NeuroImage 224 (2021) 117402

Contents lists available at ScienceDirect

NeuroImage

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

Classifyber, a robust streamline-based linear classifier for white matter bundle segmentation

Giulia Bertò[a,b], Daniel Bullock[c], Pietro Astolfi[a,b,d], Soichi Hayashi[c], Luca Zigiotto[f], Luciano Annicchiarico[f], Francesco Corsini[f], Alessandro De Benedictis[e], Silvio Sarubbo[f], Franco Pestilli[c], Paolo Avesani[a,b], Emanuele Olivetti[a,b,*]

[a] NeuroInformatics Laboratory (NILab), Bruno Kessler Foundation (FBK), Trento, Italy
[b] Center for Mind and Brain Sciences (CIMeC), University of Trento, Italy
[c] Department of Psychological and Brain Sciences, Indiana University, Bloomington, USA
[d] PAVIS, Italian Institute of Technology (IIT), Genova, Italy
[e] Neurosurgery Unit, Department of Neuroscience, Bambino Gesù Children's Hospital IRCCS, Rome, Italy
[f] Division of Neurosurgery, Structural and Functional Connectivity Lab, S. Chiara Hospital, Trento, Italy

ARTICLE INFO

Keywords:
White matter bundle segmentation
Supervised learning
Linear classification
Diffusion Magnetic Resonance Imaging (dMRI)

ABSTRACT

Virtual delineation of white matter bundles in the human brain is of paramount importance for multiple applications, such as pre-surgical planning and connectomics. A substantial body of literature is related to methods that automatically segment bundles from diffusion Magnetic Resonance Imaging (dMRI) data indirectly, by exploiting either the idea of connectivity between regions or the geometry of fiber paths obtained with tractography techniques, or, directly, through the information in volumetric data. Despite the remarkable improvement in automatic segmentation methods over the years, their segmentation quality is not yet satisfactory, especially when dealing with datasets with very diverse characteristics, such as different tracking methods, bundle sizes or data quality. In this work, we propose a novel, supervised streamline-based segmentation method, called Classifyber, which combines information from atlases, connectivity patterns, and the geometry of fiber paths into a simple linear model. With a wide range of experiments on multiple datasets that span from research to clinical domains, we show that Classifyber substantially improves the quality of segmentation as compared to other state-of-the-art methods and, more importantly, that it is robust across very diverse settings. We provide an implementation of the proposed method as open source code, as well as web service.

1. Introduction

Accurate delineation of anatomical structures in the human brain is essential to numerous scientific disciplines. In particular, white matter bundle segmentation can provide information to multiple applications, e.g. the characterization of neurodevelopmental disorders, pre-surgical planning, or connectomic studies (O'Donnell et al., 2017; Yeatman et al., 2012; Yeh et al., 2018).

In the last decade, several automatic methods for white matter bundle segmentation have been developed to mimic the manual segmentation done by expert neuroanatomists (Catani et al., 2002; Mori et al., 2005; Wakana et al., 2007), which is very time consuming and difficult to reproduce. Automatic methods can be divided into three main groups: (i) Connectivity-based, (ii) Streamline-based, and (iii) Direct.

Connectivity-based methods aim to extract bundles by filtering the entire set of streamlines with inclusion/exclusion Regions of Interest (ROIs) that the bundle is assumed to pass / not to pass through (Oishi et al., 2008; Wassermann et al., 2016; Yeatman et al., 2012; Yendiki et al., 2011; Zhang et al., 2010). These ROIs, which can be placed both in the cortex or in the white matter, frequently come from atlases that have to be registered into the individual subject space. A significant drawback to this approach is that the segmentation is inherently limited by the anatomical variability of the subjects and by the process of registration (Siless et al., 2020).

Streamline-based methods group together streamlines according to some similarity measure. *Unsupervised* streamline-based methods, such as those in Brun et al. (2004), Maddah et al. (2005), O'Donnell and Westin (2007), Guevara et al. (2012), Tunç et al. (2014), Siless et al. (2016), Siless et al. (2018), and Zhang et al. (2018), perform whole brain segmentation through clustering, without prior knowledge about the anatomy of the bundles and without leveraging examples of expert-made segmented bundles, limiting the quality of segmen-

* Corresponding author at: NeuroInformatics Laboratory (NILab), Bruno Kessler Foundation (FBK), Trento, Italy.
E-mail address: olivetti@fbk.eu (E. Olivetti).

https://doi.org/10.1016/j.neuroimage.2020.117402
Received 24 January 2020; Received in revised form 12 September 2020; Accepted 18 September 2020
Available online 23 September 2020

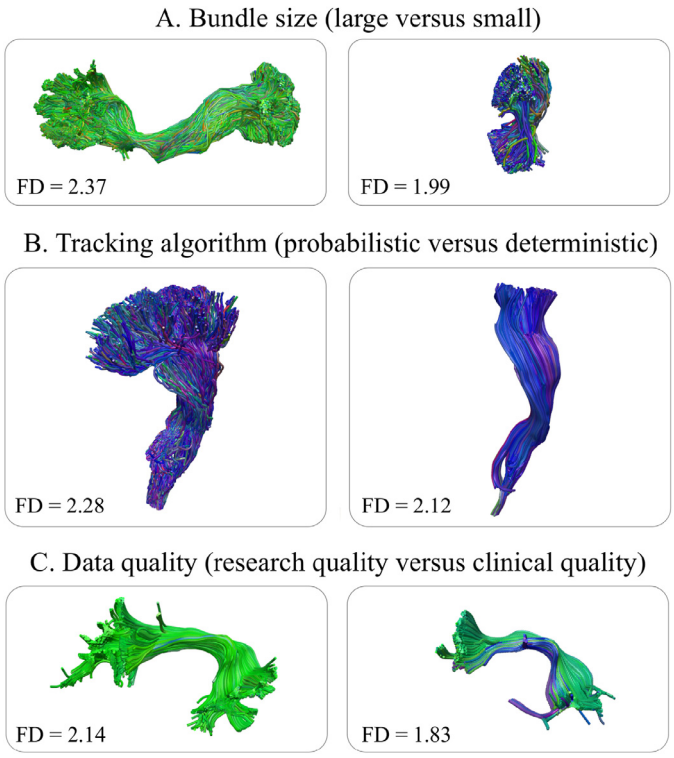

Fig. 1. Examples of different properties of bundles. A. Two bundles with different size, on the left a large bundle (inferior-fronto-occipital fascicle) and on the right a small bundle (posterior arcuate). B. Two bundles (corticospinal tracts) obtained using different tracking algorithms, on the left with probabilistic and on the right with deterministic tracking. C. Two bundles (arcuate fascicles) segmented from diffusion data of different quality, on the left at research quality and on the right at clinical quality. In each panel it is reported the fractal dimension (FD) of the voxel mask of the respective bundle.

tation. In contrast, *supervised* streamline-based methods require one or more examples of the bundle to learn from, in order to segment such bundle in the target subject, such as those in Mayer et al. (2011), Olivetti and Avesani (2011), Vercruysse et al. (2014), Yoo et al. (2015), Labra et al. (2016), Garyfallidis et al. (2018) and Sharmin et al. (2018). It has been shown that streamline-based methods like RecoBundles (Garyfallidis et al., 2018) and the one based on the Linear Assignment Problem, referred to as LAP (Sharmin et al., 2018), outperform connectivity based methods in terms of quality of segmented bundles.

Direct methods are voxel-based methods that segment bundles directly from diffusion images without the need for streamlines, see Wasserthal et al. (2018a) for a brief review. In contrast to the limited quality of segmentation reached by these methods, a recent direct method proposed in Wasserthal et al. (2018a) presented evidence of remarkably better segmentation quality in comparison with a large selection of other segmentation methods, including connectivity-based and streamline-based methods. This method, called TractSeg, is based on convolutional neural networks (Ronneberger et al., 2015) and has set the new standard in terms of quality of bundle segmentation.

Despite the remarkable improvement in automatic segmentation methods over the years, the resulting bundles can be unsatisfactory. The quality of segmentation may be strongly affected by some properties of the bundles, for example by their size; by the tractography technique, e.g. probabilistic or deterministic tracking algorithm; or by the data quality, e.g. research (high-resolution) or clinical quality, see Fig. 1 for some examples.

As of today, no single method for bundle segmentation has been demonstrated to be robust, to bundle size, tracking method and data quality[1]. The choice of the most appropriate pipeline for tractography is not unequivocal, but rather is strongly affected by the quality of the available diffusion Magnetic Resonance Imaging (dMRI) data, and changes according to the specific application, depending on the desired level of sensitivity/specificity (Thomas et al., 2014). Similarly, even though the interest in large bundles is well established in multiple applications (Pestilli, 2018; Wandell, 2016), small and short bundles, which we here call *minor* bundles, have recently received increasing attention, see Guevara et al. (2011), Wu et al. (2016b), Guevara et al. (2017) and Bullock et al. (2019). For example, the relatively smaller bundles connecting the human dorsal and posterior cortices have been recently proven to be of great help in understanding how information flows in the human brain (Bullock et al., 2019; Sani et al., 2019; Wu et al., 2016b). For these reasons, we believe that automatic methods for white matter bundle segmentation must be able to maintain a high quality of results across different settings.

The main contribution of the present work is a novel method for bundle segmentation that is robust to all properties described in Fig. 1. We call the method *Classifyber*. Classifyber is a supervised streamline-based method, and is based on a linear classification model that predicts whether or not individual streamlines belong to the bundle of interest. It combines the current knowledge in bundle segmentation, exploiting both the similarity between streamlines, typical of streamline-based methods, and the anatomical information from ROIs, typical of connectivity-based methods. In contrast to state-of-the-art automatic segmentation methods, we claim that Classifyber is robust to different data settings.

As a second contribution, we present an extensive comparison between Classifyber and multiple other automatic bundle segmentation methods available in the literature, across a diverse set of conditions: major bundles vs minor bundles, different tractography techniques, and bundles from healthy subjects vs brain tumor patients. The results of these experiments support our claims that Classifyber is able to adapt to different data settings and sets a new standard with respect to the current literature by substantially improving the segmentation quality reached by other methods.

As a third contribution, we show that some segmentation methods are deeply affected by a geometrical property of the shape of the bundles: the *fractal dimension* (FD) (Esteban et al., 2007; Zhang et al., 2006). Bundles with high fractal dimension are in general larger, more rounded, and have a smooth shape. Alternatively, bundles with low fractal dimension are generally smaller, flattened, and have a less smooth shape, see Fig. 1. We observe that the tracking algorithm used to generate the tractography and the size of the bundles are among the main factors in the change of fractal dimension. The concept of the fractal dimension of a bundle is a key concept to discuss the experiments presented in this work.

This paper is structured as follows. In Section 2, we describe the proposed method, Classifyber. Then, in Section 3 we present the materials, which are composed of four different datasets and of the atlases used to derive the ROIs for the proposed method. In Section 4, we report the design and the results of a number of experiments that we conducted to verify our hypotheses. Finally, in Section 5, we discuss the results that suggest that practitioners should adopt the proposed Classifyber method as the leading standard for bundle segmentation.

2. Methods

Classifyber is a novel method that performs automatic bundle segmentation as a supervised learning problem, meaning that the algorithm

[1] With the term *robust* we refer to the ability of the method to perform consistently well across different data settings, e.g., with different bundle sizes, tractography techniques, or data quality, rather than across repetitions of the data acquisition, e.g., test-retest.

learns how to segment from expert-based examples. The name *Classifyber* is the linguistic blend of *Classify* and *fiber*, which explains the basic principle of its functioning: to classify whether or not a given streamline/fiber[2] belongs to the bundle of interest.

Below, we provide a formal description of Classifyber, from the basic concepts to the key element of the proposed method, i.e., the vectorial representation of a streamline that merges geometrical information typically used by streamline-based segmentation methods, and anatomical information, typically used by connectivity-based segmentation methods. Afterwards, we describe the details of training and testing Classifyber. We then briefly recall the most important aspects of other bundle segmentation methods that we include in the experimental comparison of Section 4.1, together with the evaluation procedure. We conclude the section by introducing the notion of the *fractal dimension* (FD) of a bundle, which will be used to discuss the experimental results in Section 5.

2.1. Basic concepts

A *streamline* $s = [\mathbf{x}_1, \dots, \mathbf{x}_n]$ is an ordered sequence of 3D points, $\mathbf{x}_i = [x_i, y_i, z_i] \in \mathbb{R}^3$, $i = 1 \dots n$, that approximates a group of axons with a similar path in the white matter of the brain. A *tractogram* T is the entire set of streamlines of the white matter of a brain: $T = \{s_1, \dots, s_M\}$, where M typically ranges from hundreds of thousands to several millions. A white matter *bundle*, $b \subset T$, is a subset of the tractogram with a specific anatomical meaning, such as the corticospinal tract.

Experts neuroanatomists manually segment a given bundle b in a tractogram adopting several strategies, which may comprise the definition of inclusion/exclusion ROIs to obtain the desired streamlines. From the point of view of an algorithm, a convenient way to model that segmentation process is to consider each streamline individually and to decide whether or not the streamline belongs to the bundle:

$$e(s) = \begin{cases} 1 & \text{if } s \in b \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $e(s)$ denotes the expert deciding on the streamline s.

2.2. Classifyber

Classifyber implements a classifier that accurately predicts whether or not a given streamline s belongs to the bundle b. In analogy to the previous work of Olivetti and Avesani (2011), in this work we propose a linear classifier method as core algorithm for Classifyber, for multiple reasons: it is extremely well known and easy to understand, it is very fast and requires minimal resources, software implementations are commonly available and, as opposed to non-linear methods, it can be interpreted. Generally, a linear classifier c takes as input a vector of real values $\mathbf{v} \in \mathbb{R}^d$ and returns its predicted category, i.e. 0 or 1:

$$c(\mathbf{v}) = \begin{cases} 1 & \text{if } a_1 v_1 + \dots + a_d v_d + a_0 > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where the weights of the linear classifier $a_0, \dots, a_d$ are estimated by minimizing the errors in classification on a training set (plus regularization terms that may differ between different algorithms).

In order to use a linear classifier on a streamline s, it is necessary to transform the streamline into a vector v which contains the necessary information for the task of bundle segmentation. In other words, we need to define an effective *feature space* to represent streamlines as vectors. This is a key step in the proposed method, where we extract geometrical and anatomical information from the streamline and create this vector.

2.3. Vectorial representation of a streamline

Given a streamline s, we compute 4 sets of values that, concatenated, create the proposed vectorial representation v of the streamline. The first two sets refer to the geometrical aspects of the streamline, typically exploited by streamline-based segmentation methods. The remaining two sets refer to connectivity and anatomical aspects of the bundle of interest respectively, which are the main focus of connectivity-based segmentation methods.

Streamline-based segmentation methods group together streamlines according to some similarity measures, or distances. Typical distances between two streamlines are the *minimum average direct flip* (d_{MDF}) distance or the *minimum average mean* (d_{MAM}) distance, which account for the respective positions and shapes of the two streamlines, see Garyfallidis et al. (2015) and Olivetti et al. (2017). Based on such concepts, an accurate and easy way to compute a vectorial representation of streamlines has been proposed in Olivetti et al. (2012) and since been used for multiple applications, like clustering, interactive segmentation and fast nearest-neighbor queries, see Olivetti et al. (2013), Porro-Muñoz et al. (2015) and Sharmin et al. (2016). The transformation from streamline to vector is built on the general concept of *dissimilarity representation*, initially proposed for pattern recognition problems, see for example the comprehensive (Pekalska and Duin, 2005). The dissimilarity representation for streamlines described by Olivetti et al. (2012), first requires the user to define a small group of prototypical streamlines of the tractogram[3], called *landmark* streamlines, $l_1, \dots, l_L$, that summarise the tractogram and acts as a reference system. Then, given a streamline s, the set of its distances from the landmarks is its vectorial representation: $v = [d(s, l_1), \dots, d(s, l_L)]$, where d is a streamline distance, like d_{MDF} or d_{MAM}. As shown in Olivetti et al. (2012) and in Porro-Muñoz et al. (2015), a vector v or this sort is an accurate vectorial representation of the streamline s.

In this work we propose a vectorial representation for streamlines that extends the one originally proposed in Olivetti et al. (2012). The first two sets of values are two dissimilarity representations based on different landmarks: the **first** one uses $L = 100$ landmarks taken *globally* from a whole tractogram, as in Olivetti et al. (2012); the **second** one is bundle-specific and uses $L = 100$ landmarks taken *locally* in the area of bundle of interest. Both the global and local landmarks are chosen in one random subject using the subset farthest first (SFF) policy, which provides a uniform coverage of the area of interest, as suggested in Olivetti et al. (2012). Notice that, since the set of landmarks act as a reference system, it has to be the same for all subjects.

The **third** set of values represents connectivity features and is, again, a dissimilarity representation but now focused on connectivity patterns instead of the shape of the streamline. The idea is that, if a streamline represents the anatomical connection between cortical areas at its endpoints, then two streamlines with neighboring endpoints represent the same pattern of anatomical connectivity and serve the same purpose. The dissimilarity representation of this third set of values is based on a recent streamline distance that we proposed in Bertò et al. (2019), which exploits only the endpoints of the streamline: given two streamlines s_A and s_B, whose endpoints are $\{\mathbf{x}_1^A, \mathbf{x}_{n_A}^A\} \in s_A$ and $\{\mathbf{x}_1^B, \mathbf{x}_{n_B}^B\} \in s_B$, their endpoint distance is simply the mean Euclidean distance of the corresponding endpoints:

$$\begin{aligned} &d_{\text{END}}(s_A, s_B) \\ &= \frac{1}{2}(\min(||\mathbf{x}_1^A - \mathbf{x}_1^B||_2, ||\mathbf{x}_1^A - \mathbf{x}_{n_B}^B||_2) + \min(||\mathbf{x}_{n_A}^A - \mathbf{x}_1^B||_2, ||\mathbf{x}_{n_A}^A - \mathbf{x}_{n_B}^B||_2)) \end{aligned} \tag{3}$$

In this work, we propose to use this endpoint distance from the $L = 100$ global landmarks as the third set of values to describe the *connectivity* pattern of a streamline.

The **fourth** set of values refers to anatomical aspects of the bundle of interest, by means of the ROIs that define that bundle. Often, a bundle is defined by two ROIs that define its trajectory before it diverges towards the cortex, see for example (Wakana et al., 2007; Yeatman et al., 2012),

[2] In some literature, the name *fiber* refers to *axon* and in other literature to *streamline*. Here we refer to the latter for linguistic convenience.

[3] Such streamlines can be just a random subset of the tractogram.

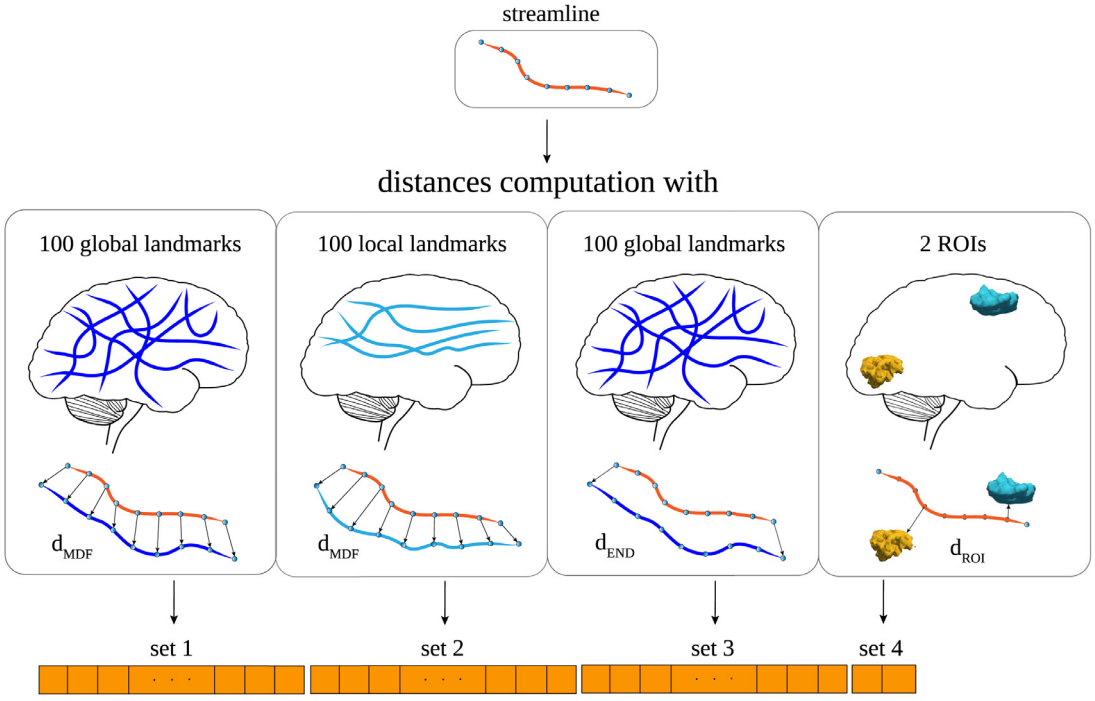

Fig. 2. Feature definition and extraction. Set 1 and set 2 contain the distances (d_{MDF}) of the streamline with 100 global and 100 local landmarks, respectively. Set 3 contains the distances (d_{END}) between the endpoints of the streamline and the 100 global landmarks. Set 4 contains the distances (d_{ROI}) between the streamline and the two ROIs pertaining to the bundle of interest.

or by the two terminal ROIs, see for example (Bullock et al., 2019). In Bertò et al. (2019), we recently proposed a streamline-ROI distance as a closest point distance: given a streamline s and one ROI represented as a voxel mask $\text{ROI} = \{\text{vox}_1, \dots, \text{vox}_M\}$, their distance is the minimum among all Euclidean distances between the points of the streamline and the voxels of the ROI:

$$d_{\text{ROI}}(s, \text{ROI}) = \min_{\mathbf{x} \in s, \text{vox} \in \text{ROI}} ||\mathbf{x} - \text{vox}||_2 \tag{4}$$

where with vox we indicate the coordinates of the center of the voxel. We use this distance to define the fourth set of values, i.e., the set of distances of the streamline s to each of the two ROIs that define the bundle.

In conclusion, given a streamline s, we compute 100 values as the dissimilarity representation from *global* landmarks (**set 1**), then 100 values as the dissimilarity representation from *local* landmarks (**set 2**), 100 values as the *endpoint* distance from global landmarks (**set 3**) and 2 values as the Euclidean distance from the 2 ROIs (or more than 2 values in case of more than 2 ROIs) relevant to the bundle of interest (**set 4**). The vector v resulting from concatenating those 302 values is the proposed vectorial representation of the streamline and these 302 variables define the proposed feature space. An illustration of the proposed feature space is given in Fig. 2.

2.4. Classifyber: training and test

Given tractograms and bundles segmented by experts, we first transform all streamlines into vectors and label them with 1 or 0, to indicate whether or not they belong to the bundle of interest, and then train a classifier to segment a specific bundle, e.g. the corticospinal tract (CST). Notice that, in order to segment different kinds of bundles, it is necessary to train different instances of Classifyber, each with a set of examples of the desired kind of bundle. Afterwards, given a tractogram of a new subject, we predict which of its streamlines belong to that bundle by first transforming them into vectors and then by applying the previously trained classifier to them. This procedure, divided into a *training phase* and a *test phase*, is described below and illustrated in Fig. 3.

2.4.1. Training phase

The training phase is composed of three steps, which are schematically illustrated in Fig. 3 (A).

Step (a1) *Bundle superset*. The entire set of streamlines in each tractogram is reduced to a subset of those proximal to the bundle of interest. The main purpose of this reduction is to avoid extremely *imbalanced data*, which decreases the accuracy of classification. Typically, the ratio between the number of streamlines of a bundle (class 1) and all the other streamlines in the tractogram (class 0) is around 1: 500, so extremely imbalanced. A typical simple technique to promote effective training is to remove examples far away from the boundary between the two classes and to get a more even class ratio. Specifically, the bundle superset of an example bundle is computed by considering the neighboring streamlines belonging to the corresponding tractogram retrieved by a k nearest neighbors (k-NN) procedure applied to each streamline of the bundle. We found $k = 2000$ to be a good compromise between computational cost reduction and size of the resulting superset with respect to the bundle and the tractogram[4]. Such operation is computationally intensive, but we adopted the very fast solution described in Sharmin et al. (2018). Moreover, this extra cost in time is massively outweighed by the 20x gain in time when computing the next steps, i.e. steps (a2) and (a3), see Section 4.2.7 for more details.

Step (a2) *Feature extraction*. Each streamline of the superset is then transformed into a vector, as described in Section 2.3. To the vector is assigned a class label 1 if it belongs to the bundle, 0 otherwise, see Fig. 3 (A), where they are represented in green and red respectively. The entire set of vectors, i.e. the *training set*, is z-scored independently for each feature.

Step (a3) *Training*. A binary Logistic Regression classifier is trained, using the stochastic average gradient (SAG) solver (Schmidt et al., 2017) available in the Python package scikit-learn (Pedregosa et al., 2011). We

[4] Usually, with $k = 2000$, the bundle superset, which is a subset of the entire tractogram, is approximately 30 times bigger that the bundle and 20 times smaller than the whole brain tractogram.

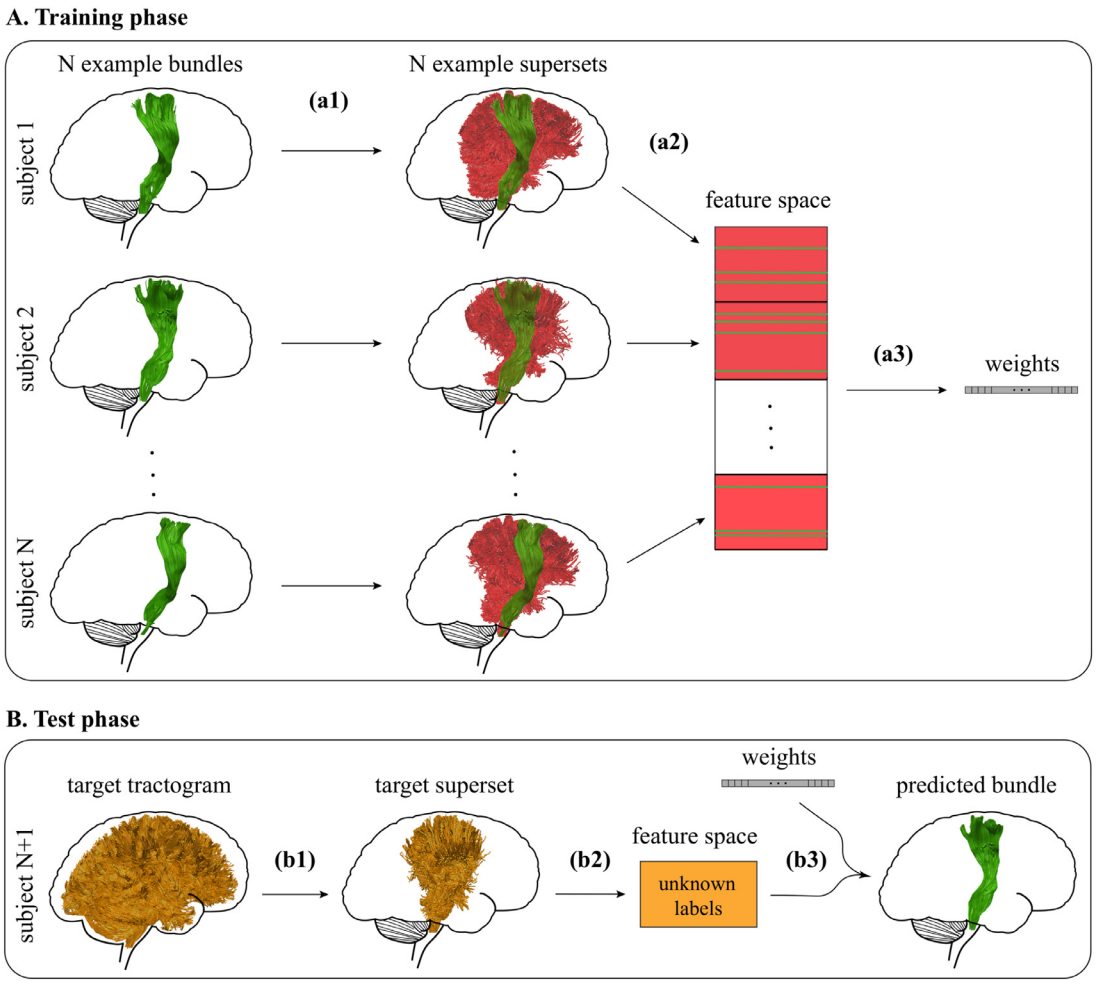

Fig. 3. Training and test of Classifyber. A. Schematic illustration of the training phase of Classifyber for a given bundle (CST) over *N* different subjects. Step (a1): bundle superset. Streamlines belonging to the bundle are depicted in green (class 1), while those not belonging to the bundle are depicted in red (class 0). Step (a2): feature extraction. Streamlines are transformed into vectors. Step (a3): training of a linear Logistic Regression (LR) classifier. The outcome of this phase is a vector of weights. B. Schematic illustration of the test phase of Classifyber on a single target subject. Step (b1): bundle superset. All the streamlines are depicted in orange because the labels are unknown. Step (b2): feature extraction. Streamlines are transformed into vectors. Step (b3): test using the resulting weights of the training phase. The outcome of this phase is the predicted bundle (CST) in the target subject.

use default parameters, except for the number of iterations of the solver, which we increase to 1000 to ensure convergence, as well as the parameter to lessen the negative effects of the residual class imbalance, which we set in all cases to 1:3. These choices are the result of a preliminary investigation on left out data and are kept for all the experiments.

2.4.2. Test phase

The test phase is performed on one subject of the test set at the time, called the *target subject*. Similarly to the training phase, the test phase comprises three steps, which are schematically illustrated in Fig. 3 (B).

Step (b1) *Bundle superset*. Similarly to step (a1) of the training phase, we reduce the whole target tractogram to a superset of the target bundle, mainly to decrease the computational cost of segmenting the target bundle. Obviously, in this case we do not know the target bundle in advance, so the superset is only *expected* to contain the target bundle, with very high probability. In this case, first, a candidate bundle superset is computed as in step (a1) but considering, in the target tractogram, the neighboring streamlines of one of the *example* bundle. This procedure is repeated using 5 of the example bundles. Then, the final bundle superset is obtained as the union of all the candidate bundle supersets[5].

Step (b2) *Feature extraction*. Each streamline of the bundle superset is embedded into a vector, as described in Section 2.3. All the vectors are z-scored feature-by-feature using means and standard deviations obtained in step (a2) of the training phase.

Step (b3) *Test*. By exploiting the linear classifier obtained from the training phase in step (a3), each streamline of the superset is predicted to be either part of the bundle (class 1) or not (class 0).

2.5. Other bundle segmentation methods

In Section 4.1 we compare Classifyber to state-of-the-art automatic segmentation methods. We selected two methods based on the recent extensive comparison presented in Wasserthal et al. (2018a), where *TractSeg* obtained the highest quality of bundle segmentation and *RecoBundles* ranked as the second best method among those freely available. In our comparison we also included *LAP*, see Sharmin et al. (2018), because it

[5] Retrospectively, in all experiments, the superset obtained in this way was approximately 40 times larger than the target bundle and always containing all the streamlines of the target bundle.

Table 1
Where to find the code and web apps of the methods considered in this work.

code / web app	web link	contribution of this work
Classifyber code	https://github.com/FBK-NILab/app-classifyber	yes (original)
Classifyber web app	https://doi.org/10.25663/brainlife.app.228	yes (original)
	https://doi.org/10.25663/brainlife.app.265	yes (original)
TractSeg code	https://github.com/MIC-DKFZ/TractSeg	no, already available
TractSeg web app	https://doi.org/10.25663/brainlife.app.186	no, already available
TractSeg-retrained web app	https://doi.org/10.25663/brainlife.app.204	yes (adapted)
	https://doi.org/10.25663/brainlife.app.205	yes (adapted)
RecoBundles(-atlas) code	http://nipy.org/dipy	no, already available
LAP web app	https://doi.org/10.25663/brainlife.app.209	yes (adapted)
Box-counting dimension	https://github.com/FBK-NILab/fractal_dimension	yes (original)

was not compared in Wasserthal et al. (2018a) but proved to be superior to nearest neighbor methods, the category to which RecoBundles belongs. In some cases, we used variants of TractSeg and RecoBundles, referred to as *TractSeg-retrained* and *RecoBundles-atlas*. We provide details on these other segmentation methods in Section 4.1.2.

2.6. Evaluation procedure

To quantitatively evaluate the performance of the different segmentation methods we use a procedure commonly adopted in this literature, see for example Garyfallidis et al. (2018), Sharmin et al. (2016) and Wasserthal et al. (2018a). We compute the degree of voxel overlap between the automatically segmented bundle $\hat{b}$ and the expert-based segmented bundle b, through the Dice Similarity Coefficient (DSC) (Dice, 1945): $DSC = 2 \cdot (|v(\hat{b}) \cap v(b)|)/(|v(\hat{b})| + |v(b)|)$ where $|v()|$ is the number of voxels in the bundle mask. The DSC ranges from 0 to 1 and the closer the score is to 1, the more the two bundles $\hat{b}$ and b are similar. The evaluation is conducted in the subject's native space.

2.7. Fractal dimension

The concept of *fractal dimension* (FD) (Mandelbrot, 1982) can be used to quantify the degree of irregularity of a 3D shape. This notion has already been applied to the shape of the brain white matter (Zhang et al., 2006) and to characterize multiple sclerosis (Esteban et al., 2007).

Intuitively, for standard objects like straight lines, a 2D flat square or a 3D cube, the FD is 1, 2 and 3, respectively. Irregular lines can have FD greater than 1 and asymptotically 2, if their resulting shape is close to a 2D surface. In the same way, a convoluted 2D shape that resembles a 3D shape, or a 3D shape with several holes, both have FD between 2 and 3. For example, Zhang et al. (2006) estimated the FD of the 3D voxel mask of the white matter of human brains and obtained values between 2.1 and 2.5.

In this work, we determine the FD of the voxel mask of white matter bundles via the box-counting dimension, see Falconer (2014). The box-counting dimension is based on the idea of covering a given shape with boxes of size σ and it quantifies how the number of boxes changes when σ changes, in double-log scale:

$$\mathrm{FD}_{\mathrm{box}} = -\lim_{\sigma \to 0} \frac{\log \mathrm{count}(\sigma)}{\log \sigma} \tag{5}$$

where $\mathrm{count}(\sigma)$ is the number of the necessary boxes. As an example, see the FD of some bundles in Figure 1.

2.8. Code availability

We provide the source code of Classifyber and the code to estimate the box-counting dimension (with examples) as open source software, see Table 1. Moreover, Classifyber can be freely used as web application on the online platform brainlife.io. In Table 1 we list the web links related to all the implementations of the bundle segmentation methods considered in this work.

3. Materials

In order to test different automatic bundle segmentation methods across a wide range of settings, we conducted extensive experiments across four different datasets of tractograms and bundles, three of which are novel. The description of these datasets, which we denote as HCP-minor, HCP-IFOF, HCP-major and Clinical, is provided in the following sections, together with the atlases used to derive the ROIs for the proposed method.

3.1. Data sources

The first three datasets are built on top of diffusion data freely available from the Human Connectome Project (HCP) (Sotiropoulos et al., 2013; Van Essen et al., 2013), 3T scanner, image resolution of 1.25 mm isotropic, 270 gradient directions with b-values=1000, 2000, and 3000 s/mm^2 and 18 volumes with b=0. Data have already been preprocessed with the minimal pipeline of Glasser et al. (2013), which includes brain extraction and correction for motion, distortion and eddy-currents. The fourth dataset is an in-house clinical dataset built from patients with brain tumors, 1.5T scanner, image resolution 0.9 x 0.9 x 2.4 mm, 60 gradient directions with b-value=1000 s/mm^2 and 1 volume with b=0. Data were corrected for eddy-current and motion, and an additional step of rescaling was applied to obtain an isotropic voxel resolution of 2 x 2 x 2 mm.

3.2. Datasets of tractograms and expert-based segmented bundles

(i) **HCP-minor**. *Number of subjects:* 105 from HCP. *Tractography:* 90 directions, single shell b=2000 s/mm^2, constraint spherical deconvolution (CSD), ensemble probabilistic tracking (Takemura et al., 2016) with curvature parameters=0.25, 0.5, 1, 2 and 4 mm, step size=0.625 mm, 750K streamlines. *Bundles:* Left and right posterior arcuate fasciculus (Left_pArc and Right_pArc), left and right temporo-parietal connection to the superior parietal lobule (Left_TP-SPL and Right_TP-SPL), left and right middle longitudinal fasciculus-superior parietal lobule component (Left_MdLF-SPL and Right_MdLF-SPL), left and right middle longitudinal fasciculus-superior angular gyrus component (Left_ MdLF-Ang and Right_MdLF-Ang). *Expert-based segmentations:* We obtained the segmentations of 192 randomly selected HCP subjects using the procedure proposed in Bullock et al. (2019). We then filtered out segmented bundles that were not considered plausible from the neuroanatomical point of view with a semi-automatic technique, as described in Appendix A, remaining with 105 subjects.

(ii) **HCP-IFOF**. *Number of subjects:* 30 from HCP. *Tractography:* 90 directions, single shell b=2000 s/mm^2, constraint spherical deconvolution (CSD), deterministic local tracking (Berman et al., 2008; Garyfallidis et al., 2014), step size=0.625 mm, white matter seeding, approximately 500K streamlines. *Bundles:* Left and right inferior fronto-occipital fasciculus (Left_IFOF and Right_IFOF). *Expert-based segmentations:* One expert neurosurgeon (A.D.B.) manually seg-

mented the bundles in 30 random HCP subjects following the guidelines in Sarubbo et al. (2013) and Hau et al. (2016), who proposed a classification of the IFOF in different subcomponents based on microdissection studies. Specifically, the bundle is composed of two layers: the first layer is superficial and antero-superiorly directed, with terminations in the inferior frontal gyrus, while the second layer is deeper and consists of three components (anterior, middle and posterior).

(iii) **HCP-major**. *Number of subjects:* 105 from HCP. *Tractography:* 270 directions, multi-shell multi-tissue (msmt) constraint spherical deconvolution (CSD), iFOD2 probabilistic anatomically constrained tractography (ACT), variable step size, white matter seeding, 10 million streamlines. *Bundles:* Left and right corticospinal tract (Left_CST and Right_CST), left and right inferior fronto-occipital fasciculus (Left_IFOF and Right_IFOF), left and right inferior longitudinal fasciculus (Left_ILF and Right_ILF), left and right uncinate fasciculus (Left_UF and Right_UF), left and right arcuate fasciculus (Left_AF and Right_AF). *Expert-based segmentations:* We considered a portion of the semi-automatically segmented bundles from the freely available benchmark dataset of Wasserthal et al. (2018a) available at Wasserthal et al. (2018)).

(iii) **Clinical**. *Number of patients:* 10 with brain tumor. *Tractography:* 60 directions, single shell b=1000 s/mm^2, diffusion tensor imaging (DTI) reconstruction, Euler Delta Crossing (EuDX) tracking method (Garyfallidis et al., 2014), 10^6 seeds, approximately 100K streamlines. *Bundles:* Left inferior fronto-occipital fasciculus (Left_IFOF) and left arcuate fasciculus (Left_AF). *Expert-based segmentations:* One expert neurosurgeon (S.S.) manually segmented the bundles in the lesioned hemisphere of the patients, who were affected by brain tumors. The lesion however did not affect the shape of the bundles consistently. Bundles where successively refined to remove outliers using the interactive segmentation tool Tractome (Porro-Muñoz et al., 2015) and visually inspected, remaining with 7 instances for each bundle.

3.3. Atlases

We exploited the following freely available atlases in order to derive the ROIs used by Classifyber, which were then registered to the MNI152 T1 template (Mazziotta et al., 2001).

MNI152_ICBM2009c_reconstructed_atlas. This atlas is a curated FreeSurfer parcellation of the ICBM2009c nonlinear asymmetric template, see Fischl (2012), Fonov et al. (2011), and Gorgolewski (2016). The parcellation is used to define the terminal regions of bundles in HCP-minor dataset.

MNI_JHU_tracts_ROIs_atlas. This atlas is composed of two planar waypoint ROIs for each of 20 major bundles, which delineate the path of each bundle before it diverges towards the cortex. Each ROI was drawn on a group-average dataset in MNI space, see Wakana et al. (2007). This atlas is used to define the waypoint ROIs of the bundles in HCP-major, HCP-IFOF and Clinical datasets.

3.4. Data preprocessing

For the three HCP datasets, we computed the non-linear warp to register the structural T1-weighted images of every subject of each dataset to the MNI152 T1 template using the Advanced Normalization Tool (ANTs) (Avants et al., 2008). For the clinical dataset, we computed a streamline linear registration (SLR) to the whole brain template of Yeh et al. (2018) (available at Garyfallidis (2018)) because non-linear registration of clinical data is debated, as reported in Garyfallidis et al. (2015). In all cases, we applied the registrations to tractograms and bundles.

3.5. Data availability

We freely share tractograms and expert-based segmented bundles of the HCP-minor dataset through the brainlife.io platform at https://doi.org/10.25663/brainlife.pub.11. The HCP-major dataset is available at Wasserthal et al. (2018). The HCP-IFOF is available upon formal data sharing agreement with the authors. The access to the Clinical dataset is limited by ethical and privacy issues and requires formal agreement with the neurosurgery unit involved in this study.

4. Experiments and results

4.1. Experiments

The experiments were conducted on the four datasets described in Section 3: HCP-minor, HCP-major, HCP-IFOF and Clinical. For each dataset, the entire pool of subjects was randomly divided into two groups: the *training set* and the *test set*. Bundles of the training set were used as examples to learn from, while bundles of the test set were used to assess the performance of the different methods. Notice that the exact same test sets were kept for all the methods compared. In this way, we could compare both the quality of segmentation obtained by each method averaged over the pool of test subjects, such as in an *unpaired* test, and the subject-by-subject comparison in segmenting each bundle, such as in a *paired* test, e.g. how frequently one method obtained better quality of segmentation than another method.

4.1.1. Classifyber: experimental setup

We retrieved the ROIs pertaining to each bundles in order to build the feature space of Classifyber, using the available atlases described in Section 3.3. For the dataset HCP-minor, the two ROIs considered for each bundle are the two terminal ROIs, i.e. the cortical regions that the bundle of interest connects, derived from Bullock et al. (2019). Specifically, the MdLF-Ang and MdLF-SPL connect the parietal region to the lateral-temporal region, while the TP-SPL and pArc connect the parietal region to the temporal region. Each region was built by merging specific cortical parcellations of the *MNI152_ICBM2009c_reconstructed_atlas*. For the other three datasets, the ROIs considered are the two planar waypoint ROIs defined in the *MNI_JHU_tracts_ROIs_atlas*, see Wakana et al. (2007).

HCP-minor We considered only subjects for which all bundles received an expert-made score of 3 or higher, according to the procedure explained in Appendix A, resulting in a set of 40 subjects. We randomly split this pool of subjects into a group of 15 for training and a group of 25 for testing. Additionally, within this dataset, we also studied how much the quality of segmentation of Classifyber was affected when changing the number of subjects in the training set from 1 to 60. In this case, we considered also subjects for which all bundles received an expert-made score of at least of 2.

HCP-IFOF We randomly split the pool of subjects into a group of 15 for training and a group of 15 for testing.

HCP-major For this dataset, which is part of the dataset used in Wasserthal et al. (2018a), we selected the same 21 test subjects used in the experiments presented there. In this way, we could directly compare our new results on the major bundles with theirs and, at the same time, we could test the reproducibility of their results. Of the 84 remaining subjects, 15 were randomly selected and used as training set for Classifyber. The kinds of bundles considered are those for which the two waypoint ROIs are available in the *MNI_JHU_tracts_ROIs_atlas*. In preliminary experiments, we observed that the 10 million streamlines of each tractogram in HCP-major were extremely redundant for training Classifyber and just using 10% of them, randomly selected, did not significantly change the results. By using just 10% of the streamlines we reduced the training time by a factor of 10 and the RAM usage by a factor of 4.

Table 2
Summary of the experimental setup of all the segmentation methods and variants, across the experiments/datasets considered in this work. For each dataset, all *training* sets are identical across the methods, except when indicated with a letter. Explanation of *a*: trained also on 15 subjects from HCP-IFOF and 15 subjects from HCP-major; *b*: 84 subjects/bundles from HCP-major; *c*: exact same training set as the other methods composed of 15 bundles, but considered individually; *d*: 1 example bundle from the bundle atlas. UNSUPPORTED: TractSeg and RecoBundles-atlas cannot segment minor bundles. UNFEASIBLE: RecoBundles and LAP required too many computational resources. Notice that in all cases, all *test* sets are identical across all methods.

Method	Description	Experiment / Dataset (# subjects in training set)			
		HCP-minor	HCP-IFOF	HCP-major	Clinical
Classifyber	linear classifier of single streamlines	15	15	15	6^a
TractSeg	voxel-based CNNs, *pre-trained*	UNSUPPORTED	84^b	84^b	84^b
TractSeg-retrained	voxel based CNNs, *retrained*	15	15	84^b	–
RecoBundles	1-NN of single streamlines	15^c	15^c	UNFEASIBLE	–
RecoBundles-atlas	1-NN of single streamlines, from atlas	UNSUPPORTED	1^d	1^d	–
LAP	linear assignment of single streamlines	15	15	UNFEASIBLE	–

Clinical Due to the small number of subjects in the dataset, instead of splitting the pool of 7 subjects into training and test sets, we ran Classifyber in two different ways: (i) we trained Classifyber on the IFOFs and AFs of the HCP-major dataset and then segmented the 7 patients in the Clinical dataset. We chose this dataset because it is part of the exact same dataset used for training TractSeg, to have fair comparison between the two methods. We refer to this case as *Classifyber-major*. (ii) We performed a cross-validation study with the leave-one-subject-out strategy, using *only* 6 subjects from the Clinical dataset as training set and the remaining subjects as test set, repeatedly. We refer to this case plainly as *Classifyber*. In this latter case we also aimed to show the ability of Classifyber to accurately segment bundles even when trained on a very small number of segmentations, in this case only 6. To conclude, for the IFOF, we ran one additional experiment where Classifyber was trained on the HCP-IFOF dataset. We refer to this case as *Classifyber-IFOF*.

4.1.2. State-of-the-art methods: experimental setup

Here we describe the state-of-the-art automatic segmentation methods that we considered in our comparison, their necessary variants to experiment on all datasets, and their experimental setup. A summary of the experimental setup of all the methods, variants, and datasets considered is given in Table 2.

TractSeg TractSeg, a voxel-based method recently proposed by Wasserthal et al. (2018a), is based on fully convolutional neural networks (FCNNs) and segments 72 bundles simultaneously. Its output are the voxel masks of the segmented bundles. We adopted the openly available pretrained network, which was trained on 84 subjects, and tested it on the dMRI data of the target subjects. Note that the pre-trained TractSeg, despite being very fast and easy to use, is expected to not perform well in some of the experimental settings because it was trained on bundles whose characteristics[6] may differ from the bundles to be segmented. We used the default parameters and the postprocessing option, which removes holes and isolated voxels in the predicted voxel mask of the bundles.

TractSeg-retrained When the bundle to be segmented was not available among those covered by TractSeg, we re-trained the FCNN on new examples with a procedure discussed with the authors of TractSeg and described in the following. We refer to this variant as *TractSeg-retrained*. First, we trained a single FCNNs per dataset with default parameters, 250 epochs, fraction of validation subjects = 0.2 and data augmentation. Then, we tested the method enabling the postprocessing option. For the HCP-minor dataset, we trained the model both with the same 15 subjects used in the other methods, and also with 69 additional subjects by considering as well those subjects for which all bundles received a score of at least 2 (84 subjects in total). We provide evidence of the successful training in Appendix B.

RecoBundles-atlas Garyfallidis et al. (2018) proposed a streamline-based segmentation method, called RecoBundles, that takes as input one example bundle which is used to estimate the corresponding bundle in a new tractogram by means of linear registration and nearest-neighboring streamlines. We contacted the authors of RecoBundles and received the indication to use the bundle models provided by the bundle atlas of Yeh et al. (2018) (available at Garyfallidis (2018)) as the example bundles and specifically 30 (out of 80) of them. We denote as *RecoBundles-atlas* this use of the RecoBundles algorithm. Note that this variant of RecoBundles, despite being very fast and easy to use, is expected to not perform well in some of the settings of the experiments, because it uses as input a single bundle model from an atlas whose characteristics[7] may differ from the bundles to be segmented. We used the best configuration of parameter values found from an extensive preliminary assessment analogous to the one reported in the supplementary materials. This configuration uses default parameter values with the exception of disabling the local streamline linear registration (SLR) option (because all the datasets were already coregistered in MNI space) and using the minimum average mean distance (d_{MAM}) instead of the minimum average direct flip distance (d_{MDF}).

RecoBundles When the bundle to be segmented is not available among the 30 selected bundles from the bundle atlas of Yeh et al. (2018), we fell back to the original indication in Garyfallidis et al. (2018) and used the same example bundles adopted as input for the other methods. We denote this use of the algorithm plainly as *RecoBundles*. Due to the fact that RecoBundles accepts only one bundle as example, to quantify the quality of segmentation when multiple bundles are available in the training set, we adopted a procedure similar to the one used in the experiments of Wasserthal et al. (2018a). Specifically, we treated the N example bundles as models for N separate runs of the algorithm over the target subject, thus obtaining N different predictions of the same bundle. We then evaluated the segmentation accuracy by computing the mean DSC across the N bundles. As for RecoBundles-atlas, we used the best configuration of parameter values found from an extensive preliminary assessment described in the supplementary materials.

LAP Sharmin et al. (2018) proposed a streamline-based segmentation method that takes as input multiple example bundles which are used to estimate the corresponding bundle in a target tractogram by means of finding corresponding streamlines through the solution of a Linear Assignment Problem (LAP) and a refinement step. We ran the algorithm following the original procedure and we set the parameter k, the only parameter of the method, corresponding to the number of nearest neighbors streamlines to compute the superset, equal to 2000 (default

[6] TractSeg was trained with bundles from dMRI data of the Human Connectome project, CSD reconstruction and probabilistic tracking, see Section 3.2.

[7] The atlas of Yeh et al. (2018) is based on dMRI data from the Human Connectome Project and deterministic tracking.

Table 3
Quantitative comparison over HCP-minor dataset: DSC (mean ± sd) across 25 target subjects for RecoBundles, TractSeg-retrained, LAP and Classifyber. Highest quality of segmentation in bold face.

	RecoBundles	TractSeg-ret.	LAP	Classifyber
Right_pArc	0.76 ± 0.04	0.77 ± 0.03	0.80 ± 0.03	**0.88** ± 0.03
Left_MdLF-Ang	0.71 ± 0.04	0.72 ± 0.06	0.79 ± 0.05	**0.87** ± 0.03
Left_pArc	0.73 ± 0.05	0.75 ± 0.03	0.79 ± 0.04	**0.85** ± 0.05
Right_MdLF-Ang	0.68 ± 0.04	0.70 ± 0.03	0.76 ± 0.03	**0.84** ± 0.03
Left_MdLF-SPL	0.63 ± 0.06	0.67 ± 0.05	0.73 ± 0.04	**0.82** ± 0.04
Right_TP-SPL	0.62 ± 0.08	0.68 ± 0.06	0.72 ± 0.05	**0.82** ± 0.05
Left_TP-SPL	0.63 ± 0.06	0.67 ± 0.04	0.70 ± 0.05	**0.81** ± 0.04
Right_MdLF-SPL	0.60 ± 0.05	0.64 ± 0.04	0.70 ± 0.03	**0.80** ± 0.04

$k = 500$), since the total number of streamlines of the tractograms considered in our experiments are approximately 4 times higher than in the original study of Sharmin et al. (2018). One limitation of LAP is that it is computationally too expensive for supersets larger than 100 thousands streamlines, both for memory and time requirements.

4.1.3. Experiments on Fractal Dimension (FD)

In this experiment, we studied how the performance of the different segmentation methods is affected by the FD of the target bundles. We computed the FD of the voxel mask of each target bundle as segmented by experts and compared it with the quality of segmentation (DSC) obtained for that bundle by each automatic segmentation method, across all experiments (approximately 500 bundles). For TractSeg and RecoBundles, that number was larger because we investigated also the variants TractSeg-retrained and RecoBundles-atlas, while for LAP it was smaller because it was not possible to execute the method on the HCP-major dataset, where supersets substantially exceeded 100 thousands streamlines.

4.2. Results

4.2.1. Results on HCP-minor dataset

In Table 3 and Fig. 7, we quantify the mean quality of segmentation in terms of DSC across the minor bundles considered in this set of experiments for RecoBundles, TractSeg-retrained, LAP and Classifyber across 25 subjects. TractSeg and RecoBundles-atlas were excluded because they do not address minor bundles. The quality of segmentation obtained by Classifyber is very high and outperforms all the other methods. Moreover, given that the target subjects are exactly the same across all methods, we can also summarize the results with a direct comparison on the individual bundles: over the 200 segmentations (8 different bundles for each of the 25 test subjects) performed by each method during the test phase, Classifyber obtained higher quality of segmentation (higher DSC) than RecoBundles and TractSeg-retrained in 100% of the cases (Wilcoxon signed-rank test, p-value=1.4×10^{-34} for both the comparisons), and than LAP in 99% of the cases[8] (Wilcoxon signed-rank test, p-value=1.7×10^{-34}).

[8] TractSeg-retrained, when trained on 84 subjects, performed better than TractSeg-retrained on 15 subjects, obtaining a marginal increase in DSC between 0 and 0.03. For a fair comparison with the other methods, this result is not reported in Table 3 and Fig. 7.

Table 5
Quantitative comparison over HCP-major dataset: DSC (mean ± sd) across 21 target subjects for RecoBundles-atlas, TractSeg and Classifyber. Highest quality of segmentation in bold face.

	RecoB.-atlas	TractSeg	Classifyber
Right_CST	0.62 ± 0.07	**0.85** ± 0.02	**0.87** ± 0.02
Left_CST	0.62 ± 0.11	**0.85** ± 0.03	**0.86** ± 0.10
Right_UF	0.57 ± 0.24	0.79 ± 0.03	**0.86** ± 0.03
Right_AF	0.53 ± 0.11	**0.83** ± 0.02	**0.86** ± 0.03
Left_UF	0.55 ± 0.27	0.77 ± 0.03	**0.84** ± 0.04
Left_IFOF	0.67 ± 0.06	0.80 ± 0.02	**0.84** ± 0.03
Left_ILF	0.57 ± 0.07	0.77 ± 0.02	**0.84** ± 0.04
Right_IFOF	0.76 ± 0.04	0.80 ± 0.02	**0.84** ± 0.03
Left_AF	0.71 ± 0.05	**0.84** ± 0.03	**0.83** ± 0.04
Right_ILF	0.42 ± 0.13	0.75 ± 0.03	**0.82** ± 0.04

The superiority of Classifyber over the other segmentation methods is also evident from the qualitative comparison in Fig. 4, in which the segmentations provided by the proposed method are, for all the bundles considered, the most anatomically similar to the expert-based segmentations. When using other methods, we observe a consistent bias in the predictions: RecoBundles and LAP tend to overestimate the bundle producing several false positives streamlines. On the other hand, for the majority of the bundles of this dataset, TractSeg-retrained correctly identifies the core part of the bundles, but fails to retrieve part of the cortical terminations. Illustrative examples of this behavior are in the last row of Fig. 4, in which the Right_MdLF-SPL is overestimated by RecoBundles (first panel), and it is missing most of the terminations in the latero-temporal ROI by TractSeg-retrained (second panel).

4.2.2. Results on HCP-IFOF dataset

In Table 4 and Fig. 7 we report the result of comparing Classifyber with all other methods and variants: RecoBundles-atlas, RecoBundles, TractSeg, TractSeg-retrained and LAP. The average DSC across 15 subjects shows the superiority of Classifyber. Moreover, in all individual cases, i.e. the 30 segmented bundles of the test set, Classifyber always obtained the highest DSC as compared to all other methods (Wilcoxon signed-rank test, p-value=1.7×10^{-6} for all the comparisons).

Additionally, a qualitative visual comparison is reported in Fig. 5, which illustrates that the Left_IFOF estimated with RecoBunldes-atlas (first panel), is clearly missing the middle and posterior subcomponents with respect to the expert-based segmented bundle (last panel). A very similar behavior is observed in the bundle predicted by TractSeg (third panel).

4.2.3. Results on HCP-major dataset

In Table 5 and Fig. 7 we report the mean quality of segmentation as DSC for RecoBundles-atlas, TractSeg and Classifyber over the major bundles considered, across 21 subjects. Over the 210 individual segmentations generated by each method in the test phase, Classifyber obtained a higher DSC than RecoBundles-atlas in 99% of the cases (Wilcoxon signed-rank test, p-value=5.4×10^{-35}) and higher than TractSeg in 86% of the cases (Wilcoxon signed-rank test, p-value=3.2×10^{-27}).

Fig. 6 shows a qualitative comparison of two of the bundles segmented with the three different methods. It is visible that Classifyber

Table 4
Quantitative comparison over the HCP-IFOF dataset: DSC (mean ± sd) across 15 target subjects for RecoBundles-atlas, RecoBundles, TractSeg, TractSeg-retrained, LAP and Classifyber. Highest quality of segmentation in bold face.

	RecoBundles-atlas	RecoBundles	TractSeg	TractSeg-retrained	LAP	Classifyber
Left_IFOF	0.45 ± 0.14	0.80 ± 0.04	0.48 ± 0.04	0.61 ± 0.03	0.81 ± 0.04	**0.91** ± 0.03
Right_IFOF	0.62 ± 0.18	0.72 ± 0.06	0.41 ± 0.06	0.57 ± 0.04	0.73 ± 0.05	**0.89** ± 0.03

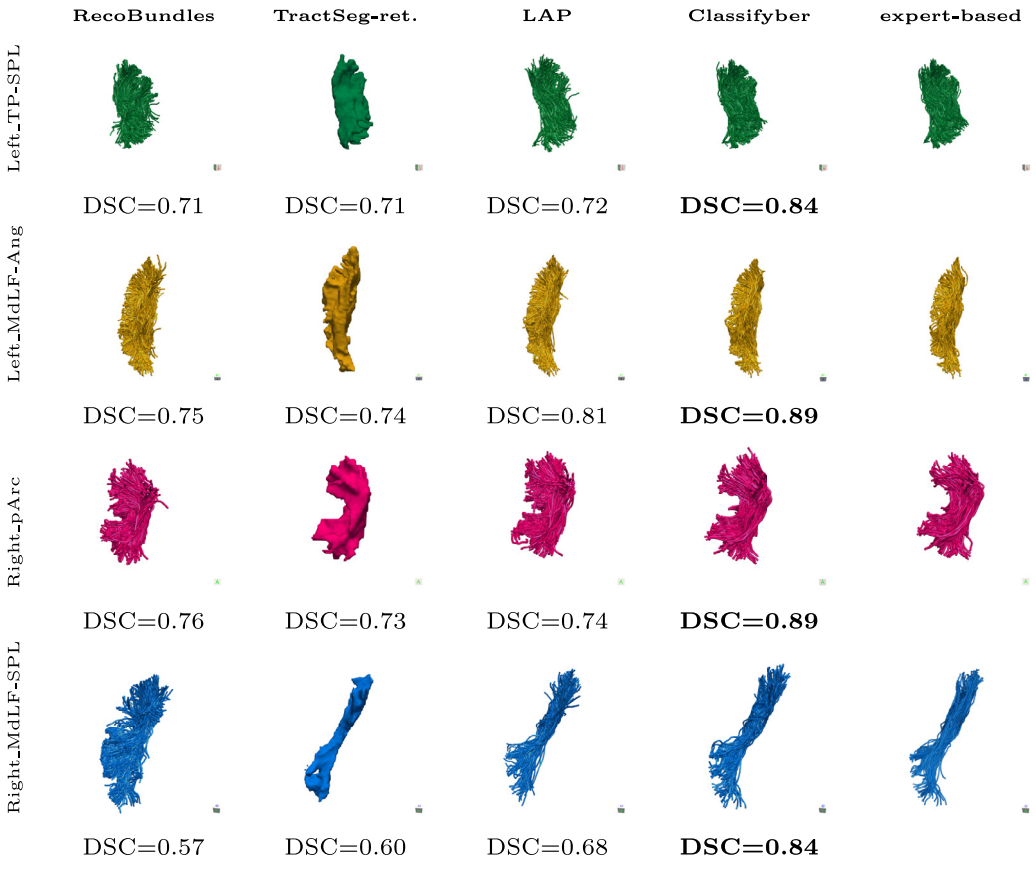

Fig. 4. Qualitative comparison of segmented bundles in one target subject. Bundles on the rows: Left_TP-SPL (first row), Left_MdLF-Ang (second row) Right_pArc (third row), and Right_MdLF-SPL (fourth row). Automatic segmentation methods on the columns: RecoBundles (first column), TractSeg-retrained (second column), LAP (third column) and Classifyber (fourth column) and expert-based segmentation (fifth column). Highest quality of segmentation in bold face.

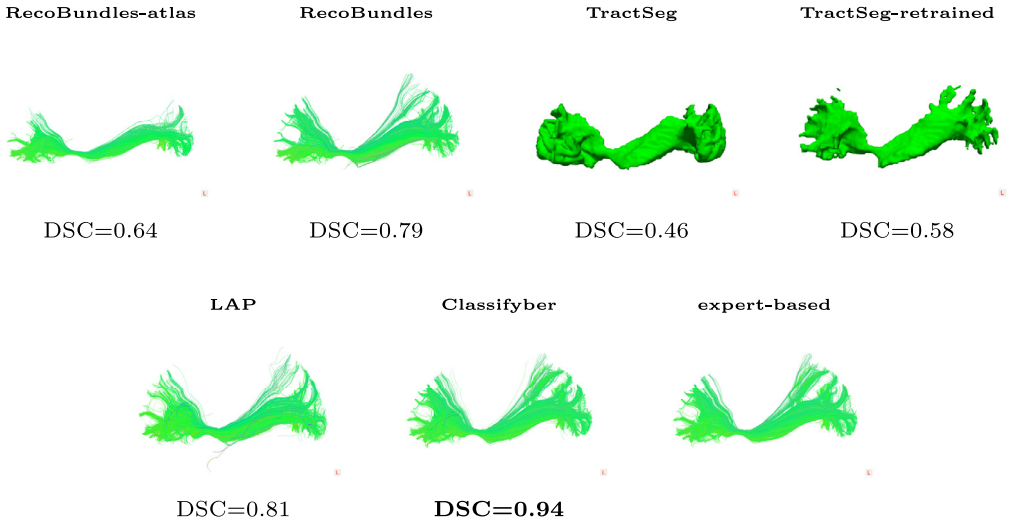

Fig. 5. Qualitative comparison of segmented bundles in one target subject. One instance of Left_IFOF for RecoBundles-atlas, RecoBundles, TractSeg, TractSeg-retrained, LAP and Classifyber with the expert-based segmented bundle. Highest quality of segmentation in bold face.

reaches a comparable quality of segmentation to TractSeg, even though it uses only 15 subjects as examples.

4.2.4. Results on clinical dataset

In Table 6 we report the quantitative comparison in terms of mean DSC for Classifyber and TractSeg. The comparison is focused on TractSeg because in Wasserthal et al. (2018a) it is stated that the method is effective on clinical quality data as well, without the need for retraining

Table 6
Quantitative comparison over the Clinical dataset: DSC (mean ± sd) across 7 target subjects for TractSeg, Classifyber-major, Classifyber-IFOF and Classifyber. Highest quality of segmentation in bold face.

	TractSeg	Classifyber-major	Classifyber-IFOF	Classifyber
Left_IFOF	0.42 ± 0.05	0.72 ± 0.09	0.81 ± 0.07	**0.89** ± 0.03
Left_AF	0.23 ± 0.02	0.74 ± 0.13	–	**0.92** ± 0.03

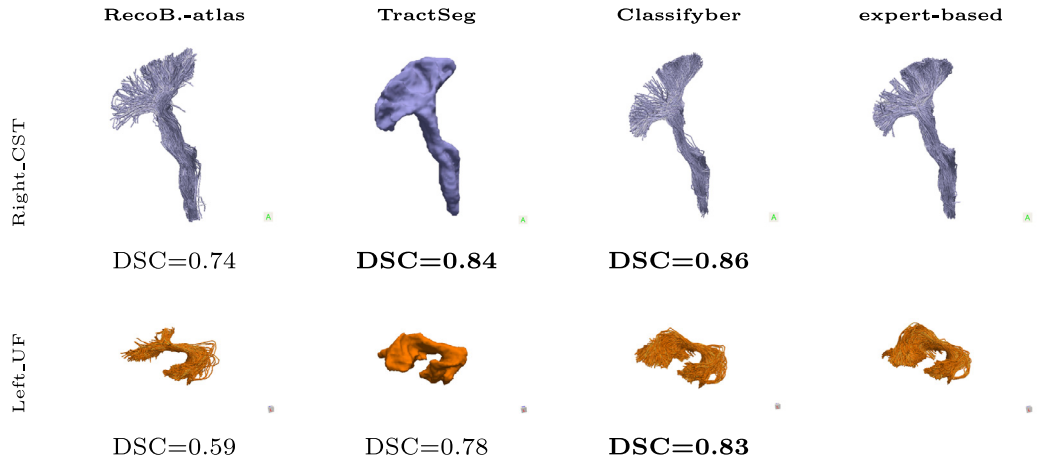

Fig. 6. Qualitative comparison of segmented bundles in one target subject. One instance of Right_CST and Left_UF for RecoBundles-atlas, TractSeg, and Classifyber with the expert-based segmented bundle. Highest quality of segmentation in bold face.

Table 7
FD values of the 4 datasets used in this work. The dataset are sorted according to their mean FD.

dataset	FD min	FD max	FD mean ± sd
HCP-major	2.09	2.44	2.30 ± 0.08
HCP-minor	1.89	2.26	2.10 ± 0.08
HCP-IFOF	1.74	2.08	1.99 ± 0.06
Clinical	1.75	1.96	1.86 ± 0.06

the network. Individually, over the 14 segmented bundles, Classifyber always obtained a higher DSC than TractSeg, for all the different training sets, i.e. for all three different variants: Classifyber-major, Classifyber-IFOF and Classifyber (Wilcoxon signed-rank test, p-value=9.8×10^{-4}, p-value=1.8×10^{-2}, and p-value=9.8×10^{-4}, respectively).

In Fig. 8 we show a qualitative comparison between the different cases.

4.2.5. Results on Fractal Dimension (FD)

In Fig. 9, we present the relationship between the FDs and the DSC scores of each method when segmenting all bundles in the experiments over the four datasets described above, i.e. on approximately 500 bundles. In the same figure, we also show the linear interpolation of such values as a summary of all experiments presented in this work, reporting the Pearson correlation coefficient (R) between FD and DSC. The results show that the quality of segmentation of TractSeg is strongly dependent on the FD of the bundle to be segmented. LAP also shows some degree of dependency, while RecoBundles and Classifyber are not affected by the FD of bundles. Additionally, in Table 7, we report the different range of FD values across the four datasets described in Section 3. Bundles of the HCP-major dataset have on average the highest FD, while bundles of the clinical dataset the lowest.

4.2.6. Classifyber: the size of the training set

For all automatic segmentation methods that learn from examples, the higher the number of training subjects, the better the resulting quality of segmentation. Nevertheless, in practice, the cost of time and effort by an expert to prepare a curated training set severely limits this number. In Fig. 10 we show how the mean DSC of Classifyber over multiple bundles changes with the number of training subjects. We observe that the quality of segmentation has no substantial increase beyond approximately 15 subjects and plateaus at 30 subjects.

4.2.7. Analysis of the computing time

In Table 8 we report the time required by each segmentation method for the training phase and for segmenting one bundle of the HCP-IFOF dataset. We chose this dataset because it is the only dataset on which we compared all segmentation methods and variants.

Table 8
Time in minutes required to train each method and to segment one IFOF for: RecoBundles-atlas, RecoBundles, TractSeg, TractSeg-retrained, LAP and Classifyber, when having 15 training examples. (*) GPU accelerated. (**) segmenting 2 kinds of bundles at the same time. (***) Training on 84 subjects to segment 72 kinds of bundles at the same time.

	Training phase	Segmentation	Total
RecoBundles-atlas	0	0.5	0.5
RecoBundles	0	3	3
Classifyber	34	3	37
LAP	0	130	130
TractSeg-ret.(*)	175(**)	5	180
TractSeg(*)	720(***)	5	725

We observed that the training time is linearly correlated with the number of training streamlines. For example, in the experiments of on the HCP-major dataset, by using only 10% of the training set, the training time was reduced 10 times as well. When trained, Classifyber segments bundles in just 3 minutes. The main cost of the computation in both the training and test phases is the preparation of the input for the classifier, i.e. steps (a1) and (a2), and steps (b1) and (b2). The actual segmentation, i.e. step (b3) only requires less than 1 second.

In contrast to Classifyber, RecoBundles and LAP do not require training time, because their underlying learning algorithms, i.e. nearest neighbor and linear assignment respectively, are *lazy learning* algorithms that postpone the computation to when the testing/segmentation step is required. In the case of RecoBundles, the segmentation step requires between 0.5 and 3 minutes, on the example discussed above. LAP requires 130 minutes and it is the slowest of the methods compared.

TractSeg adopts a different approach because it segments 72 bundles in parallel. The training time of TractSeg is vastly larger than all other methods, requiring 7 h on a GPU. When the bundle of interest is not included in those 72 bundles, or when the training examples differ from the ones used in Wasserthal et al. (2018a), we re-trained TractSeg (called TractSeg-retrained): for example, on the examples of HCP-IFOF, the training phase required approximately 3 hours on GPU, see Table 8. Both TractSeg and TractSeg-retrained required approximately 5 min to segment a new bundle.

All computations of all experiments described in this work were executed on the high-performance computing (HPC) cluster provided by Indiana University, allocating 16 cores of Intel Xeon CPU

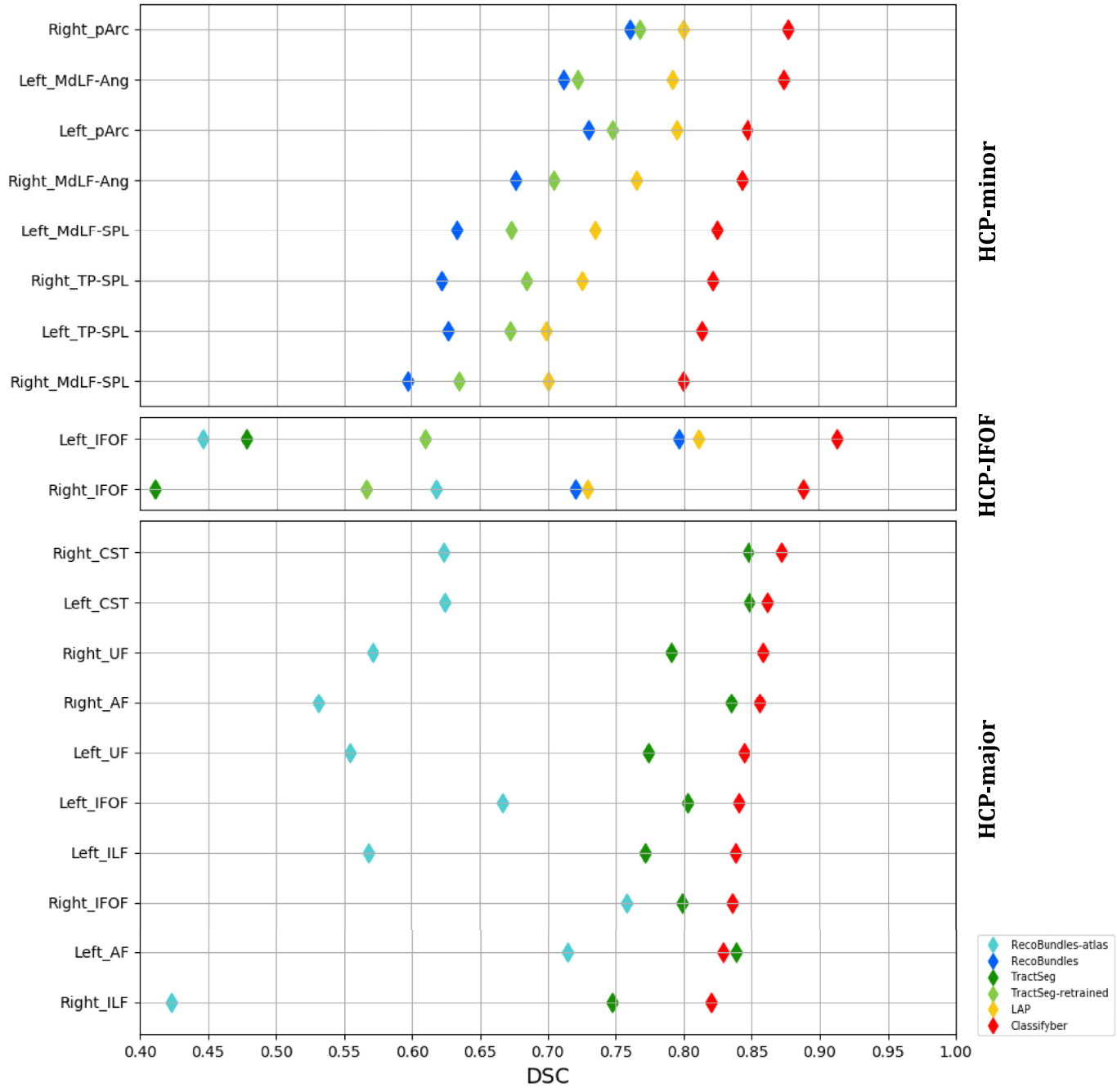

Fig. 7. Summary of the quantitative comparison across the three HCP datasets. Top: mean DSC across 25 subjects of the HCP-minor dataset. Middle: mean DSC across 15 subjects of the HCP-IFOF dataset. Bottom: mean DSC across 21 subjects of the HCP-major dataset. The methods compared are depicted in different colors: RecoBundle-atlas (light blues), RecoBundles (blue), TractSeg (green), TractSeg-retrained (light green), LAP (yellow) and Classifyber (red).

E5-2680 2.50 GHz and 32Gb of RAM, a setup equivalent to a powerful personal workstation typically available in research labs and clinics. For TractSeg, we also allocated one NVIDIA GPU RTX 2080Ti.

5. Discussion and conclusions

5.1. General comments

At the global level, all the results on the comparison among automatic segmentation methods presented in Section 4.2 indicate one main message: Classifyber clearly outperforms other methods in all cases, by a substantial margin, and segments bundles very accurately. This is observed to occur across different kinds of bundles, tractography techniques, expert-made segmentations, and quality of dMRI data, i.e., research vs clinical quality. The summary results in Fig. 9, which report on the y-axes the DSC score for each of the hundreds of individual bundles segmented across all the experiments of Section 4.1, show that Classifyber obtained scores ranging from 0.65 to 0.96, with a mean and standard deviation of 0.85 ± 0.05. This is the highest quality of segmentation among the different automatic segmentation methods by a large or substantial margin, in almost all cases, see the results at the

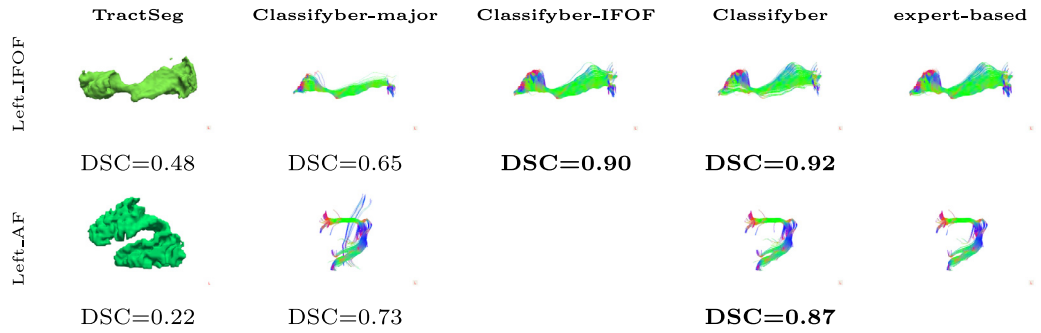

Fig. 8. Qualitative comparison of segmented bundles in one of the patients. Bundles on the rows: Left_IFOF (first row) and Left_AF (second row). Automatic segmentation methods on the columns: TractSeg (first column), Classifyber-major (second column), Classifyber-IFOF (third column), Classifyber (fourth column) and expert-based segmentation (fifth column). Highest quality of segmentation in bold face.

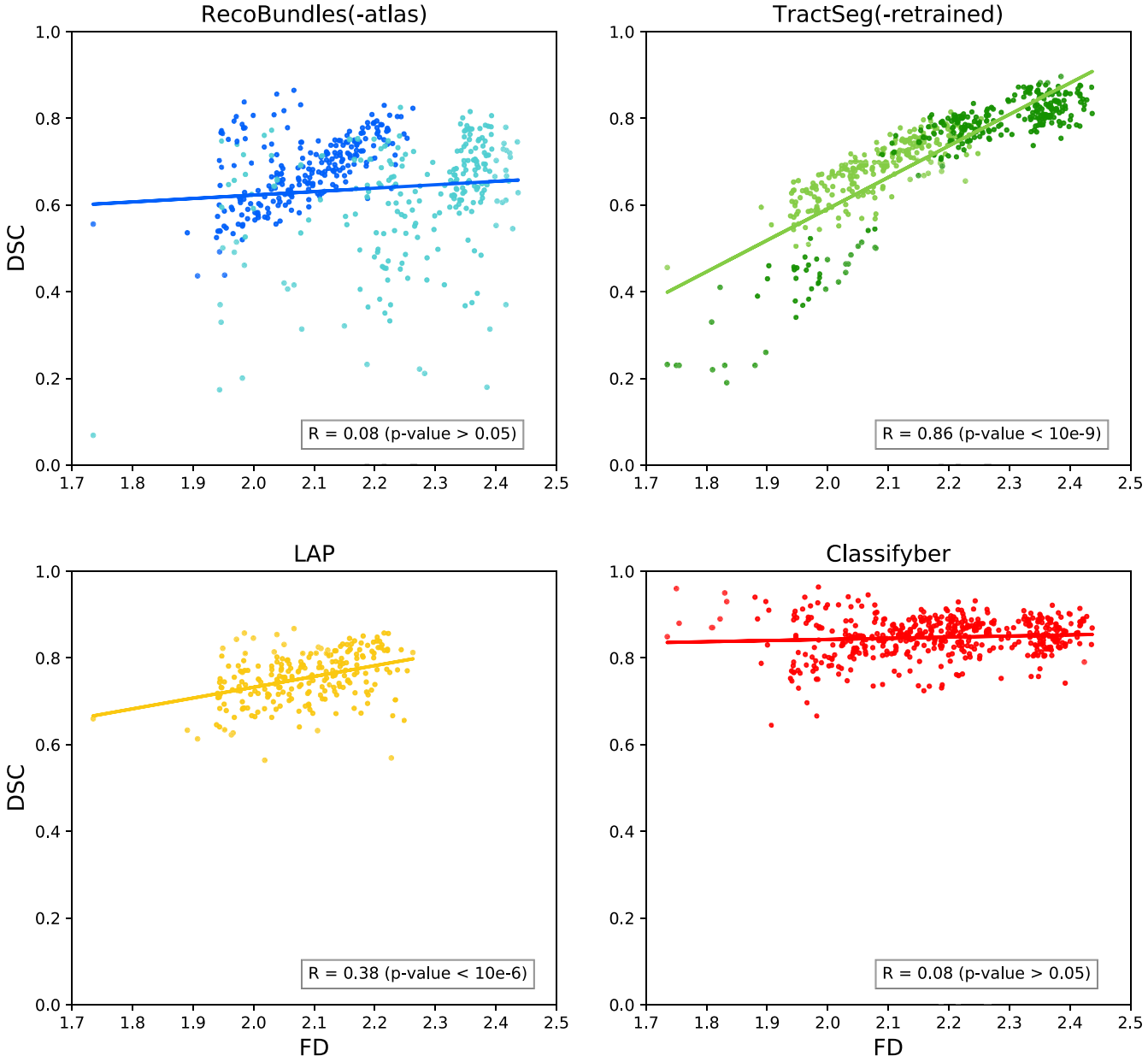

Fig. 9. DSC vs FD across all methods for all the predicted bundles of the experiments in this work. From top left: RecoBundles and RecoBundles-atlas (blue and light blue), TractSeg and TractSeg-retrained (green and light green), LAP (yellow), and Classifyber (red). R is the Pearson correlation coefficient and related *p*-value between FD and DSC over all the predicted bundles, i.e. approximately 500 segmentations.

level of individual target bundles reported in Sections 4.2.1–4.2.4 and in Fig. 7.

Fig. 9 also reports that the results obtained by LAP are consistently superior to those obtained by RecoBundles and TractSeg, at least on the datasets HCP-minor and HCP-IFOF. Moreover, the figure shows that RecoBundles and TractSeg have a large amount of variability in the quality of segmentation across the different experiments: their DSC scores range from 0.07 to 0.90, with means of 0.64 ± 0.14 and 0.71 ± 0.13 respectively. Surprisingly, TractSeg reaches a low (or very low) quality of segmentation on small bundles. We discuss this point in detail below, in Section 5.3, where we analyze the FD of bundles.

5.2. Discussion of the comparison across datasets

HCP-minor Fig. 7 and Table 3 show that the quality of segmentation obtained by Classifyber is very high (DSC ≥ 0.80) across all kinds of small bundles and distinctively superior to all other methods[9]. This result is of particular importance because minor bundles are notoriously harder to segment due to their size and high variability across subjects (Guevara et al., 2017).

In the qualitative comparison in Fig. 4 we observe that TractSeg-retrained is not very precise in segmenting fine-grained structures of the bundles, in particular their terminal portions. We believe that this is due to an inherent bias of FCNNs, which we discuss in Section 5.3.

HCP-IFOF

When segmenting the IFOFs of the HCP-IFOF dataset, Classifyber reaches an extremely high quality of segmentation, with DSC around 0.9, as reported in Table 4. RecoBundles and LAP ranked second with DSC around 0.8. TractSeg-retrained, despite being trained on the IFOFs of that dataset, ranked third with DSC around 0.6. Also in this case we believe that this is evidence of an inherent bias of the method, which we discuss in Section 5.3. TractSeg and RecoBundles-atlas ranked last with DSC around 0.5.

A possible explanation of the poor performances of TractSeg and RecoBundles-atlas is that the anatomical shape of the bundles used as examples differs from the shape of the manually expert-based segmented bundles of the HCP-IFOF dataset. Specifically, the example used by RecoBundles-atlas, i.e. the IFOF of the atlas of Yeh et al. (2018), comes from clustering followed by expert labeling. The examples used by TractSeg come from a semi-automatic refinement of the segmentation provided by TractQuerier (Wassermann et al., 2016), while the examples in HCP-IFOF are manually segmented by an expert neurosurgeon and follow the definition in Sarubbo et al. (2013) and Hau et al. (2016). These anatomical differences are justified by the fact that the anatomical definition of some white matter bundles, among which the IFOF, is in evolution (Forkel et al., 2014; Sarubbo et al., 2013; Wu et al., 2016a).

HCP-major Even for the segmentation of major bundles, Classifyber obtained very high quality of segmentation, ranging from DSC = 0.82 for the Right_ILF, to DSC = 0.87 for the Right_CST, see Table 5 and Fig. 7, outperforming in most of the cases all other methods. Nevertheless, TractSeg reached slightly inferior segmentation quality, with an average DSC ranging from 0.75 to 0.85, even though it used a much larger training set of 84 subjects instead of 15. Due to their size, major bundles are generally easier to segment (Guevara et al., 2017). On the contrary, RecoBundles-atlas obtained more modest and highly-variable results, with an average DSC ranging from 0.42 to 0.76, although we used the bundle models from Yeh et al. (2018) as suggested by the authors of RecoBundles. We believe that this result is partly motivated by the fact that the bundles used as examples by RecoBundles-atlas may have a different shape than those of the HCP-major dataset, as already discussed in the paragraph related to the HCP-IFOF dataset.

[9] The mean improvement in terms of DSC with respect to the second-best method is 0.09.

Clinical On the Clinical dataset, i.e. on the white matter of patients with a brain tumor in the same hemisphere as the bundles of interest, Classifyber reached extremely high quality of segmentation, i.e. DSC around 0.9 as reported in Table 6, when the training examples came from the same clinical dataset. When examples partly differ from the ones in the Clinical dataset, the DSC dropped accordingly to around 0.8 for Classifyber-IFOF and to 0.7 for Classifyber-major, see Table 6 and the example in Fig. 8 (first row). Specifically, in Classifyber-IFOF, the tractography of the training bundles is built on research-quality data instead of clinical-quality and the reconstruction step of the tractography is CSD instead of DTI. In Classifyber-major the differences are even greater: the training data is research-quality and the tractography is probabilistic instead of the deterministic tractography featured in the Clinical dataset. Moreover, in this case, the definition of the IFOF is the classical one provided by TractQuerier (Wassermann et al., 2016) instead of the more refined from Sarubbo et al. (2013) and Hau et al. (2016) used in the Clinical dataset.

It is well known that training classification algorithms on examples that systematically differ from the examples in the test set substantially reduces the quality of classification. This problem, called *domain shift*, was previously mentioned for bundle segmentation in Wasserthal et al. (2018a) and has no simple solution.

Although in Wasserthal et al. (2018a) they claim that their pretrained network works properly also on clinical settings, the results of TractSeg on the Clinical dataset are surprisingly low, with DSC around 0.3, as reported in Table 6. These results should be comparable to those of Classifyber-major, which instead reached a DSC around 0.7. We believe that the main reason of this behavior is the low FD of the clinical bundles, which has a strong impact on TractSeg as explained in detail below in Section 5.3.

5.3. The fractal dimension of bundles

While conducting hundreds of automatic segmentations with different methods, we noticed that TractSeg had consistent success or consistent failure on specific datasets. TractSeg very accurately segmented the bundles in HCP-major, but obtained only medium or poor results in other datasets, see Fig. 7. Fig. 9 shows that the segmentation quality reached by TractSeg is deeply affected by a specific geometric property of the voxel mask of the target bundle: its fractal dimension (FD, see Section 2.7). Tractseg accurately segmented bundles which are smooth and rounded, i.e., with high FD, while it produced poor segmentations when they are wrinkled and irregular, i.e., with low FD. By Combining the information of Table 7 and the trends in Fig. 9, we can indeed expect TractSeg to accurately segment bundles in the HCP-major dataset and to consistently fail in the HCP-IFOF or Clinical datasets.

We believe that this tendency is related to the operations of convolution and max-pooling of the fully convolutional neural networks (FCNNs) within TractSeg. In the domain of computer vision, it has been observed multiple times that FCNNs are biased towards rounded segmentations of objects, which can loose details and fine-grained structure, in particular because of the max-pooling operation, see for example (Kim et al., 2018; Sabour et al., 2017; Wei et al., 2019). This problem is inherent in U-net (Ronneberger et al., 2015), which is at the core of TractSeg.

As an example, consider the experiments related to the segmentation of the IFOFs and Fig. 5. The IFOFs in the HCP-IFOF dataset were manually segmented by experts and, according to Table 7, have FD = 1.99 ± 0.06. The IFOFs predicted by TractSeg have FD = 2.3 ± 0.1, and appear substantially more rounded and smoother than the expert-based segmented IFOFs, see Fig. 5 (third and last panels). Even retraining TractSeg only on examples of HCP-IFOF did not solve this problem but instead merely mitigated it: the IFOFs predicted by TractSeg-retrained have FD = 2.1 ± 0.1, which is still systematically higher than the expert-based segmentations, confirming the bias, see for example Fig. 5 (fourth panel).

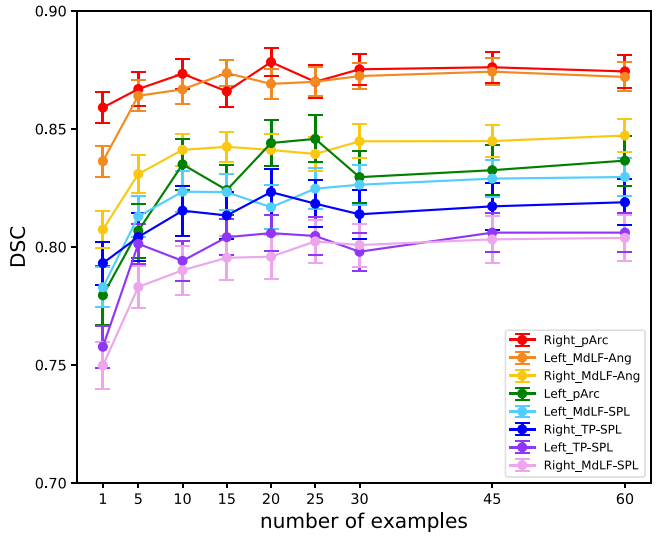

Fig. 10. Effect on the segmentation accuracy when varying the number of examples to train Classifyber. DSC (mean ± sd of the mean) across 25 test subjects of the HCP-minor dataset when varying the number of examples, from 1 to 60. Each bundle is depicted with a different color.

Fig. 9 shows that LAP is also slightly affected by the FD of bundles, though much less than TractSeg. However, such result might not be entirely reliable because a large portion of the segmentations are missing due to the limitation of LAP to address large tractograms.

In contrast to TractSeg and LAP, both RecoBundles and Classifyber are insensitive to the FD of the voxel masks of bundles, as clearly shown in Fig. 9. We speculate that the reason for this is related to the streamline-based nature of such methods and, more specifically, to the fact that they operate via *single* streamline classification. By predicting whether or not each streamline of the tractogram belongs to the target bundle, there is not a specific constraint to produce round/smooth voxel-structures as observed with TractSeg or to jointly consider all target streamlines during the prediction as in LAP.

5.4. Size of the training set

As an additional result of this work, we observed that Classifyber requires only a small number of example bundles to obtain high quality of segmentation. In fact, Fig. 10 shows that, on the HCP-minor dataset, there is no substantial gain in the quality of segmentation beyond 15 training examples. In the experiments on the Clinical dataset, Classifyber reached an extremely high segmentation quality using only 6 example subjects, with a mean DSC around 0.9, see Table 6.

Both RecoBundles and LAP require a very small number of training subjects: 1 bundle/model for RecoBundles and around 5-10 for LAP, according to Sharmin et al. (2018). On the contrary, TractSeg was trained on 84 subjects. Although in Wasserthal et al. (2018a) there are no clear guidelines on the number of subjects to be used for training, it is well known that deep learning models need a very large training set, which is often not available in clinical settings.

5.5. Time required to segment a bundle

Among the methods compared in this work, deciding which one is faster is not straightforward: on the one hand, streamline-based methods like Classifyber, RecoBundles and LAP require the tractogram as input. In our experience and applications, the tractogram is always already available and provided by neurosurgeons/neuroscientists, because they decide the reconstruction and tracking algorithms specifically for their desired task, the available MR scanner and sequence of acquisition. If only raw dMRI data is provided, the time to build the tractogram should be accounted for the total time of the computation. On the other hand, TractSeg uses the GPU and requires a specific pre-processing of dMRI data as input, which needs approximately 30 minutes of computation per subject. Moreover, to obtain the predicted bundle as streamlines, bundle-specific tracking must be computed afterwards (Wasserthal et al., 2018b).

Overall, if the target tractogram is available, RecoBundles is the fastest segmentation method in our comparison, see Table 8. Alternatively, if pre-trained methods are available, like in the case of TractSeg and Classifyber, TractSeg and Classifyber are also similarly as fast as RecoBundles. LAP is the slowest segmentation method but, if training has to be done, TractSeg ranks last.

5.6. Reproducibility

The results on large bundles that we present in Table 5 and Fig. 7 reproduce those in Wasserthal et al. (2018a) for what concerns TractSeg and RecoBundles, for the variant RecoBundles-atlas: TractSeg has a distinctively higher quality of segmentation than RecoBundles. However, when considering other datasets, the situation is different. On the dataset HCP-minor, see Fig. 7 and Table 3, RecoBundles shows comparable quality of segmentation to TractSeg, while on the dataset HCP-IFOF, see Table 4, RecoBundles has better quality of segmentation than TractSeg. These results are novel because Wasserthal et al. (2018a) did not consider bundles with low FD.

The better performances of LAP than those of RecoBundles and Tractseg on the dataset HCP-minor, and of TractSeg on the dataset HCP-IFOF, are shown in the same tables and figures just mentioned. With respect to RecoBundles, this result is consistent with what was demonstrated in Sharmin et al. (2018), i.e., that LAP outperforms the nearest-neighbor-based segmentation, which is the category to which RecoBundles belongs. With respect to TractSeg, the result of our comparison is novel, because LAP was not included in the extensive comparison presented in Wasserthal et al. (2018a).

The sharing of code and data is becoming standard practice in neuroscience and facilitates both accelerated scientific discovery and reproducibility, see Avesani et al. (2019). For this reason, Classifyber is freely available on the online platform https://brainlife.io both as the full algorithm that implements the training and test phases, and as a pre-trained method ready to segment bundles in the highest quality fashion available.

5.7. Conclusions

In this work we present Classifyber, a streamline-based linear classifier that segments white matter bundles from dMRI data and expert-made examples. Classifyber is the first automatic classification-based segmentation method that exploits both the shape of streamlines, obtained with tractography techniques from dMRI data, and the anatomical information of the bundles, in the form of connectivity patterns and specific ROIs. Classifyber substantially raises the quality of segmentation as compared to the current state-of-the-art methods described in the literature, by a large margin, and more importantly, across very diverse settings. Maintaining a high quality of bundle segmentation regardless of the type of input tractography or the quality of dMRI data is nowadays of paramount importance for a vast number of applications. For example, the practitioner may not be able to anticipate whether the bundle to be segmented will have high or low FD.

As opposed to voxel-based methods like TractSeg, we believe that accurate segmentation of bundles from dMRI data must leverage tractography techniques and also include information about streamlines. Streamlines represent a spatial statistic of the dMRI signal that approximates the underlying anatomical connectivity, though it does so with a substantial problem of false positives (Daducci et al., 2015; Jeurissen et al., 2019; Maier-Hein et al., 2017; Pestilli et al., 2014).

Additionally, Classifyber is fast to train on new datasets/bundles and requires only a small number of examples. This specific feature is of great importance for bundle-specific applications like in pre-surgical planning, because Classifyber can be tailored to the specific task, dMRI data and tractography technique at the cost of a small amount of manual segmentation by expert neuroanatomists.

In future, we plan to test nonlinear classification algorithms in order to investigate potential improvements in the segmentation quality of Classifyber. The current linear model used within Classifyber is indeed a limitation of the proposed method. Nevertheless, linear models are fast and light and, according to the results presented in this work, sufficient to substantially advance the state-of-the-art in automatic white matter bundle segmentation.

Declaration of Competing Interest

The authors declare no competing financial interests.

Credit authorship contribution statement

Giulia Bertò: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing - original draft. **Daniel Bullock:** Data curation, Software, Validation, Writing - review & editing. **Pietro Astolfi:** Software, Writing - review & editing. **Soichi Hayashi:** Resources, Software, Writing - review & editing. **Luca Zigiotto:** Data curation, Validation. **Luciano Annicchiarico:** Data curation. **Francesco Corsini:** Data curation. **Alessandro De Benedictis:** Data curation, Validation. **Silvio Sarubbo:** Data curation, Validation, Writing - review & editing. **Franco Pestilli:** Conceptualization, Resources, Software, Supervision, Writing - review & editing, Funding acquisition. **Paolo Avesani:** Conceptualization, Data curation, Software, Supervision, Writing - review & editing, Funding acquisition. **Emanuele Olivetti:** Conceptualization, Formal analysis, Investigation, Methodology, Software, Project administration, Supervision, Writing - original draft.

Acknowledgements

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.

F.P. was supported by NSF IIS-1636893, NSF BCS-1734853, NSF AOC-1916518, a Microsoft Investigator Fellowship, Microsoft Research Azure Award, Google Cloud Platform, and the Indiana University Areas of Emergent Research initiative “Learning: Brains, Machines, Children.”.

The authors acknowledge the Indiana University Pervasive Technology Institute for providing HPC (Big Red 3, Karst, Carbonate), visualization, database, and storage resources that have contributed to the research results reported within this paper.

Appendix A. Semi-automatic technique to curate the HCP-minor bundle dataset

In this Section, we describe the semi-automatic technique adopted to filter out bundles considered not anatomically plausible in the HCP-minor dataset.

First, we automatically discarded those subjects which had at least one bundle that deviated more than $\pm$ 2 standard deviations from the mean of the bundle distribution of the number of voxels and number of streamlines of the population across the 192 subjects. After this step, the number of subjects retained was 121. Then, an expert (D.B.) performed visual inspection of each individual bundle to detect anomalies in the segmentations. Bundles were assigned an omnibus score corresponding to their degree of anatomical plausibility. These scores ranged from 1 to 5 such that 1 indicated a rating of *bad*, 2 indicated a rating of *poor*, 3 indicated a rating of *OK*, 4 indicated a rating of *good*, 5 indicated a rating of *great*. Finally, we kept those subjects whose *all* bundles obtained a score of 2 or higher, remaining with a total of 105 subjects.

Appendix B. TractSeg-retrained metrics on HCP-minor dataset

In Fig. B.11, we report the training metrics obtained when training TractSeg-retrained on HCP-minor dataset as explained in Section 2.5. Red lines represent the value of the loss function obtained across all the epochs (y-axis labels on the left side), while green lines represent the f1 score (y-axis labels on the right side). The graph shows that we reached convergence in 250 iterations.

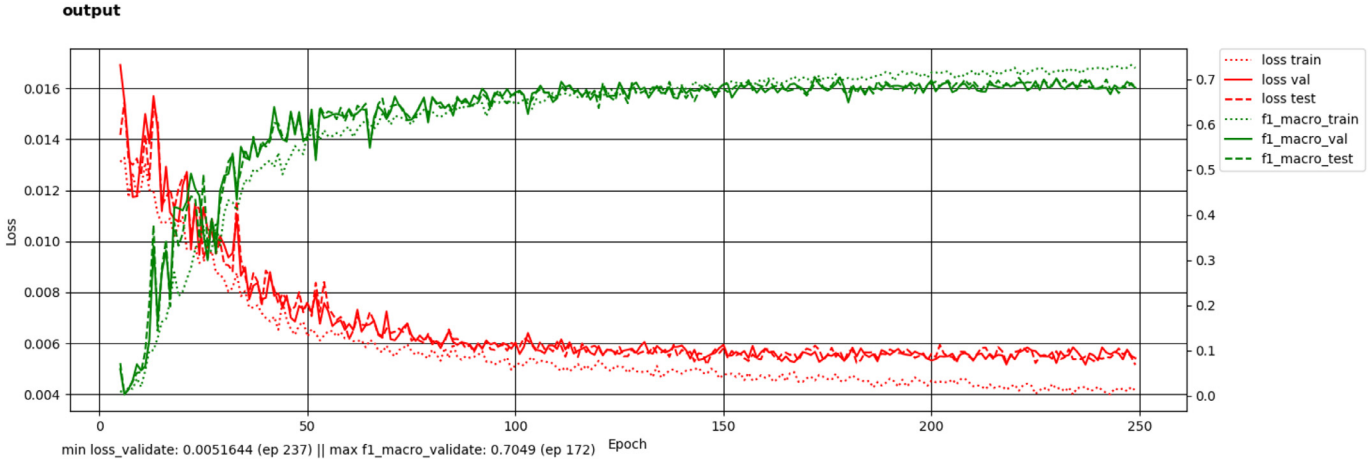

Fig. B1. Metrics to train 15 subjects for TractSeg-retrained on the HCP-minor dataset, data augmentation, 250 epochs.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.neuroimage.2020.117402.

References

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.

Avesani, P., McPherson, B., Hayashi, S., Caiafa, C.F., Henschel, R., Garyfallidis, E., Kitchell, L., Bullock, D., Patterson, A., Olivetti, E., Sporns, O., Saykin, A.J., Wang, L., Dinov, I., Hancock, D., Caron, B., Qian, Y., Pestilli, F., 2019. The open diffusion data derivatives, brain data upcycling via integrated publishing of derivatives and reproducible open cloud services. Sci. Data 6 (1), 1–13. doi:10.1038/s41597-019-0073-y.

Berman, J.I., Chung, S., Mukherjee, P., Hess, C.P., Han, E.T., Henry, R.G., 2008. Probabilistic streamline q-ball tractography using the residual bootstrap. NeuroImage 39 (1), 215–222. doi:10.1016/j.neuroimage.2007.08.021.

Bertò, G., Avesani, P., Pestilli, F., Bullock, D., Caron, B., Olivetti, E., 2019. Anatomically-informed multiple linear assignment problems for white matter bundle segmentation. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 135–138. doi:10.1109/ISBI.2019.8759174.

Brun, A., Knutsson, H., Park, H.-J., Shenton, M.E., Westin, C.-F., 2004. Clustering fiber traces using normalized cuts. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2004. Springer, pp. 368–375.

Bullock, D., Takemura, H., Caiafa, C.F., Kitchell, L., McPherson, B., Caron, B., Pestilli, F., 2019. Associative white matter connecting the dorsal and ventral posterior human cortex. Brain Struct. Funct. doi:10.1007/s00429-019-01907-8.

Catani, M., Howard, R.J., Pajevic, S., Jones, D.K., 2002. Virtual in vivo interactive dissection of white matter fasciculi in the human brain.. NeuroImage 17 (1), 77–94.

Daducci, A., Dal Palù, A., Lemkaddem, A., Thiran, J.-P., 2015. COMMIT: convex optimization modeling for microstructure informed tractography. IEEE Trans. Med. Imaging 34 (1), 246–257.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26 (3), 297–302.

Esteban, F.J., Sepulcre, J., de Mendizábal, N.V., Goñi, J., Navas, J., de Miras, J.R., Bejarano, B., Masdeu, J.C., Villoslada, P., 2007. Fractal dimension and white matter changes in multiple sclerosis. NeuroImage 36 (3), 543–549. doi:10.1016/j.neuroimage.2007.03.057.

Falconer, K.J., 2014. Fractal Geometry: Mathematical Foundations and Applications, third ed. John Wiley & Sons Inc, Hoboken.

Fischl, B., 2012. FreeSurfer. NeuroImage 62 (2), 774–781. doi:10.1016/j.neuroimage.2012.01.021.

Fonov, V., Evans, A.C., Botteron, K., Almli, C.R., McKinstry, R.C., Collins, D.L., Brain Development Cooperative Group, 2011. Unbiased average age-appropriate atlases for pediatric studies. NeuroImage 54 (1), 313–327. doi:10.1016/j.neuroimage.2010.07.033.

Forkel, S.J., Thiebaut de Schotten, M., Kawadler, J.M., Dell'Acqua, F., Danek, A., Catani, M., 2014. The anatomy of fronto-occipital connections from early blunt dissections to contemporary tractography. Cortex 56, 73–84. doi:10.1016/j.cortex.2012.09.005.

Garyfallidis, E., 2018. Simple model bundle atlas for RecoBundles. figshare. Dataset. 10.6084/m9.figshare.6483614.v1

Garyfallidis, E., Brett, M., Amirbekian, B., Rokem, A., van der Walt, S., Descoteaux, M., Nimmo-Smith, I., Contributors, D., 2014. Dipy, a library for the analysis of diffusion MRI data. Front. Neuroinform. 8 (8), 1+.

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

Garyfallidis, E., Ocegueda, O., Wassermann, D., Descoteaux, M., 2015. Robust and efficient linear registration of white-matter fascicles in the space of streamlines. NeuroImage 117, 124–140. doi:10.1016/j.neuroimage.2015.05.016.

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., Van Essen, D.C., Jenkinson, M., 2013. The minimal preprocessing pipelines for the Human Connectome Project. NeuroImage 80, 105–124. doi:10.1016/j.neuroimage.2013.04.127.

Gorgolewski, K., 2016. FreeSurfer reconstruction of the MNI152 ICBM2009c asymmetrical non-linear atlas. figshare. Dataset. 10.6084/m9.figshare.4223811.v1

Guevara, M., Román, C., Houenou, J., Duclap, D., Poupon, C., Mangin, J.F., Guevara, P., 2017. Reproducibility of superficial white matter tracts using diffusion-weighted imaging tractography. NeuroImage 147, 703–725.

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.

Guevara, P., Duclap, D., Poupon, C., Marrakchi-Kacem, L., Houenou, J., Leboyer, M., Mangin, J.-F., 2011. Segmentation of short association bundles in massive tractography datasets using a multi-subject bundle atlas. In: San Martin, C., Kim, S.-W. (Eds.), Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Springer, Berlin Heidelberg, pp. 701–708.

Hau, J., Sarubbo, S., Perchey, G., Crivello, F., Zago, L., Mellet, E., Jobard, G., Joliot, M., Mazoyer, B.M., Tzourio-Mazoyer, N., Petit, L., 2016. Cortical terminations of the inferior fronto-occipital and uncinate fasciculi: anatomical stem-based virtual dissection. Front. Neuroanat. 10. doi:10.3389/fnana.2016.00058.

Jeurissen, B., Descoteaux, M., Mori, S., Leemans, A., 2019. Diffusion MRI fiber tractography of the brain. NMR Biomed. 32 (4), e3785. doi:10.1002/nbm.3785.

Kim, S., Bae, W.C., Masuda, K., Chung, C.B., Hwang, D., 2018. Fine-grain segmentation of the intervertebral discs from MR spine images using deep convolutional neural networks: BSU-Net. Appl. Sci. 8 (9). doi:10.3390/app8091656.

Labra, N., Guevara, P., Duclap, D., Houenou, J., Poupon, C., Mangin, J.-F., Figueroa, M., 2016. Fast automatic segmentation of white matter streamlines based on a multi-subject bundle atlas. Neuroinformatics 15 (1), 71–86. doi:10.1007/s12021-016-9316-7.

Maddah, M., Mewes, A.U.J., Haker, S., Grimson, Warfield, S.K., 2005. Automated atlas-based clustering of white matter fiber tracts from DTMRI. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2005. Springer, Berlin, Heidelberg, pp. 188–195.

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. Nat. Commun. 8 (1). doi:10.1038/s41467-017-01285-x.

Mandelbrot, B.B., 1982. The Fractal Geometry of Nature. W.H. Freeman, San Francisco.

Mayer, A., Zimmerman-Moreno, G., Shadmi, R., Batikoff, A., Greenspan, H., 2011. A supervised framework for the registration and segmentation of white matter fiber tracts. IEEE Trans. Med. Imaging 30 (1), 131–145. doi:10.1109/TMI.2010.2067222.

Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., Woods, R., Paus, T., Simpson, G., Pike, B., Holmes, C., Collins, L., Thompson, P., MacDonald, D., Iacoboni, M., Schormann, T., Amunts, K., Palomero-Gallagher, N., Geyer, S., Parsons, L., Narr, K., Kabani, N., Le Goualher, G., Feidler, J., Smith, K., Boomsma, D., Pol, H.H., Cannon, T., Kawashima, R., Mazoyer, B., 2001. A four-dimensional probabilistic atlas of the human brain. J. Am. Med. Inform.Assoc. 8 (5), 401–430.

Mori, S., Wakana, S., Van Zijl, P.C.M., Nagae-Poetscher, L.M., 2005. MRI atlas of human white matter, 16. Am Soc Neuroradiology, Amsterdam, The Netherlands.

O'Donnell, L.J., Suter, Y., Rigolo, L., Kahali, P., Zhang, F., Norton, I., Albi, A., Olubiyi, O., Meola, A., Essayed, W.I., Others, 2017. Automated white matter fiber tract identification in patients with brain tumors. NeuroImage 13, 138–153.

O'Donnell, L.J., Westin, C.-F.F., 2007. Automatic tractography segmentation using a high-dimensional white matter atlas. In: IEEE Trans. Med. Imag, pp. 1562–1575.

Oishi, K., Zilles, K., Amunts, K., Faria, A., Jiang, H., Li, X., Akhter, K., Hua, K., Woods, R., Toga, A.W., Pike, G.B., Rosa-Neto, P., Evans, A., Zhang, J., Huang, H., Miller, M.I., van Zijl, P.C.M., Mazziotta, J., Mori, S., 2008. Human brain white matter atlas: Identification and assignment of common anatomical structures in superficial white matter. NeuroImage 43 (3), 447–457. doi:10.1016/j.neuroimage.2008.07.009.

Olivetti, E., Avesani, P., 2011. Supervised segmentation of fiber tracts. In: Proceedings of the First international conference on Similarity-based pattern recognition. Springer-Verlag, Berlin, Heidelberg, pp. 261–274. doi:10.1007/978-3-642-24471-1_19. Event-place: Venice, Italy

Olivetti, E., Bertò, G., Gori, P., Sharmin, N., Avesani, P., 2017. Comparison of distances for supervised segmentation of white matter tractography. In: 2017 International Workshop on Pattern Recognition in Neuroimaging (PRNI). IEEE, pp. 1–4. doi:10.1109/prni.2017.7981502. Event-place: Toronto, ON, Canada

Olivetti, E., Nguyen, T.B., Garyfallidis, E., 2012. The approximation of the dissimilarity projection. In: IEEE International Workshop on Pattern Recognition in NeuroImaging, 0, pp. 85–88. doi:10.1109/prni.2012.13.

Olivetti, E., Nguyen, T.B., Garyfallidis, E., Agarwal, N., Avesani, P., 2013. Fast clustering for interactive tractography segmentation. In: Pattern Recognition in Neuroimaging (PRNI), 2013 International Workshop on. IEEE, pp. 42–45. doi:10.1109/prni.2013.20. Event-place: Philadelphia, PA

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.

Pekalska, E., Duin, R.P.W., 2005. The Dissimilarity Representation for Pattern Recognition: Foundations And Applications (Machine Perception and Artificial Intelligence). World Scientific Publishing Co., Inc., River Edge, NJ, USA.

Pestilli, F., 2018. Human white matter and knowledge representation. PLoS Biol. 16 (4). doi:10.1371/journal.pbio.2005758.

Pestilli, F., Yeatman, J.D., Rokem, A., Kay, K.N., Wandell, B.A., 2014. Evaluation and statistical inference for human connectomes. Nat. Methods 11 (10), 1058–1063. doi:10.1038/nmeth.3098.

Porro-Muñoz, D., Olivetti, E., Sharmin, N., Nguyen, T., Garyfallidis, E., Avesani, P., 2015. Tractome: a visual data mining tool for brain connectivity analysis. Data Min. Knowl. Discov. 29 (5), 1258–1279. doi:10.1007/s10618-015-0408-z.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing, pp. 234–241.

Sabour, S., Frosst, N., Hinton, G.E., 2017. Dynamic routing between capsules. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc., USA, pp. 3859–3869. Event-place: Long Beach, California, USA

Sani, I., McPherson, B.C., Stemmann, H., Pestilli, F., Freiwald, W.A., 2019. Functionally defined white matter of the macaque monkey brain reveals a Dorso-Ventral attention network. eLife 8, e40520. doi:10.7554/eLife.40520.

Sarubbo, S., De Benedictis, A., Maldonado, I., Basso, G., Duffau, H., 2013. Frontal terminations for the inferior fronto-occipital fascicle: anatomical dissection, DTI study and functional considerations on a multi-component bundle. Brain Struct. Funct. 218 (1), 21–37. doi:10.1007/s00429-011-0372-3.

Schmidt, M., Le Roux, N., Bach, F., 2017. Minimizing finite sums with the stochastic average gradient. Math. Program. 162 (1), 83–112. doi:10.1007/s10107-016-1030-6.

Sharmin, N., Olivetti, E., Avesani, P., 2016. Alignment of tractograms as linear assignment problem. In: Fuster, A., Ghosh, A., Kaden, E., Rathi, Y., Reisert, M. (Eds.), Computational Diffusion MRI. Springer International Publishing, pp. 109–120. doi:10.1007/978-3-319-28588-7_10.

Sharmin, N., Olivetti, E., Avesani, P., 2018. White matter tract segmentation as multiple linear assignment problems. Front. Neurosci. 11. doi:10.3389/fnins.2017.00754.

Siless, V., Chang, K., Fischl, B., Yendiki, A., 2016. Hierarchical clustering of tractography streamlines based on anatomical similarity. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 184–191. Event-place: Athens, Greece

Siless, V., Chang, K., Fischl, B., Yendiki, A., 2018. AnatomiCuts: hierarchical clustering of tractography streamlines based on anatomical similarity. NeuroImage 166, 32–45. doi:10.1016/j.neuroimage.2017.10.058.

Siless, V., Davidow, J.Y., Nielsen, J., Fan, Q., Hedden, T., Hollinshead, M., Beam, E., Vidal Bustamante, C.M., Garrad, M.C., Santillana, R., Smith, E.E., Hamadeh, A., Snyder, J., Drews, M.K., Van Dijk, K.R.A., Sheridan, M., Somerville, L.H., Yendiki, A., 2020. Registration-free analysis of diffusion MRI tractography data across subjects through the human lifespan. NeuroImage 214, 116703. doi:10.1016/j.neuroimage.2020.116703.

Sotiropoulos, S.N., Jbabdi, S., Xu, J., Andersson, J.L., Moeller, S., Auerbach, E.J., Glasser, M.F., Hernandez, M., Sapiro, G., Jenkinson, M., Feinberg, D.A., Yacoub, E., Lenglet, C., Van Essen, D.C., Ugurbil, K., Behrens, T.E., WU-Minn HCP Consortium, 2013. Advances in diffusion MRI acquisition and processing in the Human Connectome Project.. NeuroImage 80, 125–143.

Takemura, H., Caiafa, C.F., Wandell, B.A., Pestilli, F., 2016. Ensemble tractography.. PLoS Comput. Biol. 12 (2).

Thomas, C., Ye, F.Q., Irfanoglu, M.O., Modi, P., Saleem, K.S., Leopold, D.A., Pierpaoli, C., 2014. Anatomical accuracy of brain connections derived from diffusion MRI tractography is inherently limited. Proc. Natl. Acad. Sci. USA 111 (46), 16574–16579. doi:10.1073/pnas.1405672111.

Tunç, B., Parker, W.A., Ingalhalikar, M., Verma, R., 2014. Automated tract extraction via atlas based adaptive clustering. Neuroimage 102, 596–607.

Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E.J., Yacoub, E., Ugurbil, K., 2013. The WU-Minn human connectome project: an overview. NeuroImage 80, 62–79. doi:10.1016/j.neuroimage.2013.05.041.

Vercruysse, D., Christiaens, D., Maes, F., Sunaert, S., Suetens, P., 2014. Fiber bundle segmentation using spectral embedding and supervised learning. In: O'Donnell, L., Nedjati-Gilani, G., Rathi, Y., Reisert, M., Schneider, T. (Eds.), Computational Diffusion MRI. Springer International Publishing, pp. 103–114.

Wakana, S., Caprihan, A., Panzenboeck, M.M., Fallon, J.H., Perry, M., Gollub, R.L., Hua, K., Zhang, J., Jiang, H., Dubey, P., Blitz, A., van Zijl, P., Mori, S., 2007. Reproducibility of quantitative tractography methods applied to cerebral white matter.. NeuroImage 36 (3), 630–644. doi:10.1016/j.neuroimage.2007.02.049.

Wandell, B.A., 2016. Clarifying human white matter. Annu. Rev. Neurosci. 39 (1), 103–128. doi:10.1146/annurev-neuro-070815-013815.

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

Wasserthal, J., Neher, P., Maier-Hein, K., 2018. High quality white matter reference tracts (Version 1.2.0) [Data set]. Zenodo. Type: dataset. 10.5281/zenodo.1477956

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

Wasserthal, J., Neher, P.F., Maier-Hein, K.H., 2018. Tract orientation mapping for bundle-specific tractography. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Springer International Publishing, Cham, pp. 36–44. doi:10.1007/978-3-030-00931-1_5.

Wei, Z., Zhang, J., Liu, L., Zhu, F., Shen, F., Zhou, Y., Liu, S., Sun, Y., Shao, L., 2019. Building detail-sensitive semantic segmentation networks with polynomial pooling. pp. 7115–7123.

Wu, Y., Sun, D., Wang, Y., Wang, Y., 2016. Subcomponents and connectivity of the inferior fronto-occipital fasciculus revealed by diffusion spectrum imaging fiber tracking. Front. Neuroanat. 10. doi:10.3389/fnana.2016.00088.

Wu, Y., Sun, D., Wang, Y., Wang, Y., Wang, Y., 2016. Tracing short connections of the temporo-parieto-occipital region in the human brain using diffusion spectrum imaging and fiber dissection – ScienceDirect. Brain Res. 1646 (0006-8993), 152–159.

Yeatman, J.D., Dougherty, R.F., Myall, N.J., Wandell, B.A., Feldman, H.M., 2012. Tract Profiles of White Matter Properties: Automating Fiber-Tract Quantification. PLoS ONE 7 (11), e49790+. doi:10.1371/journal.pone.0049790.

Yeh, F.-C., Panesar, S., Fernandes, D., Meola, A., Yoshino, M., Fernandez-Miranda, J.C., Vettel, J.M., Verstynen, T., 2018. Population-averaged atlas of the macroscale human structural connectome and its network topology. NeuroImage 178, 57–68. doi:10.1016/j.neuroimage.2018.05.027.

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. doi:10.3389/fninf.2011.00023.

Yoo, S.W., Guevara, P., Jeong, Y., Yoo, K., Shin, J.S., Mangin, J.-F., Seong, J.-K., 2015. An example-based multi-atlas approach to automatic labeling of white matter tracts. PLoS One 10 (7). doi:10.1371/journal.pone.0133337.

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. doi:10.1016/j.neuroimage.2018.06.027.

Zhang, L., Liu, J.Z., Dean, D., Sahgal, V., Yue, G.H., 2006. A three-dimensional fractal analysis method for quantifying white matter structure in human brain. J. Neurosci. Methods 150 (2), 242–253. doi:10.1016/j.jneumeth.2005.06.021.

Zhang, Y., Zhang, J., Oishi, K., Faria, A.V., Jiang, H., Li, X., Akhter, K., Rosa-Neto, P., Pike, G.B., Evans, A., Others, 2010. Atlas-guided tract reconstruction for automated and comprehensive examination of the white matter anatomy. Neuroimage 52 (4), 1289–1301.