMC Med. Imag. 15. doi:10.1186/s12880-015-0068-x.

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. doi:10.1016/j.neuroimage.2007.02.016.

Veraart, J., Novikov, D.S., Christiaens, D., Ades-aron, B., Sijbers, J., Fieremans, E., 2016. Denoising of diffusion MRI using random matrix theory. NeuroImage 142, 394–406. doi:10.1016/j.neuroimage.2016.08.016.

Wassermann, D., Makris, N., Rathi, Y., Shenton, M., Kikinis, R., Kubicki, M., Westin, C.-F., 2016. The white matter query language: a novel approach for describing human white matter anatomy. Brain Struct. Function 221 (9), 4705–4721.

Wasserthal, J., Neher, P., Maier-Hein, K., 2018. Tract Orientation Mapping for Bundle-Specific Tractography: 21st International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Granada, Spain, September 16–20, 2018, Proceedings, Part III, pp. 36–44. doi:10.1007/978-3-030-00931-1_5.

Wasserthal, J., Neher, P., Maier-Hein, K.H., 2018. Tractseg - fast and accurate white matter tract segmentation. NeuroImage doi:10.1016/j.neuroimage.2018.07.070.

Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biometrics Bull. 1 (6), 80–83. doi:10.2307/3001968.

Yendiki, A., Panneck, P., Srinivasan, P., Stevens, A., Zöllei, L., Augustinack, J., Wang, R., Salat, D., Ehrlich, S., Behrens, T., Jbabdi, S., Gollub, R., Fischl, B., 2011. Automated probabilistic reconstruction of white-matter pathways in health and disease using an atlas of the underlying anatomy. Front. Neuroinform. 5.

cicek, O., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3d U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016. Springer, Cham, pp. 424–432. doi:10.1007/978-3-319-46723-8_49.