



**HAL**  
open science

# Supervised tractogram filtering using Geometric Deep Learning

Pietro Astolfi, Ruben Verhagen, Laurent Petit, Emanuele Olivetti, Silvio Sarubbo, Jonathan Masci, Davide Boscaini, Paolo Avesani

► **To cite this version:**

Pietro Astolfi, Ruben Verhagen, Laurent Petit, Emanuele Olivetti, Silvio Sarubbo, et al.. Supervised tractogram filtering using Geometric Deep Learning. *Medical Image Analysis*, 2023, 90, pp.102893. 10.1016/j.media.2023.102893 . hal-04232375

**HAL Id: hal-04232375**

**<https://hal.science/hal-04232375v1>**

Submitted on 8 Oct 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Supervised Tractogram Filtering using Geometric Deep Learning\*

Pietro Astolfi<sup>a,b,c</sup>, Ruben Verhagen<sup>a</sup>, Laurent Petit<sup>d</sup>, Emanuele Olivetti<sup>a,c</sup>, Silvio Sarubbo<sup>f,c</sup>, Jonathan Masci<sup>e</sup>, Davide Boscaïni<sup>a</sup>, Paolo Avesani<sup>a,c,\*</sup>

<sup>a</sup>NILab, TeV, Fondazione Bruno Kessler, Trento, Italy

<sup>b</sup>PAVIS, Istituto Italiano di Tecnologia, Genova, Italy

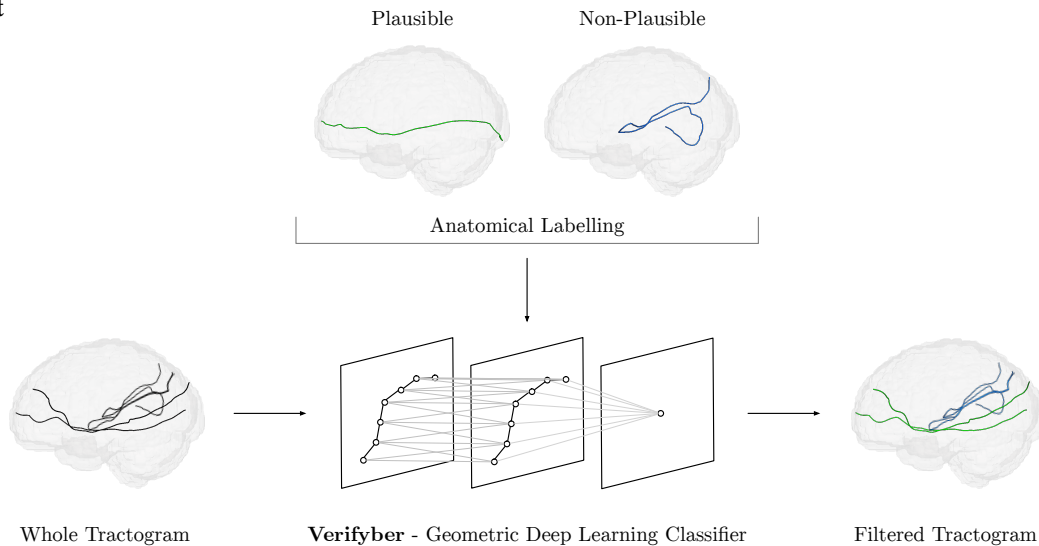
<sup>c</sup>Center for Mind/Brain Sciences (CiMeC), University of Trento, Rovereto, Italy

<sup>d</sup>GIN, IMN, CNRS, CEA, Université de Bordeaux, Bordeaux, France

<sup>e</sup>NNAISENSE, Lugano, Switzerland

<sup>f</sup>Department of Neurosurgery, Azienda Provinciale per i Servizi Sanitari, “Santa Chiara” Hospital, Trento, Italy

## Abstract



A tractogram is a virtual representation of the brain white matter. It is composed of millions of virtual fibers, encoded as 3D polylines, which approximate the white matter axonal pathways. To date, tractograms are the most accurate white matter representation and thus are used for tasks like presurgical planning and investigations of neuroplasticity, brain disorders, or brain networks. However, it is a well-known issue that a large portion of tractogram fibers is not anatomically plausible and can be considered artifacts of the tracking procedure. With Verifyber, we tackle the problem of filtering out such non-plausible fibers using a novel fully-supervised learning approach. Differently from other approaches based on signal reconstruction and/or brain topology regularization, we guide our method with the existing anatomical knowledge of the white matter. Using tractograms annotated according to anatomical principles, we train our model, Verifyber, to classify fibers as either anatomically plausible or non-plausible. The proposed Verifyber model is an original Geometric Deep Learning method that can deal with variable size fibers, while being invariant to fiber orientation. Our model considers each fiber as a graph of points, and by learning features of the edges between consecutive points via the proposed sequence Edge Convolution, it can capture the underlying anatomical properties. The output filtering results highly accurate and robust across an extensive set of experiments, and fast; with a 12GB GPU, filtering a tractogram of 1M fibers requires less than a minute. Verifyber implementation and trained models are available at <https://github.com/FBK-NILab/verifyber>.

**Keywords:** Tractography, Deep Learning, Graph Neural Networks, Tractogram Filtering

## 1. Introduction

The purpose of this work is to leverage the brain anatomical knowledge to design a method for tractogram filtering based on the notion of non-plausible pathways of the white matter. This challenge is approached with a geometric deep learning model to better capture and learn the structural properties of the brain fibers and to provide fast tractogram filtering at run time.

A tractogram provides an explicit representation of the brain connectivity structure in the white matter (Basser et al., 2000). It is composed of a collection of fibers, usually of the order of  $10^{5-6}$ , which encode the main axonal pathways. Each fiber, sometimes also called streamline, is represented as a 3D polyline by a sequence of points of variable length. Tractograms play a key role in both neuroanatomical studies (Henderson et al., 2020; Maffei et al., 2018; Jeurissen et al., 2017; Hau et al., 2017; De Benedictis et al., 2016) and brain network studies (Zhang et al., 2022; Yeh et al., 2020). The characterization of the structural brain connectivity aims to identify the bundles of fibers with a specific functional purpose. A neuroanatomical bundle is obtained by segmenting the relevant portion of fibers from the tractogram. The detection of main bundles represents an important step in the process of neurosurgical planning (Yang et al., 2021; Henderson et al., 2020; Essayed et al., 2017). Also brain network studies take as input a tractogram to compute the connectome. Given a parcellation of the brain cortex, the connectome is obtained by computing the adjacency matrix containing in each cell an estimate of the connectivity between two parcels based on the fibers connecting them. Both bundle segmentation and connectome computation are very sensitive to the accuracy of the tractogram (Buchanan et al., 2014; Rheault et al., 2020a). For this reason the quality assurance of the fiber pathways is being an open challenge for the scientific community (Jörgens et al., 2021; Zhang et al., 2022; Rheault et al., 2020b; Jeurissen et al., 2017; Maier-Hein et al., 2017; Thomas et al., 2014).

Tractograms are data derivatives. This type of data is the outcome of a complex pipeline of data processing. After a step of diffusion MRI preprocessing (Glasser et al., 2013;

Fischl, 2012; Jenkinson et al., 2012; Tournier et al., 2019) and a step of diffusivity model reconstruction (Pierpaoli et al., 1996; Tournier et al., 2007; Descoteaux, 2015), a further step of tracking (Mori et al., 1999; Basser et al., 2000; Jeurissen et al., 2017) is in charge of the computation of the fiber pathways. The process of tracking requires the definition of many parameters like the policy of seeding, the strategy of stepping, the stop criterion, the constraints on curvature and length of fibers. Slightly different choices may produce quite different tractograms (Thomas et al., 2014), and the evaluation of their accuracy is not straightforward (Neher et al., 2015).

In the last years the issues of tractography reliability and reproducibility have been approached with several data analysis contests: FiberCup on 2011 (Fillard et al., 2011), Fiberfox on 2015 (Maier-Hein et al., 2017), Traced on 2017 (Nath et al., 2020) and Votem on 2018 (Schilling et al., 2019). The design of these contests is quite similar and it is based on the use of a phantom to define the ground truth in advance (Côté et al., 2013). The evaluation is carried out by measuring the mismatch between the synthetic model and the tractograms computed using the state of the art methods. These initiatives achieved a general agreement on the main limitations of the tractography techniques.

A recurrent weakness of tracking algorithms is the generation of false positive fibers (Maier-Hein et al., 2017). Indeed, the evolution of the diffusivity model reconstruction from DTI (Pierpaoli et al., 1996) to HARDI (Descoteaux et al., 2007) improved the sensitivity of tracking (missing less existing pathways), but at the cost of increasing the number of fibers that are not anatomically plausible (Thomas et al., 2014). Also the common practice of overtracking contributes to the decreasing of tracking specificity. One of the reasons for this issue is that fiber density is not coherent with the physiological distribution of axons (Raffelt et al., 2012; Yeh et al., 2016). While for neuroanatomical studies it is crucial to have high sensitivity tractograms, brain connectivity studies might be meaningfully affected by low specificity tractograms where false positive fibers may impact twice with respect to false negatives (Zalesky et al., 2016). Finding a balance between sensitivity and specificity in fiber tracking algorithms is still an open problem.

*State of the art of tractogram filtering.*

The recent literature provides many contributions to im-

---

\*Extended version of MICCAI publication Astolfi et al. (2020).

\*Corresponding author: e-mail: [avesani@fbk.eu](mailto:avesani@fbk.eu)

prove the accuracy of the tractograms. The proposed methods can be divided into two main groups according to their strategy. The first strategy could be referred to as *ex-ante* and includes all the attempts to revise the tracking algorithms with the goal of reducing the generation of non-plausible fibers. The other strategy is approaching the problem *ex-post*, by filtering out artifactual fibers once the tractogram is computed.

According to the *ex-ante* strategy, a quantitative study (Bastiani et al., 2012) was carried out to investigate how the choice of hyperparameters might impact the results of tracking algorithms. The tuning of heuristics, especially in stochastic methods for probabilistic tracking, introduces significant source of variability and a critical dependency from the choice of parameters’s values. The general trend to improve the tracking is to make sure that such heuristics are anatomically informed. For example, seeding and stopping criteria have been revised to be driven by the gray matter and white matter interface (Smith et al., 2012; Girard et al., 2014; Lemkaddem et al., 2014) or the tracking has been constrained using topographic regularity (Aydogan and Shi, 2018). The challenge becomes how to inject anatomical priors in the tracking algorithms. It turned out to be easier and effective to elicit anatomical constraint in the case of bundle specific tracking (Yendiki et al., 2011; Thomas et al., 2014; Chamberland et al., 2017; Rheault et al., 2019). In these restricted cases the fiber pathways are driven by volumetric ROI defined according to the anatomical knowledge of a specific bundle. However, *ex-ante* strategies remains non-successful for the whole brain tractography (Schilling et al., 2020).

The task of tractogram filtering, which approaches the problem with an *ex-post* strategy, adopts a global view and considers as input the whole brain tractogram. We may consider tractogram filtering as a complementary step to be combined with a better tuning of the tracking algorithms, since *ex-ante* and *ex-post* strategies are not in contrast or mutually exclusive. We may distinguish two types of filtering solutions: signal-based and tractography-based. The former solutions formulate the filtering task as an inverse problem of signal reconstruction, the latter solutions adopt filtering criteria based only on the tractography data. Both of them carry out an unsupervised strategy.

In the signal-based solutions the plausibility of fibers is estimated by computing how much their pathways are explained by the diffusion signal. The most common meth-

ods cast the filtering task as a global regularization problem by assigning a weight to each fiber. In SIFT (Smith et al., 2013) and SIFT2 (Smith et al., 2015a), the weights are a proxy of the fibers density. In LiFE (Pestilli et al., 2014) the weights capture how much the fiber pathways are related to the diffusion signal. COMMIT (Daducci et al., 2015) extends the estimate of weights by including microstructure information. In all these methods the thresholding of weights to discriminate between plausible and non-plausible fibers is managed with heuristics. Nevertheless, as remarked by Smith et al. (2020); Frigo et al. (2020); Rheault et al. (2019), the filtering operated with a regularization approach might remove fibers whose pathway is anatomically plausible.

The alternative approaches are based only on tractography. Their basic assumption is that the topographic regularity of tractogram structures across individuals might be a good proxy of anatomical plausibility. Different unsupervised methods, by leveraging the groupwise consistency of fiber bundles, have been proposed to detect outlier pathways (O’Donnell and Westin, 2007; Wang et al., 2018; Xia and Shi, 2020). They differ in the definition of the proximity metrics for the computation of topographic regularity. The filtering in these cases is subject to the population bias due to the lack of general anatomically-informed priors. For this reason, the groupwise analysis is usually limited to the anatomy of well-known fiber bundles. While a large population of tractograms provides a more robust estimate of structure regularity, the consistency constraints tend to eliminate inter-individual differences. To contrast the smoothing effect due to population averaging, other unsupervised methods refer to the fiber density map (Yeh et al., 2019) or to the local fiber agreement (Chandio et al., 2022) as a proxy of anatomical plausibility. The focus in these methods is to exploit the intra-individual information.

The two distinct approaches, signal-based and tractography-based, have been combined to design mixed solutions (Aydogan and Shi, 2015; Neher et al., 2018; Nie and Shi, 2019; Schiavi et al., 2020; Ocampo-Pineda et al., 2021). The basic intuition is that where a priori knowledge of neuroanatomical bundles is not available, the fibers are regularized by a signal-based filtering, while along the pathways of known bundles, the fibers are filtered out when they do not meet the expected topographic regularity.

The most recent trend in unsupervised methods for tractogram filtering is the investigation of deep learning tech-

niques. FINTA (Legarreta et al., 2021) proposes a convolutional neural network to learn an embedded representation of the fibers. After the learning procedure, the fibers are projected into a new latent space where the computation of nearest neighbors might easily detect the similarity as proximity. However, the learning of the embedding is not driven by neuroanatomical knowledge and the subsequent filtering of fiber is not guaranteed to properly capture the notion of anatomical plausibility.

As a general remark of the state of the art, we may notice that all methods following an *ex-post* strategy to tractogram filtering are adopting an unsupervised approach, and the design of a loss function suitable to capture the notion of anatomical plausibility remains an open challenge. For this reason, the common trend is to integrate additional neuroanatomical constraints, both in *ex-post* (Neher et al., 2018; Nie and Shi, 2019; Schiavi et al., 2020; Ocampo-Pineda et al., 2021) and *ex-ante* (Rheault et al., 2019) strategies.

#### *Our contributions.*

In this work we propose Verifyber, a novel tractography-based method to perform *ex-post* filtering of non-plausible fibers from a tractogram. The task of tractogram filtering is shaped as a supervised learning problem where a binary classifier takes in input a fiber and provides as output either the category anatomically plausible or anatomically non-plausible. We present an original learning model based on geometric deep learning (GDL) (Masci et al., 2016; Bronstein et al., 2017), which better fits the learning on 3D data without forcing Euclidean vector representations. The notion of anatomical plausibility is derived from fiber examples, labelled either as anatomically plausible or anatomically non-plausible.

We envision the task of elicitation of brain knowledge as a binary labelling of fibers. Despite the evolutionary nature of the knowledge of the human brain, we may conceive two main labelling policies: *inclusive* and *exclusive*. The *inclusive* policy leans to be more conservative and aims to prevent false positives. According to this prior only fibers following the pathways of well known bundles are labelled as anatomically plausible, non-plausible otherwise. Conversely, the *exclusive* policy is more sensitive to the false negative, in this case only fibers with clear artifactual pathways are labelled anatomically non-plausible. It is out of the scope of this work to establish which policy

might be more effective and appropriate. Our goal is to investigate whether the proposed method is equally robust for the two policies.

Our empirical analysis is considering datasets labelled with both *inclusive* and *exclusive* policies. As a reference example of *inclusive* policy we point to an anatomically curated white matter atlas (Zhang et al., 2018). This atlas provides a whole brain tractogram averaged over 100 individuals. A team of experts manually curated the annotation of 74 bundles. For our purpose we considered anatomically plausible all the fibers of those bundles, non-plausible otherwise.

In the literature the *exclusive* policy is less common. As an instance of this kind we consider Extractor (Petit et al., 2022). In such a work, the notion of anatomical non-plausibility is defined by a set of heuristic rules based on the current knowledge of the human white matter. Well-known artifactual pathways based on geometric properties or brain locations are labelled as anatomically non-plausible, usually half portion of the whole tractogram. This declarative knowledge can be applied to any tractogram enabling the annotation of training and test sets for learning purposes.

Regardless of the source of labelling, the challenge of supervised learning is to train a binary classifier based only on a digital representation of fibers. The choice of an appropriate representation of a fiber, suitable for the learning process, is a crucial step. Usually a fiber is encoded as an ordered sequence of a variable number of 3D points. Previous works on supervised learning for tractography had to deal with the constraint of learning algorithms that require a fixed length embedding. The most common solutions are the computation of an Euclidean embedding such as dissimilarity representation (Olivetti and Avesani, 2011; Bertò et al., 2021). Unfortunately these fiber embeddings are lossy.

To overcome such limitations and to preserve the full geometrical information encoded in the fiber pathways, we propose to investigate the use of Geometric Deep Learning models like PointNet (Qi et al., 2017) and Dynamic Graph CNN (Wang et al., 2019), which by construction can deal with variable size inputs like point clouds and graphs. GDL architectures are based on layers of permutation invariant/equivariant operators whose combination allows a model to perform convolution in a non-grid (and non-Euclidean) representation. Our working hypothesis is

that GDL might be more accurate in capturing the geometrical properties of pathways associated with the notion of anatomical plausibility.

Verifyber is designed as a novel end-to-end trainable GDL model to deal with traditional encoding of fibers as ordered sequences of a variable number of 3D points. Our proposal extends the Edge Convolution (EC) layer by Wang et al. (2019) to take into account the information encoded by the edges between two subsequent 3D points in a fiber. Then, the architecture is composed by a global pooling layer that compresses each fiber to a single descriptor and a classification head to discriminate between the two categories, either anatomically plausible or non-plausible. Differently from EC, our model is sequence sensitive, i.e., not permutation invariant, while it remains invariant to the orientation of the input fibers.

We provide the results of a broad set of experiments aimed at proving the properties of the proposed model and assessing the efficacy in discriminating anatomically plausible and non-plausible fibers. We show that Verifyber outperforms in accuracy competing deep learning methods such as bidirectional LSTM (Graves and Schmidhuber, 2005; Huang et al., 2015), PointNet (Qi et al., 2017), and Dynamic Graph CNN (Wang et al., 2019). These results are robust with respect to different types of tractography and equally effective on *inclusive* and *exclusive* policies to elicit the notion of neuroanatomical plausibility. An additional comparison is aimed to show the different behaviour of supervised and unsupervised filtering approaches, these lasts represented by FINTA (Legarreta et al., 2021). We also investigate how a trained model behaves across different sources of tractograms, when the computation of tracking is not necessarily homogeneous. We show some preliminary results of this kind of analysis on a clinical dataset. An additional simulation study allows the evaluation of the behaviour of the method when the labeling of fiber is dynamically evolving over time.

## 2. Method

In this section, we describe our method, Verifyber. For the sake of comprehension, we also summarize the Edge Convolution (EC) layer (Wang et al., 2019), which is a building block of our model on top of which we built the proposed *sequence EC*.

### 2.1. Edge Convolution Layer

Considering a point cloud  $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ ,  $\mathbf{x}_i \in \mathbb{R}^3$ , an Edge Convolution (EC) layer first induces a graph structure for  $\mathcal{X}$  by retrieving for each point  $\mathbf{x}_i$  the set of  $k$  nearest neighbors,  $\text{knn}(\mathbf{x}_i) = \{\mathbf{x}_{j_{i_1}}, \dots, \mathbf{x}_{j_{i_k}}\}$ , using the Euclidean distance as metrics (see Figure 1a). The result is a  $k$ -nn graph composed of  $\mathcal{V}$  nodes and  $\mathcal{E}$  edges:

$$\mathcal{G}(\mathcal{V}, \mathcal{E}), \mathcal{V} = \mathcal{X}, e_{ij} \in \mathcal{E}: \mathbf{x}_i \rightarrow \mathbf{x}_j \iff \mathbf{x}_j \in \text{knn}(\mathbf{x}_i). \quad (1)$$

Then, each point representation,  $\mathbf{x}_i$ , is enriched with the representation of each of its neighbors,  $\mathbf{x}_{j_i}$ , to obtain edge features,  $\mathbf{e}_{ij}$ , which are learnt through a neural network  $h_{\Theta}$ . Specifically,

$$\mathbf{e}_{ij} = h_{\Theta}(\mathbf{x}_i \oplus (\mathbf{x}_j - \mathbf{x}_i)), \quad (2)$$

where  $\oplus$  denotes the concatenation operator. Finally, a new representation of a point,  $\mathbf{x}'_i$ , is obtained by aggregating all the learned edge features with a pooling operator, i.e.,  $\mathbf{x}'_i = \text{pool}(\mathbf{e}_{ij})$ ,  $j: (i, j) \in \mathcal{E}$ , where *pool* is either *max* or *mean*.

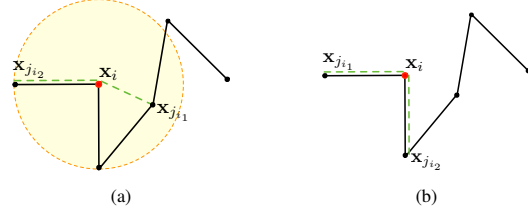


Figure 1: Comparison between Euclidean  $k$ -nn (a), and graph  $k$ -nn on the streamline (b).

### 2.2. Sequence Edge Convolution Layer

A remarkable property of the EC layer is the invariance to the permutation of the points in the input point cloud. Indeed, the layer contains only operators invariant to the points ordering, e.g., FC layers, *max* / *mean* pooling, Euclidean  $k$ -nn. Although this property is fundamental in the point cloud domain, it becomes undesired if the input is a sequence as in our case. To solve this issue, we propose a simple but well-motivated modification: we substitute the Euclidean  $k$ -nn, which was inducing a graph structure based on euclidean distance of the points, with a graph-based  $k$ -nn (see Figure 1b) that instead computes neighbors

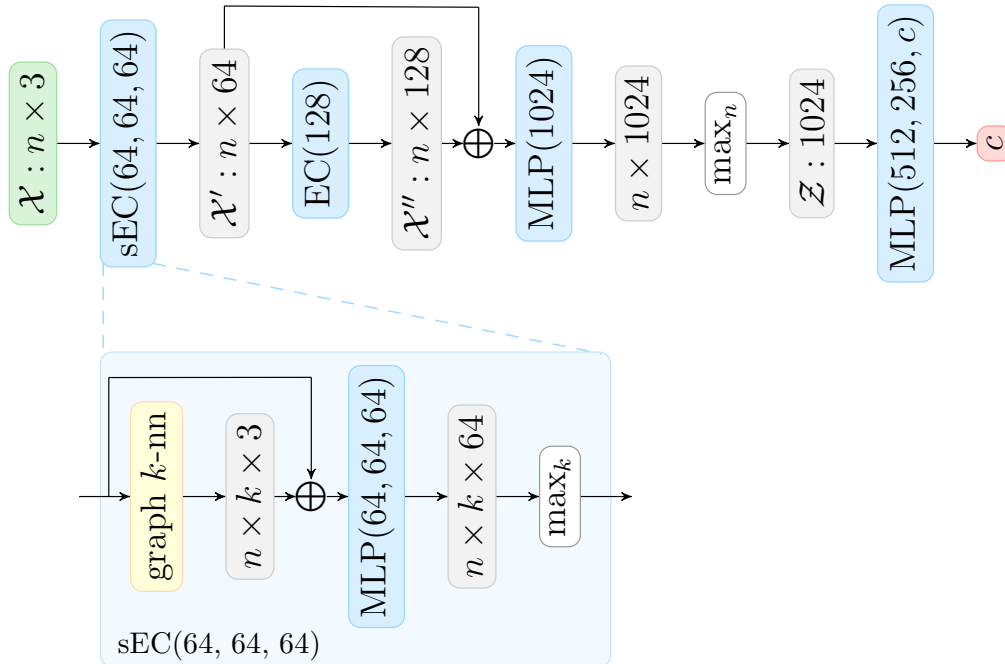


Figure 2: Block diagram of the Verifyber architecture. Green, gray, and red blocks represent input, intermediate, and output tensors, respectively. Parametric layers are colored in blue, while non-parametric layers in white. In yellow we highlight the graph  $k$ -nn that allows the model to be sequence sensitive.

along the existing input graph. Considering the streamline structure, the graph-based  $k$ -nn preserves the input graph to be a bidirectional sequence of points where each non-terminal point,  $\mathbf{x}_{i \neq 0, n}$ , has two neighbors: the previous and the next point in the sequence, while the terminal points,  $\mathbf{x}_0, \mathbf{x}_n$ , have just one neighbor:

$$\mathcal{G}(\mathcal{V}, \mathcal{E}'), e'_{ij} \in \mathcal{E}' : \mathbf{x}_i \rightarrow \mathbf{x}_j, \iff j = i + 1 \vee j = i - 1. \quad (3)$$

By using this graph structure an EC layer loses the invariance to the input permutations, while maintaining the invariance with respect to the input flipping (a crucial property when dealing with streamlines), thanks to the bidirectionality of the edges. For this reason, we define this modified EC layer as *sequence* EC (sEC) layer.

### 2.3. Verifyber Model

The Verifyber (VF) model is characterized by the stacking one sEC layer with one or more EC layers. Figure 2

shows the block diagram of the architecture of the Verifyber model we used for our experiments in Section 5. One sEC layer and one EC layer are stacked to produce new representations  $\mathcal{X}'$  and  $\mathcal{X}''$  with 64 and 128 features respectively. The stacking of these two layers guarantees the model to be both sequence sensitive and *dynamic*: as shown in (Wang et al., 2019), the computation of knn in latent space allows a dynamic adjustment of the local neighborhood of points guided by the optimization of the task at hand. Then,  $\mathcal{X}'$  and  $\mathcal{X}''$  are concatenated, encoded to 1024 features with a learning layer  $g_{\Phi}$ , and pooled to obtain a single descriptor of the whole point cloud,

$$\mathbf{z} = \text{pool}(g_{\Phi}(\mathcal{X}' \oplus \mathcal{X}'')). \quad (4)$$

Finally, the 1024-dimensional feature vector  $\mathbf{z}$  is classified using a fully connected (FC) network composed of three layers, which decreases the number of features to 512, 256, and  $c$  (number of classes), respectively.

### 3. Related Works

The proposed Verifyber model is the result of a step-by-step investigation aiming to solve a well-known problem in tractography analysis: finding a data representation compliant with computational requirements. Performing automated analysis of a tractogram requires a method able to deal with the structure of streamlines. Such a structure presents some characteristics which differ from the typical neuroimaging data representations like images and volumes: a streamline is a sequence of points with variable length and no orientation.

*Streamline embedding and traditional models.* Streamlines' properties prevent traditional machine learning methods from being directly applied to them. One common requirement of machine learning methods is to have fixed length vectors as input, and thus existing works resorted to different preprocessing solutions to match such a requirement. A widely adopted heuristics consists in resampling the streamlines to a fixed number of points (Garyfallidis et al., 2012, 2018; Gupta et al., 2017; O'Donnell and Westin, 2007; Legarreta et al., 2021). However, for traditional methods like support vector machines or linear classifiers, the resulting 3-dimensional vectors with fixed length, ( $\#$  points,  $(x, y, z)$ ), are not suitable, and they need to employ a more advanced embedding technique to project streamlines in a new convenient space. An example are embeddings based on dissimilarity representation like Olivetti and Avesani (2011) and more recently Bertò et al. (2021) that also considers handcrafted features based on the white matter anatomy, or using non-linear dimensionality reduction techniques, e.g., t-SNE (Van der Maaten and Hinton, 2008), like in Chandio et al. (2022). These embeddings enable the training of traditional classifiers at the cost of losing some geometrical/structural information of the streamlines, e.g., the presence of a loop.

*Convolutional Neural Networks for streamlines.* The limitation of lossy embeddings might be bypassed using deep learning techniques that are able to learn embeddings based on the target task. Given the breakthrough of CNNs in computer vision, there have been recent attempts to apply them also to streamlines. FINTA (Legarreta et al., 2021) proposes an unsupervised approach, where the embedding is learned by means of a convolutional autoencoder, and then it is used as input for the downstream

task, e.g., tractogram filtering. Even though the learned embedding might preserve the structural information of streamlines, the lack of a task-specific supervision does not guarantee its optimality for the downstream task. FiberNet (Gupta et al., 2017) and FiberNet 2.0 (Gupta et al., 2018), Deep CNN (DCNN) tract classification (Xu et al., 2019; Lee et al., 2020) and Deep White Matter Analysis (deepWMA) (Zhang et al., 2020), instead, train standard CNN models like AlexNet (Krizhevsky et al., 2012) and ResNet (He et al., 2016) directly on streamlines using bundle supervision. In this way they learn an embedding specific for the bundle segmentation task. However, we notice a controversial use of convolutional filters in FiberNet and DCNN, as they treat streamlines of size ( $\#$  points,  $(x, y, z)$ ) like images of size (height, width), i.e., width = 3. More correctly, as operated in FINTA and DeepWMA,  $(x, y, z)$  should be considered channels like  $(r, g, b)$  in images so that different filters are learned for each channel. Finally, we also notice a general drawback when using standard CNNs to perform learning on the streamline structure. Indeed, CNNs are translation invariant (or equivariant), a property that is crucial for the image domain. However, streamlines, unlike images, can be drastically affected by translating points, and thus the translation invariance/equivariance of CNNs is not a desired property. Streamlines, instead, being unoriented, require flip invariance, but unfortunately this is not a property of standard CNNs despite it is neglected by the approaches mentioned above.

*Recurrent Neural Networks and Geometric Deep Learning.* Based on these premises we investigate supervised deep learning approaches different from standard CNNs. We seek for neural network architectures more suitable for the streamline structure. Guided by the sequentiality of the streamline structure, we start our investigation from setting a baseline using a Recurrent Neural Network (RNN) model. Then, with the aim to have a flip invariant model able that deals with size-varying input, we explore some methods from the family of Geometric Deep Learning (GDL) (Masci et al., 2016; Bronstein et al., 2017). GDL comprises all the methods that extend convolution principles to non-Euclidean data, e.g., non grid-based data like point clouds and graphs. To deal with such data, GDL models exploit modules and operators that permutation invariant instead of translation invariant and that can be



applied to batches of size-varying samples.

*Bidirectional LSTM (Graves and Schmidhuber, 2005).* In the literature of RNN methods, especially in the field of Natural Language Processing where data has a sequential structure, a large number of methods is based on Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997). Among all, we individuate the bidirectional LSTM (bLSTM) (Graves and Schmidhuber, 2005; Huang et al., 2015) as a reference deep learning method to analyze streamlines. bLSTM is characterized by two LSTM layers, each of them fed with a different orientation of the input. It learns a shared embedding of both orientations by combining the two LSTM outputs with an aggregator operator (concatenation) and then forwards it to a FC network, which performs classification. However, there are two main limitations of the bLSTM method when applied to streamlines. First, it requires a fixed-length vector as input to its LSTM layers, and second it is not invariant to the input flipping despite the bidirectional architecture. Indeed, the two LSTM layers learn two different set of parameters, which may produce different hidden states if fed with the same sequence. Nevertheless, the use of the two directions is still beneficial for the network because it improves the learning of local context information. Eventually, bidirectionality combined with an augmented training where streamlines are given in both orientations might mitigate the lack of flip invariant layers.

The limitations of bLSTM are not present in GDL methods, which by construction are flip invariant (special case of permutation invariance) and can be fed with size-varying point clouds or graphs. Hence, if we model a streamline as a point cloud we are neglecting its sequential structure (since point clouds do not assume any ordering of their vertices), but we maintain both the streamline spatial information and its invariance to the flipping of orientation.

*PointNet (Qi et al., 2017).* In our experiments, we investigate the pioneer and most adopted model for point cloud, namely PointNet (PN) (Qi et al., 2017). PN is characterized by a simple architecture composed only of FC layers and pooling layers that are by construction permutation invariant. In particular, for the task of classification PN presents a series of FC layers as encoder, a max pooling layer that generates a single global feature vector of the input point cloud, and another series of FC layers performing

the output classification. However, learning on streamlines using PN could be limited due to the non-consideration of point relations. Indeed, PN is only able to consider a global relation among all the points by performing the max pooling in latent space. For this reason we decided to investigate also a GDL model that consider points relation as encoded by graph structures, namely Dynamic Graph CNN (DGCNN) (Wang et al., 2019).

*Dynamic Graph CNN (Wang et al., 2019).* The DGCNN model is, according to Wang et al. a generalization of PN. Instead of considering a single global (all to all) relation, DGCNN considers multiple local neighborhood relations, like in a  $k$ -nn graph structure, computed at different depths of the network, i.e., in different latent spaces. The model is based on Edge Convolution layers (explained in Section 2.1) which have deeply inspired our Verifyber. However, since DGCNN makes only use of EC layers (plus the classification decoder), it is a permutation invariant model as well as PN. These models cannot distinguish two streamlines whose points are randomly shuffled, and this is an undesired behaviour for the tractogram filtering task.

*Verifyber.* Our contribution sEC allows the proposed Verifyber model to overcome the permutation invariance limitation while remaining orientation invariant. Also, VF inherits the other good properties of PN and DGCNN, resulting able to work with size-varying input and to consider point relations.

#### 4. Material

In this section, we present the datasets used for the empirical analyses. A summary is reported in Table 1. For each dataset, we report the data source and the type of labelling, either with inclusive or exclusive policies. The source data are mainly drawn from the Human Connectome Project (HCP) (Van Essen et al., 2013) repository.

*Dataset HCP-EP.* This dataset is composed of 20 individuals randomly selected from HCP repository. The processing pipeline carried out the estimation of the diffusivity model using the Constrained Spherical Deconvolution (CSD) (Tournier et al., 2007), and the fiber tracking using the algorithm for Particle Filtering Anatomically Constrained Tractography (PF-ACT) (Girard et al., 2014).

Table 1: Summary of the adopted datasets. B: number of bundles. #: number of tractograms. T: number of fibers in a tractogram.

Name	Source	$p/np$ label	B	#	T	Track	DWI
<i>HCP-EP</i> <sup>*</sup>	HCP	Exclusive Petit et al. (2022)	-	20	1M	CSD PF-ACT	3T DWI, 1.25mm, 270g multi- b=(1,2,3)K
<i>HCP-IZ</i> <sup>†</sup>	HCP	Inclusive Zhang et al. (2018)	74	1 <sup>§</sup>	1M	HARDI UKF	3T DWI, 1.25mm, 108g single-b=3K
<i>HCP-IW</i> <sup>‡</sup>	HCP	Inclusive Wasserthal et al. (2018)	72	23	10M	CSD iFOD2	3T DWI, 1.25mm, 270g multi- b=(1,2,3)K
<i>APSS-IS</i>	APSS	Inclusive expert: S.S.	4	5	100K	DTI EuDX	1.5T DWI, 2.5mm, 60g single-b=1K

<sup>\*</sup> <https://doi.org/10.25663/brainlife.pub.13>    <sup>†</sup> <https://github.com/SlicerDMRI/ORG-Atlases>

<sup>‡</sup> <https://zenodo.org/record/1477956#.Ya67UyzMKL8>    <sup>§</sup> averaged from 100 subjects.

More in detail, the tracking generated around  $\sim 1M$  streamlines for each tractogram by seeding 16 points for each voxel with step size 0.5mm. Tractograms were normalized to the same space via non-linear co-registration to the MNI152 standard brain (Fonov et al., 2011). For computational purposes, all the streamlines have been compressed to the most significant points (Presseau et al., 2015).

For this dataset, the labelling was performed using Extractor (Petit et al., 2022), a rule-based method that implements an *exclusive* policy to elicit the notion of anatomical plausibility. In particular, the criteria encoded by the rules are driven by the definition of non-plausibility. Non-plausible streamlines are identified with several heuristics: (i) those streamlines that are shorter than 20 mm, or contain a loop, or are truncated, i.e., they terminate before reaching the WM/GM interface; (ii) outlier streamlines with respect to clustering (Côté et al., 2015) of the three main categories of pathways, i.e., associative, projective, commissural. According to a conservative approach, all the remaining streamlines are labeled as anatomical plausible.

*Dataset HCP-IZ.* The second dataset adopted to test our method is again HCP-based, but in this case, composed of only one averaged brain: tractogram and structural T1 image. The average comes from a set of 100 HCP subjects, for which the tracking has been performed on the estimated diffusivity model (Descoteaux et al., 2007) using

the Unscented Kalman Filter (UKF) Tractography (Reddy and Rathi, 2016). 10K streamlines were randomly selected from each subject’s brain, resulting in a merged tractogram composed of  $\sim 1M$  streamlines. The merge was possible after a step of streamline-based linear registration (O’Donnell et al., 2012), which moved all the tractogram to the space of one arbitrarily picked subject. The same affine transformation was applied to the structural T1w images of subjects. Finally, an average T1w was computed by merging all the subjects’ T1w through a simple mean operation.

This dataset is presented in Zhang et al. (2018) as an atlas of white matter bundles. The bundles are extracted from the average tractogram using the White Matter Analysis clustering (O’Donnell and Westin, 2007). First, 800 clusters are generated, and then they are visually inspected and merged to obtain 74 different classes of bundles (see Zhang et al. (2018) for the full list), including 16 classes of superficial U-shape streamlines. However, in this procedure, almost 300 clusters are not merged into a bundle because composed of unknown pathways. We consider all these unknown streamlines as anatomically non-plausible. Moreover, we use the very high number of bundle classes to create multiple split of  $p$  and  $np$  streamlines, emulating a real-world scenario where the labeling evolves incrementally in time. Each split considers the streamlines belonging to certain classes of bundles as plausible and all

the others as non-plausible.

*Dataset HCP-IW.* With the aim of proving the impact of our filtering method, we adopt a third HCP-based dataset published along with TractSeg (Wasserthal et al., 2018, 2019). This dataset is composed of 23 tractograms non-overlapping with the ones of the other HCP-based datasets. The tractograms were obtained using multi-shell multi-tissue CSD model estimation and the Second-order Integration over Fiber Orientation Distributions (iFOD2) probabilistic tracking with MRtrix (Tournier et al., 2019). The tracking was performed by: (i) random seeding within the masked brain, (ii) pruning streamlines shorter than 40mm, (iii) cropping streamlines at the GM/WM interface, and (iv) stopping after reconstructing 10M streamlines. Moreover, the tracking was executed twice per subject, once considering anatomical constraints and once not.

This HCP-IW dataset is one of the few benchmark datasets for bundle segmentation. It contains the labeling of 72 white matter bundles per tractogram. Given the lack of  $p$  versus  $np$  categorization in this dataset (as in all the other publicly available datasets), we use bundles as proxy evaluators to quantitatively and qualitatively show the impact of our method.

*Dataset APSS-IS.* The last dataset we adopt is a clinical dataset obtained from the Department of Neurosurgery at the Santa Chiara Hospital (APSS) in Trento (Italy). It comprises 5 patients affected by brain tumors. For each subject, we have available the DWI and the reconstructed tractogram. The DWI was acquired with a 1.5T MR scanner using 60 directions. Then, a single shell  $b=1000$  s/mm<sup>2</sup> was extracted to reconstruct the diffusion model with DTI (Pierpaoli et al., 1996). The tracking was performed using Euler Delta Crossing (EuDX) (Garyfallidis et al., 2014) and produced approximately 100K fibers.

An expert neurosurgeon (S.S.) manually segmented bundles for clinical purposes in both the healthy and lesioned hemispheres of the patients. The manual segmentation followed an ROI-based procedure operated with TrackVis (Wang et al., 2007). Due to the different sizes and locations of tumors, the types of the segmented bundle were not consistent across patients or hemispheres. Among the available segmentations, we selected the ones that were in common with the 5 subjects. The selection resulted in three types of bundles for the healthy hemisphere: the

Arcuate Fascicle (AF), the Superior Longitudinal Fascicle (SLF), and the Inferior Fronto-Occipital Fascicle (IFOF), and one bundle in the lesioned hemisphere: Pyramidal Tract (PT).

## 5. Experiments and Results

The empirical analysis is organized into model-related and task-related experiments. Model-related experiments aim to assess the properties of the proposed model; task-related experiments are designed to investigate the effectiveness of Verifyber as a solution for the task of tractogram filtering. The performance of Verifyber are compared with a selection of the state of the art methods. The sensitivity to the sequential structure of the streamlines is carried out with an ablation study. Finally, the impact of the proposed solution on tractogram filtering is estimated both quantitatively and qualitatively by: (i) looking at the distribution of the misclassification error, (ii) considering the behavior on different types of tractograms, (iii) simulating the evolving definition of anatomical plausibility to test the adaptation in the case of concept drift, and (iv) comparing with unsupervised deep learning tractogram filtering.

### 5.1. Model related experiments

*Cross-validation analysis.* The first experiment is designed to measure the learning performances of Verifyber according to the usual setting of cross-validation for a supervised task. For this purpose, we consider the HCP-EP dataset with 20 annotated tractograms. The train and test splitting follows a 5-fold cross-validation scheme, where each fold is composed of 4 tractograms and 3.5 million fibers. For each run, the remaining 4 folds are randomly split into 4 buckets, 3 devoted to training and 1 to validation. The training procedure is designed as follows: 1K epochs; cross-entropy loss to optimize the classification; Adam optimizer with default alpha and beta momentum (0.9, 0.99); initial learning rate of  $10^{-3}$  multiplied by a factor of 0.7 every 90 epochs until a minimum value of  $5 \cdot 10^{-5}$  is reached. In each epoch, we define a mini-batch composed of 16K streamlines, randomly sampled from two subjects, 8K from each of them. A subject is sampled only once for each epoch. The evaluation of the binary classification task is carried out by measuring the accuracy, the precision, the recall, and the Dice-Sørensen coefficient (DSC). Results are reported in Table 2.

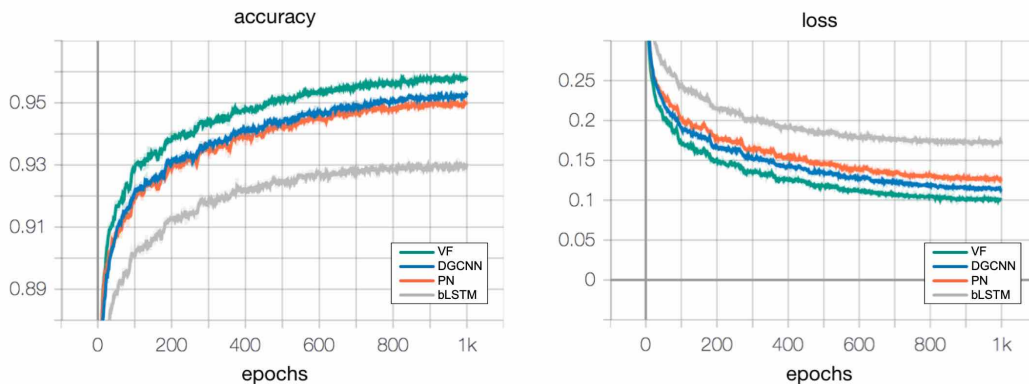


Figure 3: Training curves of the four methods compared. Despite the similar number of parameters of all the models we can observe that bLSTM (grey) has clear negative gap with respect to all the other models. Conversely, VF (green) has always a positive gap compared to all the others.

Table 2: Accuracy, precision, recall, and DSC 5-fold cross validation scores on HCP-EP dataset. Reported values are the mean and standard deviation across the 5-folds. Each fold of 4 subjects has been used once as a test.

Method	Accuracy	Precision	Recall	DSC
bLSTM	93.0 ( $\pm 0.1$ )	93.8 ( $\pm 0.1$ )	96.2 ( $\pm 0.2$ )	95.0 ( $\pm 0.1$ )
PN	94.7 ( $\pm 0.1$ )	95.5 ( $\pm 0.2$ )	96.9 ( $\pm 0.2$ )	96.2 ( $\pm 0.1$ )
DGCNN	94.4 ( $\pm 0.1$ )	95.4 ( $\pm 0.2$ )	96.5 ( $\pm 0.1$ )	96.0 ( $\pm 0.1$ )
<b>VF</b>	<b>95.2 (<math>\pm 0.1</math>)</b>	<b>96.1 (<math>\pm 0.2</math>)</b>	<b>96.9 (<math>\pm 0.1</math>)</b>	<b>96.6 (<math>\pm 0.1</math>)</b>

Table 3: Models architecture used for experiments.

Method	Architecture	params
bLSTM	MLP(128)→LSTM(256)⊕LSTM <sup>-1</sup> (256)→MLP(256,128)→FC(2)	800K
PN	MLP(64,64,64,128,1024)→MAX→MLP(512,256,40)→FC(2)	800K
DGCNN	ec1:EC(64,64,64)→EC(64,64,64,128)→ec1⊕ec2→MLP(1024)→MAX→MLP(512,256)→FC(2)	800K

*Comparison with deep learning models.* The second experiment aims to compare Verifyber with the competing deep learning models, namely bLSTM, PN, and DGCNN. As reference dataset we consider HCP-EP as above. In this experiment, we operate the 5-fold cross-validation setting adopted for Verifyber to all other methods and we measure the same evaluation metrics. We report the behavior of training curves for all models, both accuracy and loss, in Figure 3.

Regarding the input representation, GDL models can deal with size-varying input, e.g., streamlines with dif-

ferent numbers of points, while bLSTM requires a fixed vectorial representation in input, like common learning models. For this reason, we need to resample the points of all streamlines to be a fixed number. According to previous works (O’Donnell and Westin, 2007; Garyfallidis et al., 2012), the common choices are resampling to 12, 16, or 20 points per streamline. Since a side empirical assessment analysis did not provide any significant difference in performance, we operate a resampling to 16 points for all the subsequent experiments. To prevent the bias of heterogeneous size of models, we set the architecture of the

Table 4: Permutation test results on HCP-EP. Reported values refer to only one split of the 5-fold. The mean and standard deviation are computed across the 4 test subjects. Lower values are better. As expected PN and DGCNN proved the permutation invariance maintaining the same results of Table 2 of the paper.

Method	Accuracy	Precision	Recall	DSC
bLSTM <sub>perm</sub>	64.1 ( $\pm 1.1$ )	89.8 ( $\pm 1.0$ )	55.1 ( $\pm 1.1$ )	68.3 ( $\pm 0.9$ )
PN <sub>perm</sub>	94.5 ( $\pm 0.1$ )	95.4 ( $\pm 0.2$ )	96.8 ( $\pm 0.2$ )	96.1 ( $\pm 0.2$ )
DGCNN <sub>perm</sub>	94.3 ( $\pm 0.1$ )	95.4 ( $\pm 0.3$ )	96.5 ( $\pm 0.2$ )	95.9 ( $\pm 0.2$ )
<b>VF<sub>perm</sub></b>	<b>30.0</b> ( $\pm 2.8$ )	<b>87.7</b> ( $\pm 0.7$ )	<b>00.3</b> ( $\pm 0.0$ )	<b>00.6</b> ( $\pm 0.1$ )

different methods with a uniform number of parameters, as reported in Table 3.

The results of the comparison on the test set are reported in Table 2. According to the related results at the end of the training illustrated in Figure 3, we may conclude that there is no overfitting of the model during the learning process. While the performance of geometric deep learning models (PN, DGCNN, VF) are quite similar, there is a meaningful gap with respect to the recurrent neural network model. A t-test between bLSTM and PN reports a p-value  $< 10^{-3}$ .

*Permutation invariance test.* A requirement for the learning model based on streamlines is the sensitivity to the sequence order of points. Both Verifyber and DGCNN capture the notion of context by taking into account the neighbors of a point. Nevertheless, the working assumption is that Verifyber is exploiting more carefully the sequential relation of points in a fiber. For this purpose, we design a simple permutation test where the order of points in a fiber is randomly permuted. The side effect is to generate pathways potentially anatomically non-plausible. We then operate the inference on this new test set using just one split of 5-fold cross-validation since this experiment is not sensitive to the selection of the individuals. The results evaluated with the previous metrics are shown in Table 4. As expected, the performance of Verifyber drops to 30.0% of accuracy and 0.0% of recall because permuted fibers are classified as anatomically non-plausible. On the other hand, both DGCNN and PN preserve the previous scores, 94.3% and 94.5% respectively, because these models are invariant to the order of points.

## 5.2. Task Related Experiments

*Misclassification analysis.* The geometrical properties of fibers might be captured considering two features: the

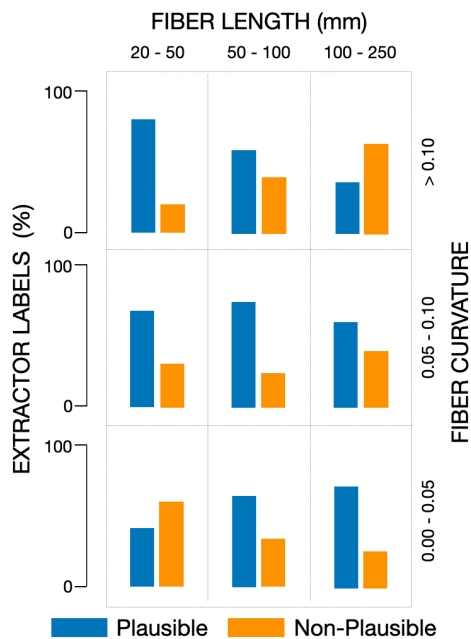


Figure 4: Distribution of Extractor labels per streamline category.

length and the curvature. The combination of these features may represent a good proxy of the anatomical properties. To analyze the misclassification error with respect to these features, We define a partition of fibers, according to their length, into three intervals: short [0, 50] mm, medium [50,100] mm, long [100,300] mm. Similarly, we operate a partition over the mean curvature: straight [0.0,0.05], curved [0.05, 0.10], very curved [0.10, 0.20]. The partitions are designed to have at least 15% of fibers in each interval. Combining the intervals of length and curvature,

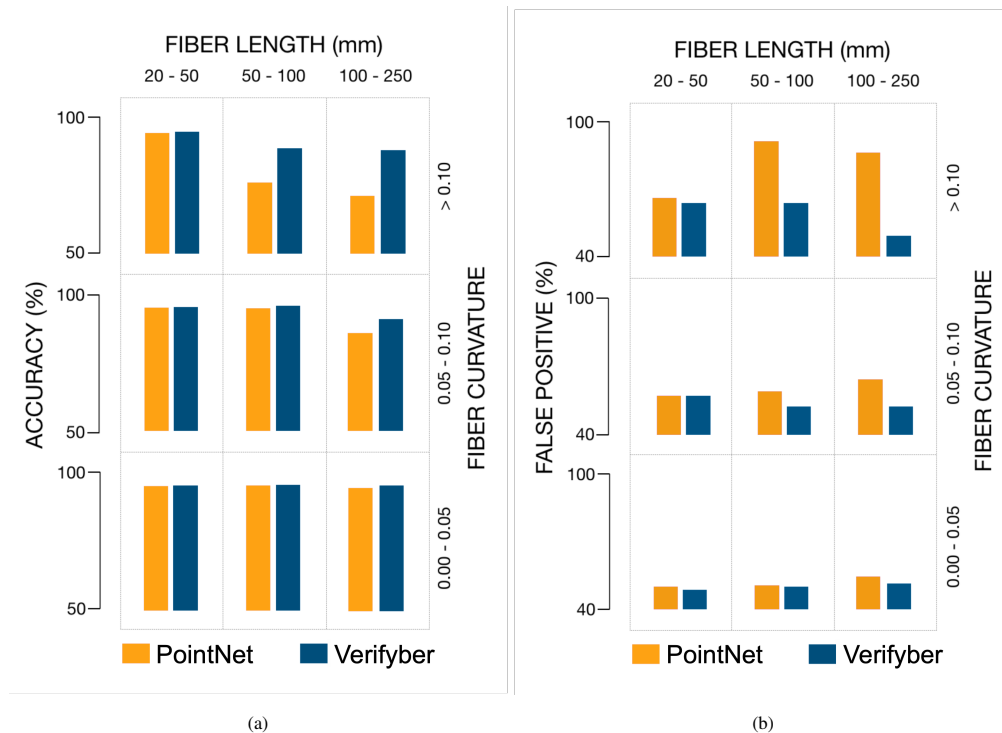


Figure 5: Comparison between PN and VF with respect to 9 streamlines categories that differ for curvature and length. We report the comparison in terms of accuracy (a) and false positive percentage (b).

we obtain 9 groups for fibers in the HCP-EP dataset, as reported in Figure 4.

We are interested to investigate how the misclassification error differs between Verifyber and PointNet, more precisely the difference in considering the point cloud only with respect to the edges. We proceed by looking at these groups and inspecting where the predictions fail to discriminate between anatomically plausible and non-plausible fibers. Despite a similar score of classification accuracy, the two methods share only 60% of the error while the remaining 40% concerns different fibers. The distribution of the error with respect to these 9 groups is reported in Figure 5a.

Figure 5a highlights that on longer and more curved fibers, Verifyber outperforms PointNet. An interpretation of the source of such a difference can be obtained by looking at the internal representation of the two models. The relevance weights associated with each fiber point are uniformly distributed in PointNet, whereas in Verifyber the

learning process clearly identifies a few more discriminating points. We show visual evidence of this difference for long and more curved fibers in Figure 6.

*In-depth error characterization.* We deepen our analysis by focusing our attention on the false positive rate, i.e., the tendency of misclassifying non-plausible fiber as plausible. We neglect the false negative rate in this analysis because Extractor (Petit et al., 2022) adopts an exclusive policy to label the fibers as anatomically non-plausible, i.e., is more sensitive to false positive error. False negative error is qualitatively investigated later. In Figure 5b, we show how false positive error differs between Verifyber and PointNet. Even in this case, Verifyber behaves better than PointNet, meaning lower false positive rate when fibers are long and curved. PointNet has a clear bias to classify those fibers as anatomically plausible, while Verifyber is more robust and keeps the false positive rate consistently in the range of 50–60% across all the groups of fibers. More in detail, the

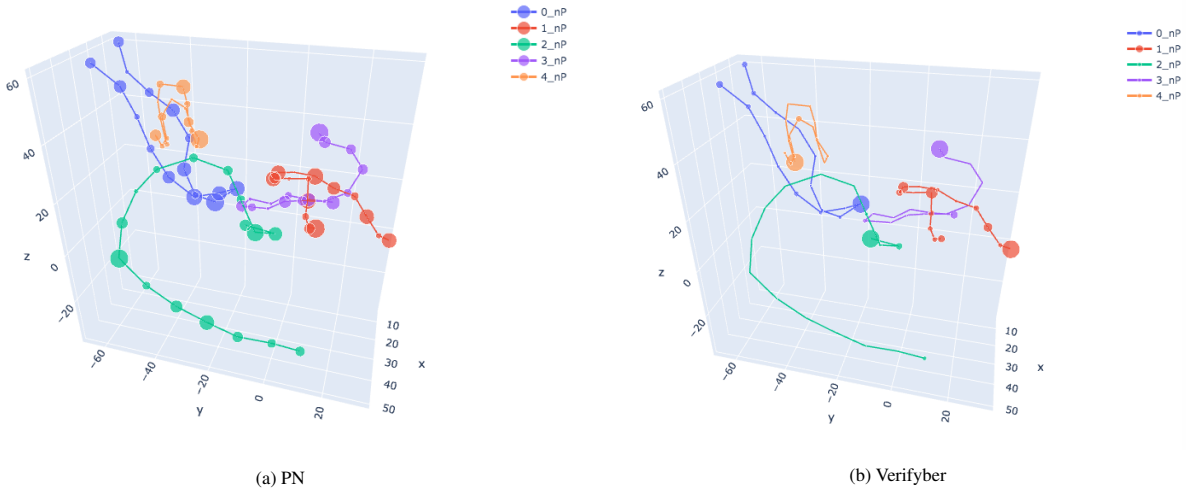


Figure 6: Given a set of streamlines, we compare the per-point contribution to the classification performed by PN and VF. The reported streamlines are all non-plausible and belong to the category of long and curved where the misclassification error is greater. We analyzed the amount of contribution of each point to the global max pooling present in both PN and VF. The size of points indicate their importance with respect to the single fiber descriptor generated by the pooling. We observe how PN tends to maintain a uniformly distributed importance, while VF seems able to individuate few strategic points for the filtering task.

worst performance of PointNet are for medium length and very curved fibers (91.4%), long and very curved (86.5%), long and curved (64.2%), where the rates of Verifyber are 64.1%, 49.5% and 51.1% respectively.

Considering the conservative approach adopted by exclusive labelings, it might be interesting to inspect the false negative error qualitatively. In this case, the goal is to evaluate whether anatomically plausible fibers misclassified as non-plausible can be considered controversial due to the noisy process of ground truth definition. For this purpose we operate a visual inspection on a random sample of misclassified fibers, as reported in Figure 7. Although those fibers are labeled as anatomically plausible, — probably because considered unknown by the exclusive policy — a manual survey by an expert anatomist confirms the classification of Verifyber as anatomically non-plausible.

*Incremental learning.* The qualitative analysis of false negative fibers highlights the issue of ground truth. Inconsistencies or mistakes in the definition of anatomical plausibility are not only related to the manual labeling process. The debate on human brain anatomy is an ongoing challenge, and the knowledge of white matter pathways

is constantly evolving. In the machine learning literature, this circumstance is known as concept drift. For this reason, we need to investigate how Verifyber might be robust when the ground truth is incrementally updated.

We design a simulation where the labeling of fibers is revised at different stages by adding new knowledge following an inclusive policy. For this analysis, we refer to the HCP-IZ dataset and the categories of bundles defined in the related atlas: association, projection, commissural, cerebellar, and superficial. In the first stage, only fibers of association bundles are labeled as anatomically plausible, non-plausible otherwise. In the second, third, and fourth stages, we add the fibers of projection, commissural and cerebellar bundles, respectively. Finally, in the fifth stage, we consider the fibers of all bundles defined in the atlas, i.e., deep and superficial bundles. Even though the HCP-IZ dataset is the result of processing hundred of individuals, it is composed of only a single average tractogram. For this reason, we organize the training set by randomly picking 80% of fibers and the test set with the remaining 20%. For all stages, we carry out a training process using the same hyperparameters described above.

Table 5 shows the results of the incremental learning

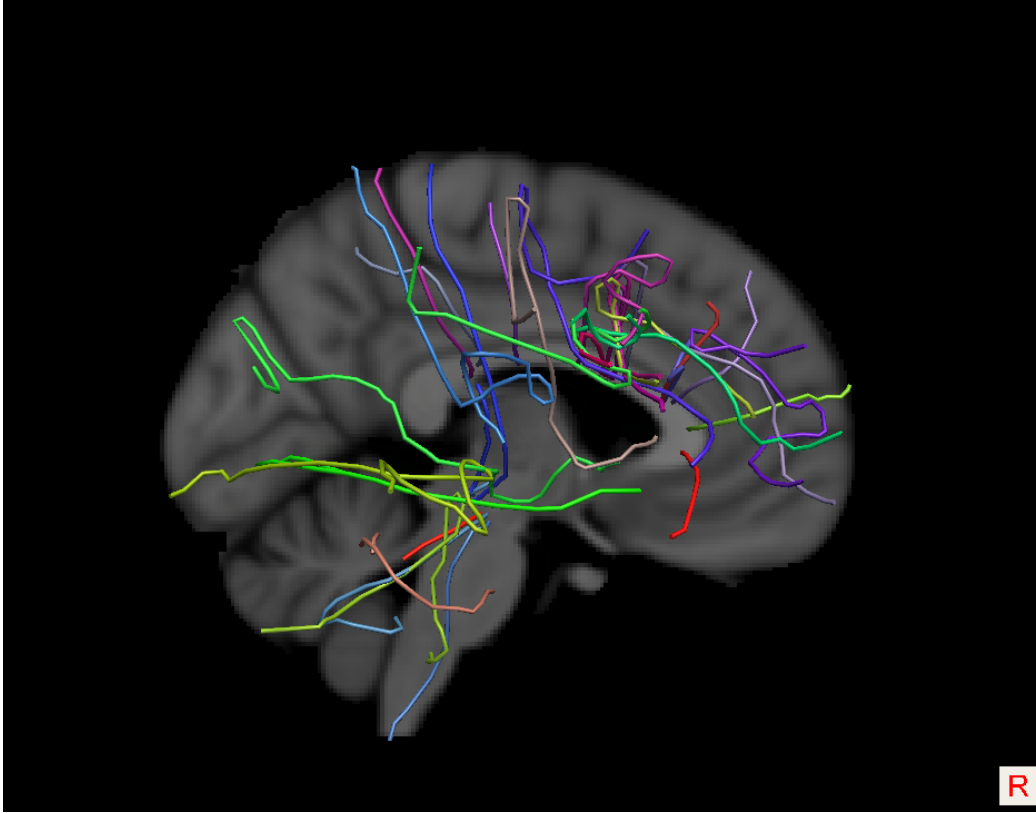


Figure 7: Example of FN i.e., labeled as plausible but classified as non-plausible, produced by VF. The streamlines shown are clearly non-plausible showing the presence of noise in the rule-based labeling, and the capacity of VF of extending the concept learnt by the rules.

Table 5: VF results on incremental learning setting in HCP-IZ dataset. The model is trained always with the same configuration and hyperparameters. The labeling changes incrementally: first row considers the streamlines of Association (A) bundles as plausible and the rest as non-plausible; in the second row the labeling of plausible is incremented considering also the streamlines of Projection (P) bundles; similarly the third and fourth rows add the streamlines of Commissural (Co) and Cerebellar (Ce) bundles. Finally, the last row adds superficial U-shape fibers (Sup) as plausible. This corresponds to considering the whole Zhang et al. (2018) atlas as plausible and the rest as non-plausible.

Method	Plausible	Accuracy	Precision	Recall	DSC
<b>Verifyber</b>	A	98.8	96.4	96.0	96.2
	A+P	98.0	96.4	95.9	96.1
	A+P+Co	97.9	97.0	96.6	96.8
	A+P+Co+Ce	97.8	97.1	96.4	96.7
	A+P+Co+Ce+Sup	97.1	97.6	98.0	97.8

for the five stages. The scores confirm that Verifyber is considerably stable and does not suffer the drift of anatomical plausibility. However, we may notice a small decrease

in accuracy, compensated by an increase of DSC, when new groups of bundles are added to the ground truth. This behavior is explainable by the balance shift between the



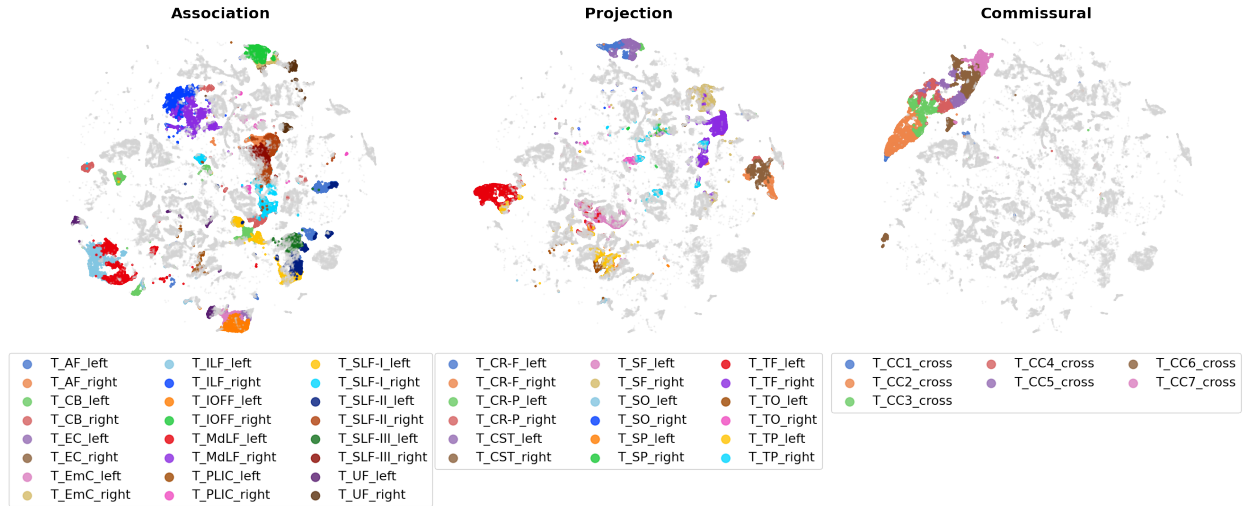


Figure 8: Latent space learned by  $VF^{IZ}$ . Each point corresponds to a streamline in the 1024D space reduced to 2D by means of t-SNE (Van der Maaten and Hinton, 2008). The three plots show the non-plausible streamlines in light-gray, together with one macro category of bundle, i.e., association, projection, commissural at a time.

number of plausible and non-plausible fibers.

*T-SNE analysis of learned features.* A deeper analysis of the results can be carried out by looking at the latent space learned by Verifyber after the training process. In the latent space, each fiber is encoded into a vector of 1024 dimensions. We may visualize this space by projecting all the fibers into a two-dimensional plot by means of t-SNE (Van der Maaten and Hinton, 2008) as reported in Figure 8. Using a color scheme, we highlight the proximity of fibers that belong to the same bundle. It is worth noting that even the proximity among the bundles is preserved, e.g., AF is close to SLF-II and SLF-III, CC[1-7] are almost consecutive, MdLF is close to ILF. Lateralized bundles are well separated from each other, e.g., IFOF left and right. There is consistency in the lateral grouping of similar bundles, i.e., if a left bundle is close to another left bundle, the corresponding right bundles are close too.

*Comparison with an unsupervised method.* An interesting open question is investigating how our supervised model behaves compared to a method such as FINTA (Legarreta et al., 2021); a state-of-the-art approach for unsupervised tractogram filtering based on deep learning. Specifically, FINTA employs a convolutional autoencoder, which, simi-

larly to our case, is trained directly on the raw streamline structure. We believe that such a comparison might be relevant to clarify the difference between unsupervised and supervised approaches. Unfortunately, neither the code nor the data used in Legarreta et al. (2021) has been publicly distributed. For this reason, we re-implemented FINTA, following the methodological description provided by the Authors in their article. We publish our FINTA implementation, see Section 5.3.

The tractogram filtering in FINTA is designed as a two steps process: (i) computation of a latent space for streamlines representation; (ii) filtering with a lazy classifier based on a thresholded nearest neighbor rule. Analogously to what is carried out in (Legarreta et al., 2021), we train the autoencoder of FINTA with a single average tractogram, the one of HCP-IZ, whose streamlines are randomly split into 80/20 partition for train and test respectively. The training is performed using the hyperparameters reported by the authors where possible; default values are used otherwise. We trained the autoencoder until convergence (see Appendix). In addition, we performed a qualitative assessment of streamlines that were reconstructed with the trained autoencoder to double-check that the reproduced approach worked as expected (see Appendix). After the computation of the latent space on HCP-

IW we proceed with the inference both on HCP-IZ and HCP-EP. For each of the two datasets we tuned the choice of the threshold value for the nearest neighbor rule. The tuning is carried out on a random subsample of streamlines from the validation set.

In Table 6, we report the values of accuracy, precision, recall, and DSC for the HCP-IZ and the HCP-EP dataset. We observe that on HCP-IZ FINTA achieves the scores in line with the results published in Legarreta et al. (2021). Nevertheless, the performance of Verifyber is  $\sim 6.5\%$  higher in terms of DSC and  $\sim 9.1\%$  in accuracy. However, the gap becomes even more consistent if we look at the result on HCP-EP, where FINTA has a significant drop in precision.

Table 6: FINTA performance evaluation. We report the test score of accuracy, precision, recall, and DSC on HCP-IZ all ( $p = A+P+Co+Ce+Sup$ ), and HCP-EP. The training of the FINTA autoencoder is performed using the dataset HCP-IZ. Then, we select a portion of plausible ( $p$ ) fibers, which are used as anchors for the embedded nearest neighbor step of FINTA. The selection is performed from the training set of the dataset at hand.

Method	$p/np$ label	Acc	Prec	Rec	DSC
HCP-IZ all					
FINTA	Zhang et al. (2018)	88.0	87.3	95.8	91.3
<b>Verifyber</b>		<b>97.1</b>	<b>97.6</b>	<b>98.0</b>	<b>97.8</b>
HCP-EP					
FINTA	Petit et al. (2022)	74.3	75.1	94.6	83.8
<b>Verifyber</b>		<b>95.2</b>	<b>96.2</b>	<b>96.9</b>	<b>96.6</b>

*Generalization across data sources.* In the previous experiments, we trained Verifyber on HCP-EP and HCP-IZ datasets, in which tractography and ground truth policy differ. An open question is to assess how these models behave on unseen tractograms. For this purpose, we design an experiment to perform inference on a new dataset, namely HCP-IW, using the models trained on HCP-EP and HCP-IZ. Since a ground truth is unavailable on HCP-IW, we need to revise the evaluation procedure. The segmented bundles in HCP-IW might be considered fiducial regions, where to focus for quantitative and qualitative analysis. We limit our analysis to 40 most common bundles, those shared with HCP-IZ, out of the 72 available bundles. The list of chosen bundles is shown in Figure 10.

We carry out the inference on the fibers of these bundles, then perform a quantitative and qualitative analysis of potential false negative error since the expected prediction should be only anatomically plausible by design. As a quantitative measure, we compute the volumetric DSC score between the mask of the original bundle and the mask of fibers classified as anatomically plausible. The working assumption is that a moderate false negative error would not affect the estimate of the volumetric region of a bundle. We deepen our evaluation with a qualitative analysis by visually inspecting a sample of fibers classified as anatomically non-plausible, i.e., drawn from the portion of potential false negative fibers. We replicate this procedure both on HCP-IZ and the 23 individuals distributed as test set from HCP-IW datasets. In addition, we investigate the agreement of the two models trained on two different ground truths, i.e.,  $VF^{EP}$  and  $VF^{IZ}$ .

In Figures 9 and 10, we report the volumetric DSC score obtained with Verifyber trained on HCP-EP and HCP-IZ respectively. In both the plots, the average value of DSC is above 0.95 while the minimum does not fall below 0.80. The qualitative analysis of the filtering is depicted in Figures 11 and 12. A small sample of fibers misclassified as false negative is selected from a few common bundles. Even though these fibers belong to segmented bundles, the visual inspection by an expert confirms that their pathways are anatomically non-plausible. It is worth noting that even if a fiber falls in the volumetric mask of a bundle not necessarily the related pathway is anatomically plausible. The color scheme highlights those fibers that are classified as anatomically non-plausible from both models.

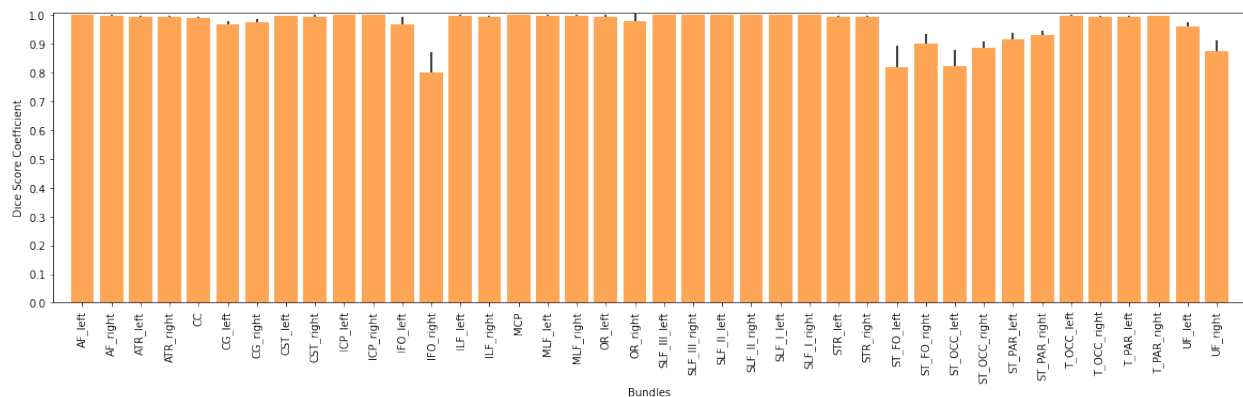


Figure 9: Volumetric DSC of Tractseg bundles after filtering using  $VF^{EP}$ . Reported mean and standard deviation refers to 23 subjects. The filtering does not impact significantly the shape of the bundle guaranteeing at least 80% of DSC. The mean DSC is  $96.8 \pm 5.4$ .

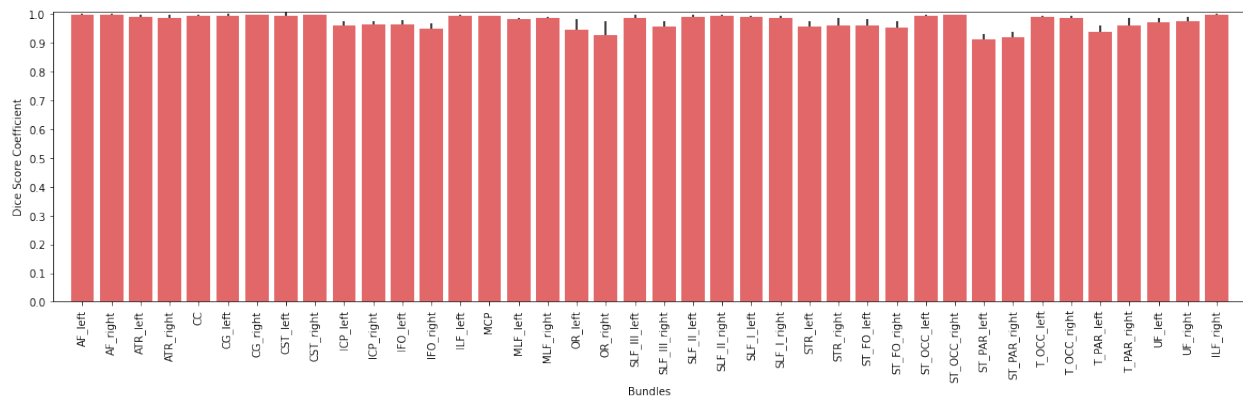


Figure 10: Volumetric DSC of Tractseg bundles after filtering using  $VF^{LZ}$ . Reported mean and standard deviation refers to 23 subjects. The filtering does not impact significantly the shape of the bundle guaranteeing at least 85% of DSC. In most of the cases the DSC is above 95%. The mean DSC is  $97.5 \pm 2.4$ .

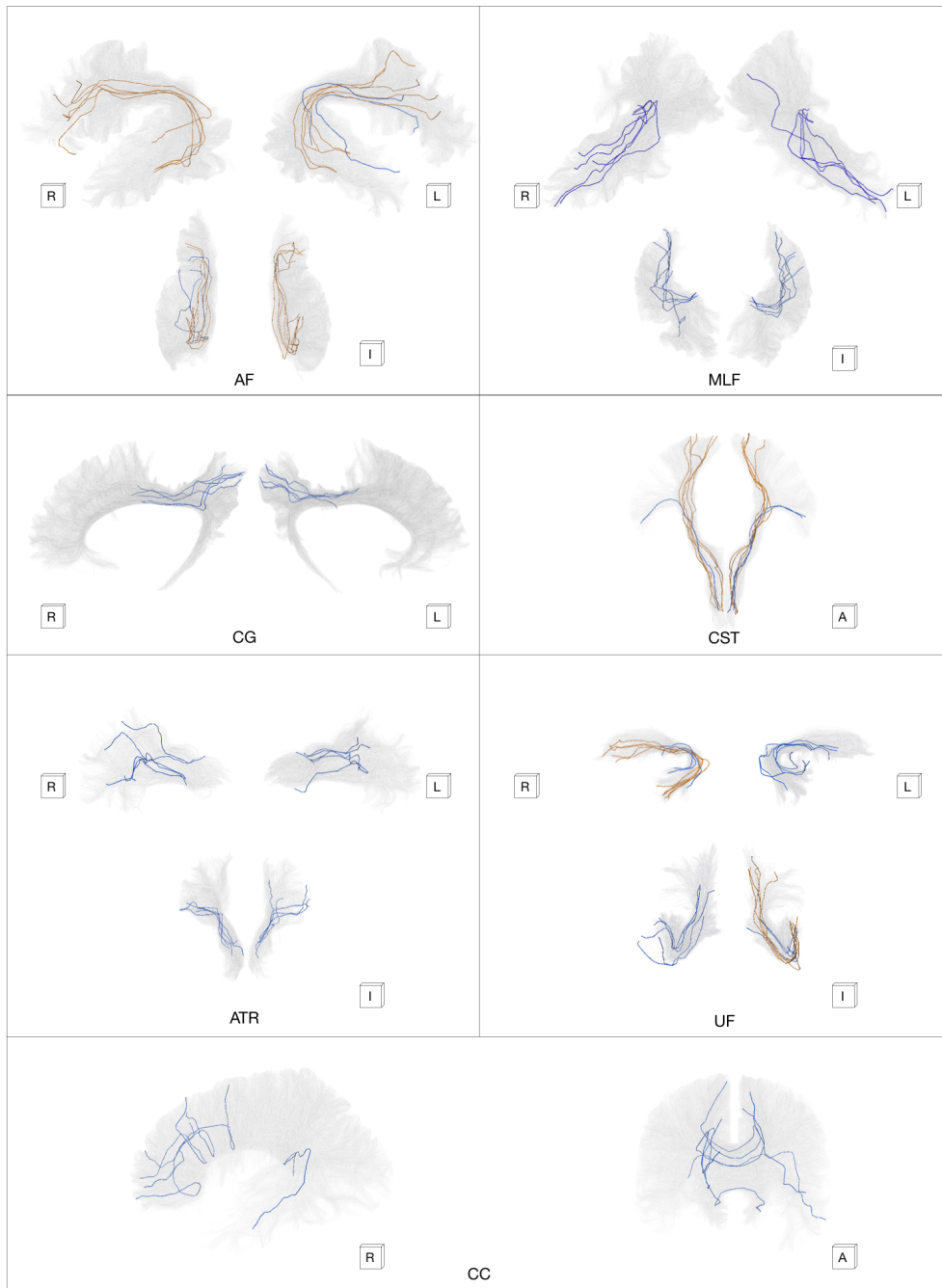


Figure 11: Qualitative example of non-plausible streamlines belonging to Tractseg bundles individuated using VF<sup>EP</sup>. The figure shows in blue streamlines individuated as non-plausible both by VF<sup>EP</sup> and VF<sup>IZ</sup> (shared non-plausible). The orange streamlines are instead exclusive of VF<sup>EP</sup>. In some bundles there is only a very small number of shared non-plausible, because the amount of non-plausible found by both model on that bundle is globally very low e.g., the AF bundles has 0.001% of non-plausible fibers found globally both when we use VF<sup>EP</sup> and when we use VF<sup>IZ</sup>.

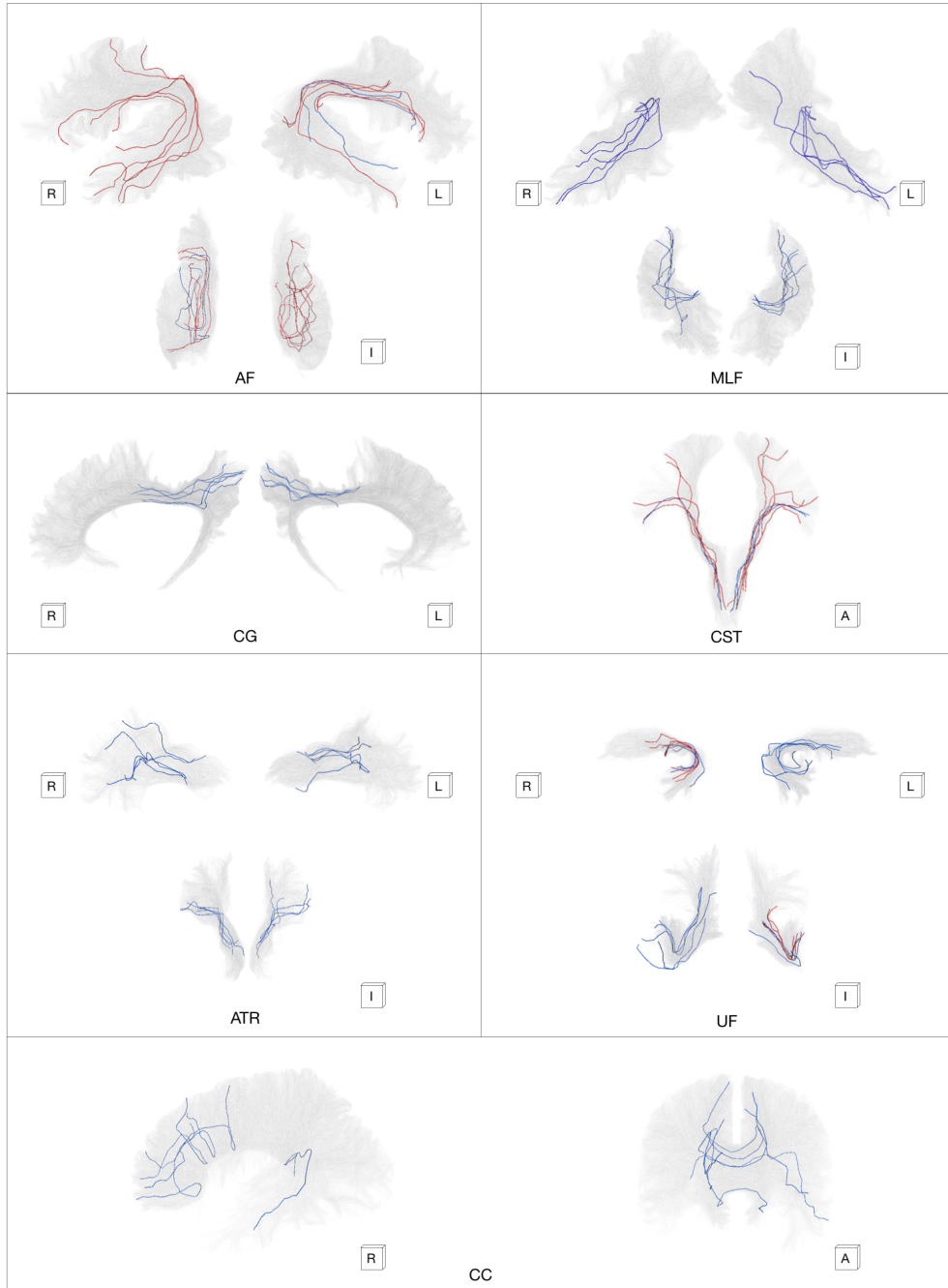


Figure 12: Qualitative example of non-plausible streamlines belonging to Tractseg bundles individuated using  $VF^{IZ}$ . Blue streamlines are shared non-plausible examples, while red streamlines are exclusive non-plausible of  $VF^{IZ}$ .

*Model deployment on clinical data.* As an additional evaluation, we investigate how the VF learnt model behaves when inference is carried out on clinical data. This experiment considers the patients with tumors of the APSS-IS dataset, and we filter their tractograms with  $\text{VF}^{\text{EP}}$ . Note that the choice of filtering with  $\text{VF}^{\text{EP}}$  rather than  $\text{VF}^{\text{IZ}}$  is driven by the more conservative approach of the underneath labeling. Similar to the experiments involving the HCP-IW dataset, we refer to the segmented bundles as a proxy evaluator of plausible streamlines. However, differently from HCP-IW, the quality of bundles in clinical data is lower since the time restrictions on DWI sequence acquisition. The shape of the segmented bundles is more sensitive to the removal of hypothetical non-plausible streamline, therefore we cannot carry out the quantitative analysis previously adopted for HCP-IW. Hence, we proceed with a qualitative analysis by visual inspection of three different bundles from the healthy hemisphere and one bundle from the tumored hemisphere for each of the 5 subjects.

To illustrate the results of the qualitative analysis we re-

port two figures. Figure 13 depicts the filtering of bundles segmented from the healthy hemisphere of the subjects, while Figure 14 shows one bundle segmented from the tumored hemisphere. In both cases, the individuated non-plausible streamlines, colored in black, appear to be clearly artificial and poorly compliant with the expected shape of the bundle. In many cases, such streamlines either do a strict U-turn like in the AF of subject S2 or are truncated like in the PT of S3.

### 5.3. Code and reproducibility

Verifyber has been implemented using PyTorch (Paszke et al., 2019) and the extension for geometric deep learning, PyG (Fey and Lenssen, 2019). Our implementation, along with the trained models used in the experiments, is available at <https://github.com/FBK-NILab/verifyber>. The Github repository also contains our FINTA (Legarreta et al., 2021) implementation. In addition, to simplify the deployment of  $\text{VF}^{\text{EP}}$  and  $\text{VF}^{\text{IZ}}$ , we published an App on the BrainLife platform (Avesani et al., 2019), accessible at <https://doi.org/10.25663/brainlife.app.390>.

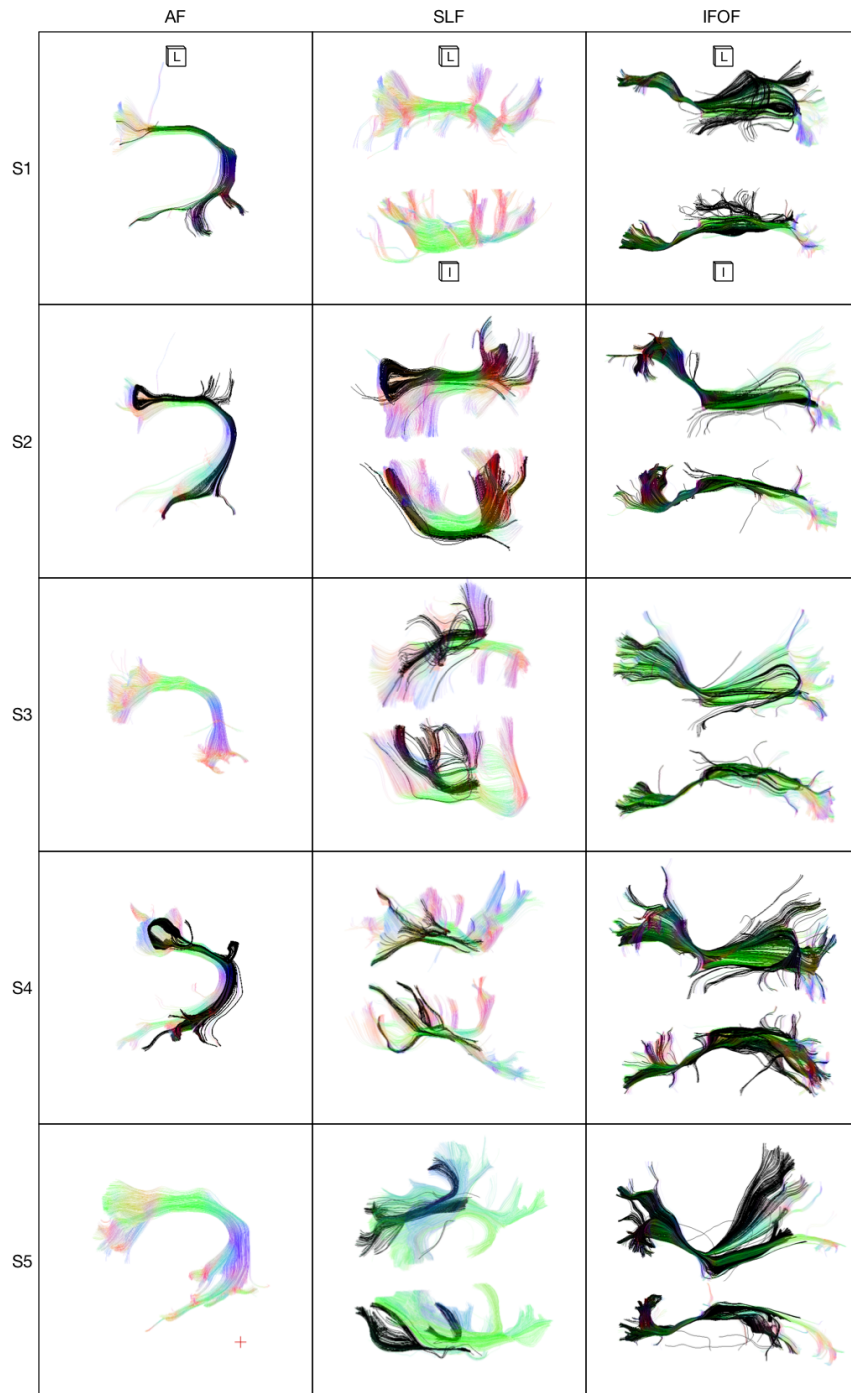


Figure 13: Qualitative results after filtering APSS clinical dataset using  $VF^{EP}$ . Black streamlines are classified as non-plausible. Bundles reported in this figure belong to the healthy hemisphere of patients.

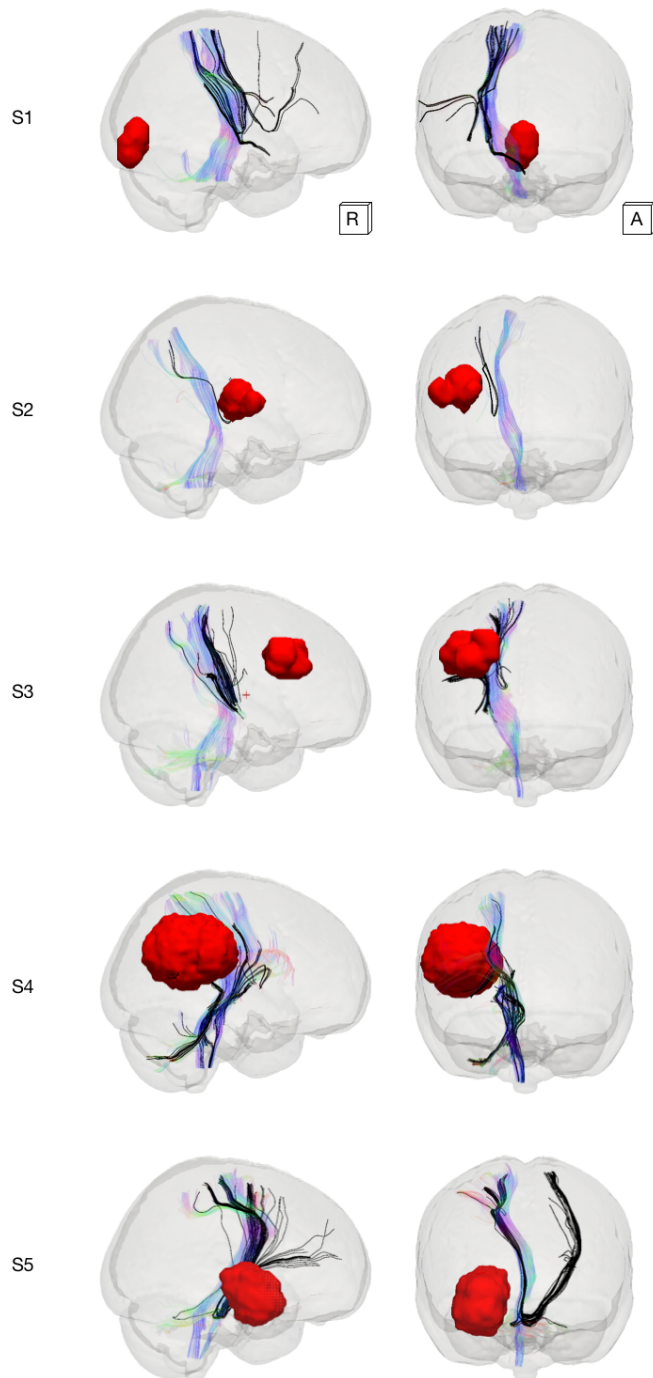


Figure 14: Qualitative results of the tumored hemisphere after filtering APSS clinical dataset using  $VF^{EP}$ . The bundle shown is the pyramidal tract (PT). Black streamlines are classified as non-plausible. The red ROI is the segmented tumor.



## 6. Discussion and Conclusions

### 6.1. Verifyber Performance

*Verifyber is effective for tractogram filtering.* The proposed method, Verifyber, resulted successful for the tractogram filtering task. It can learn from an external supervision, e.g. Extractor labeling (see Figure 3), and then preserving the high performance reached during training also for inference on unseen data, see Table 2. In all the reported metrics the high scores are coupled with a low standard deviation in the cross-validation setting, e.g., accuracy  $95.2\% \pm 0.01$  and DSC  $96.6\% \pm 0.01$ , an empirical evidence of the robustness with respect to the training and testing partition. The higher recall compared to the precision measure suggests that the model operate a more conservative filtering, in agreement with the bias of the *exclusive* labelling policy. Additionally, VF revealed fast inference at test time, classifying 1M streamlines in less than a minute (46.2 sec using a GPU NVIDIA Titan XP 12GB), a two order of magnitude less than more common filtering approaches, as signal-based and rule-based methods.

*Sequence edge convolution benefits.* Performing tractogram filtering using a state of the art GDL model, e.g., PN or DGCNN, allows the use of raw streamline representation, i.e., varying number of points without forcing an arbitrary orientation, at the cost to be invariant to the permutation of streamline points. In VF we overcome such a drawback as confirmed by results of ablation study reported in Table 4 while remaining invariant to the fiber orientation. Furthermore, we show that considering the edges of the streamlines matters and enables VF to achieve a competitive advantage over PN when streamlines are longer and curved. In these groups of streamlines, VF has only a small drop in accuracy while significantly outperforming PN: 89% vs. 76% when long and very curved, and 88% vs. 71% when long and middle curved (see Figure 5). This gap can be explained by the different property of representation learning in the two models. Sequence edge convolutional layers of VF is more effective to capture the long range spatial information into the learned embedding of fibers. VF successfully encodes the most salient point in complex pathways while PN struggles to identify the salient points because the learned global embedding is not informative for local patterns (see Figure 6).

*Error analysis.* In *exclusive* labeling policies, the false positive error (FP) is more relevant than the false negative. Observing the FP analysis in Figure 5, VF disclosed robustness despite the uneven distribution of the plausible and non-plausible labels, as observable in Figure 4. Compared to PN, which seems to be influenced by the higher number of plausible streamlines, the FP rate of VF is always lower or equal. Again, the difference is greater for curve and long streamlines where considering the edges helps the classification. In the false negative analysis, the presence of artifactual fibers (see Figure 7) labelled as plausible proves to some extent the ability of the model to generalize the rules beyond the labeling. Although some noise in the annotation process that brings to overestimate the plausibility of some streamlines, the large number of streamlines given in input to the training (around 10M) provides a good generalization ability of VF. We may remark the importance of this property as manual labeling is intrinsically error-prone and always leads to having some noisy labels.

*Incremental learning.* The lack of a complete WM knowledge and the constant effort of neuroanatomy community lead to a continuous evolution of the notion of anatomical plausibility, which reflects on tractogram labelings, a phenomenon known as *concept drift*. New bundles are added to the definition of plausibility as soon as the neuroanatomist consolidates their definitions. We emulated such a real-world scenario in the incremental learning experiments, where VF obtained convincing results, (see Table 5). The iterative addition of new categories of bundles, e.g., adding projection bundles to association bundles, has not affected the performance of VF. VF learns a meaningful latent space where fibers of a bundle are grouped, and similar bundles are close (see Figure 8). This result suggests that VF effectively captures the trajectory of fibers. Therefore, we may expect better robustness when an *inclusive* policy of labelling is incrementally adopted in a real world setting. Moreover, in the same experiment, we unveiled the possibility to learn the filtering by training VF with a single averaged tractogram. Single-subject training works because our method scales with the number of streamlines and not with the number of subjects. Note that a single-subject training might lead to low inter-individual

generalization, but this is unlikely to be the case in HCP-IZ as the tractogram already contains information of 100 subjects.

*Supervised vs. unsupervised.* One of our working hypotheses is the use of a supervised deep learning approach, in contrast with unsupervised, to guarantee higher flexibility to different definitions of anatomical plausibility. The comparison with FINTA, chosen as representative of unsupervised methods, in Table 6 highlights a significant difference between the two types of approaches. The performance of Verifyber is high and stable on both the *inclusive* and *exclusive* labeling policies, i.e., only 2%p of accuracy and 1%p of DSC difference. On the contrary, FINTA is not robust to the two labeling policies, achieving 14%p of accuracy and 7%p of DSC difference. We remark that such a difference occurs although we tuned the filtering threshold of FINTA specifically for each dataset. In our experience, adopting the same threshold across different datasets highly worsen the results.

An additional observation concerning FINTA is that it performs considerably better on the inclusive labeling such as HCP-IZ dataset. This labeling policy discriminates between plausible and non-plausible using the concept of bundle or cluster. Non-plausible fibers are either unknown bundle or outliers manually removed from clusters. Both the situations are favorable for solutions such as FINTA based on autoencoders, as they are known to perform well in the task of anomaly/outlier detection (Pang et al., 2021). The low dimensional bottleneck of the autoencoder combined with the reconstruction loss acts as a regularizer of the latent space and promotes a *denoising* of the input fiber trajectory (Vincent et al., 2010). As a result, in FINTA, the loss based on fiber reconstruction is effective for detecting non-plausible fibers, considered outliers. However, the denoising action, especially in cases of local geometrical distortion of the pathway (see Appendix), lead to producing false positives during filtering, as highlighted by lower precision *w.r.t.* recall in Table 6.

When we move from inclusive to exclusive labelings, the hypothesis of non-plausible as outliers bundles no more holds. In exclusive labelings, the anatomical plausibility concerns the artifactual geometry of fibers and the anatomical regions in which fibers pass through or terminate. However, such characteristics might be altered by the denoising effect of AEs (see Appendix), e.g., a non-plausible sharp

turn within a fiber is deleted by denoising. The excessive denoising may explain the drop of performance of FINTA on HCP-EP, 14%p lower accuracy. In general, we may state that unsupervised filtering approaches are accurate when the premises of their loss holds in the labeling at hand. Instead, supervised approaches like Verifyber are more flexible to different labeling policies as the classification loss directly optimizes the labeling at hand. For this reason, VF substantially outperforms FINTA, +20%p accuracy and +13%p DSC on the exclusive labeling.

## 6.2. Model Deployment

*Generalization across data sources.* The reproducibility analysis aims to assess the behavior of a model on different datasets with respect to those used for training. For this purpose, we investigated the inference of VF on HCP-IW dataset, a large collection of high quality data. We considered both models trained with the two labelling policies,  $VF^{EP}$  and  $VF^{IZ}$ , *exclusive* and *inclusive* respectively. It is worthwhile to remark that for HCP-IW we do not have a labeling for anatomical plausibility but only the segmentation of major bundles, which we use as a proxy for plausible fibers. From the results reported in Figure 9 and 10, we notice that in both cases there is no meaningful alteration of volumetric bundle masks before and after the filtering despite the removal of around 20% of the streamlines. In some specific bundles such as the IFOF, our method detected a higher percentage of non-plausible streamlines, around 50 – 60%. However, given the absence of a ground truth labeling to refer to, such high percentages of non-plausible fibers could be an error of our model. For this reason, we needed to operate an additional visual check to assess the source of the error, as illustrated in Figure 15. The assessment revealed a premature termination of the fiber in the frontal lobe. This anomaly is probably due to the lack of anatomical constraints (ACT) in the tracking procedure that Wasserthal et al. (2018) adopted for the tracking of IFOF. For all the other bundles, an expert visually investigated a portion of the predicted non-plausible fibers confirming that such fibers were having a non-plausible pathway, see Figure 11 and 12. However, the partial visual check is not enough to guarantee a 100% correct filtering — which we do not expect —, but the preserved bundles masks reveal a conservative filtering of VF. Testing the learnt models,  $VF^{EP}$  and  $VF^{IZ}$ , on the same third party test dataset allows us

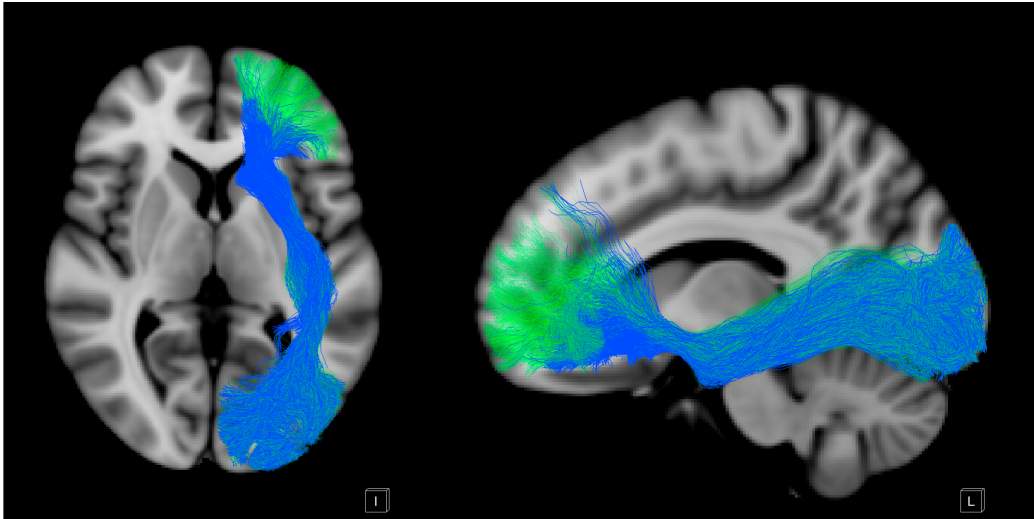


Figure 15: Filtering of Tractseg IFOF using  $\text{VF}^{\text{IZ}}$ . Blue streamlines are classified as non-plausible, while transparent green are classified as plausible.

to estimate the labeling policies overlaps computing the agreement of the predictions. The result is 84% of DSC of plausible streamlines and 37% of DSC of non-plausible streamlines. The low overlap of non-plausible is explainable by the different approaches of Zhang et al. (2018) and Extractor, i.e., inclusive vs. exclusive.

*Clinical application.* The last validation of VF on clinical data, i.e., APSS dataset, shows that the method might be effective even in the context of clinical data. Although the completely different data quality, tracking pipeline, and the presence of alteration caused by tumors, our method filters streamlines that after visual inspection confirmed to be truly non-plausible, see 13 and 14. The characteristics of the detected non-plausible streamlines agree with the principles of some of the Extractor rules, e.g., truncation, loop and strict U-turn. In most cases, it is straightforward to notice the disagreement of filtered streamlines compared to the retained streamlines of the respective bundles. One could argue that a similar strong disagreement may happen when plausible streamlines belonging to other bundles are wrongly present in the APSS dataset segmentations. To answer, we can assert that the inspection shows, on the contrary, that such streamlines are truly non-plausible. Given the clinical circumstances, we may claim that filtering such non-plausible streamlines could simplify and

speed up clinicians' manual process of bundle extraction. With fewer false positives, they do not have to draw several ROIs of exclusion.

### 6.3. Possible weaknesses and limitation

*Biases of labelling policies.* Some cautions must be taken into account when deploying Verifyber to new data, particularly considering the choice of  $\text{VF}^{\text{EP}}$  versus  $\text{VF}^{\text{IZ}}$ . In the case of  $\text{VF}^{\text{EP}}$ , the filtering is highly dependent on a proper non-linear co-registration of the target brain to standard MNI as many rules of Extractor concern the anatomy of that standard. The same strict requirement is not present for  $\text{VF}^{\text{IZ}}$ , as the HCP-IZ training dataset contains fibers that are only linearly registered to a shared anatomical reference. However,  $\text{VF}^{\text{IZ}}$  has other limitations. First, it is biased towards the bundle definitions followed by Zhang et al.. Secondly, the averaged tractogram of the atlas does not contain spurious streamlines that terminate in the WM or contain loops, and thus one cannot expect the removal of these in a new tractogram.

*Verifyber is not designed for signal-based filtering.* In case one is interested in re-training VF using new labelings, it is important to remind that VF is not meant to learn signal-based filtering. The reason is that signal-based criteria are based on spatial regularization of fibers, i.e., fiber

density map, and not on fiber anatomy. For example, considering SIFT2 (Smith et al., 2015b) weights, the anatomy of high-weight fibers is often non-plausible. For the sake of completeness, in additional experiments, we explored using VF to learn the SIFT2 filtering (see Appendix).

*Quantitative generalization experiments are missing.* Despite the extensive set of experiments carried out to test Verifyber, we miss a second labeled dataset where to quantitatively confirm the generalization performance of VF. The availability of labeled tractograms is still limited, and in our case, we had available only a single additional non-HCP tractogram labeled with Extractor. Despite that tractogram contains more than 1.5M fibers, we considered a single-subject test not statistically significant. We attach the result on that dataset in Appendix.

## CRediT

**Pietro Astolfi:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration. **Ruben Verhagen:** Software, Formal analysis, Investigation, Writing - Review & Editing. **Laurent Petit:** Resources, Validation, Data Curation, Writing - Review & Editing. **Emanuele Olivetti:** Software, Writing - Review & Editing. **Silvio Sarubbo:** Resources, Data Curation, Writing - Review & Editing. **Jonathan Masci:** Methodology, Writing - Review & Editing. **Davide Boscaini:** Conceptualization, Methodology, Writing - Review & Editing, Visualization. **Paolo Avesani:** Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing, Visualization, Supervision, Project administration, Funding acquisition.

## Acknowledgements

This work was partially supported by the grant PAT Reg. n. 764/2021 NeuSurPlan. Finally, we gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

## References

Astolfi, P., Verhagen, R., Petit, L., Olivetti, E., Masci, J., Boscaini, D., Avesani, P., 2020. Tractogram Filtering of Anatomically Non-plausible Fibers with Geometric Deep Learning, in: Medical Image

Computing and Computer Assisted Intervention – MICCAI 2020, pp. 291–301. doi:10.1007/978-3-030-59728-3\_29.

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. *Scientific Data* 6, 1–13. doi:10.1038/s41597-019-0073-y.

Aydogan, D.B., Shi, Y., 2015. Track Filtering via Iterative Correction of TDI Topology, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pp. 20–27. doi:10.1007/978-3-319-24553-9\_3.

Aydogan, D.B., Shi, Y., 2018. Tracking and validation techniques for topographically organized tractography. *NeuroImage* 181, 64–84. doi:10.1016/j.neuroimage.2018.06.071.

Basser, P.J., Pajevic, S., Pierpaoli, C., Duda, J., Aldroubi, A., 2000. In vivo fiber tractography using DT-MRI data. *Magnetic Resonance in Medicine* 44, 625–632.

Bastiani, M., Shah, N.J., Goebel, R., Roebroeck, A., 2012. Human cortical connectome reconstruction from diffusion weighted MRI: The effect of tractography algorithm. *NeuroImage* 62, 1732–1749.

Bertò, G., Bullock, D., Astolfi, P., Hayashi, S., Zigioto, L., Annicchiarico, L., Corsini, F., De Benedictis, A., Sarubbo, S., Pestilli, F., Avesani, P., Olivetti, E., 2021. Classifyber, a robust streamline-based linear classifier for white matter bundle segmentation. *NeuroImage* 224, 117402. doi:10.1016/j.neuroimage.2020.117402.

Bronstein, M.M., Bruna, J., LeCun, Y., Szlam, A., Vandergheynst, P., 2017. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine* 34, 18–42.

Buchanan, C.R., Pernet, C.R., Gorgolewski, K.J., Storkey, A.J., Bastin, M.E., 2014. Test–retest reliability of structural brain networks from diffusion MRI. *NeuroImage* 86, 231–243. doi:10.1016/j.neuroimage.2013.09.054.

Chamberland, M., Scherrer, B., Prabhu, S.P., Madsen, J., Fortin, D., Whittingstall, K., Descoteaux, M., Warfield, S.K., 2017. Active delineation of Meyer’s loop using oriented priors through MAGNETic tractography (MAGNET). *Human Brain Mapping* 38, 509–527.

Chandio, B.Q., Chattopadhyay, T., Owens-Walton, C., Reina, J.E.V., Nabulsi, L., Thomopoulos, S.I., Garyfallidis, E., Thompson, P.M., 2022. FiberNeat: Unsupervised White Matter Tract Filtering, in: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 5055–5061. doi:10.1109/EMBC48229.2022.9870877.

Côté, M.A., Garyfallidis, E., Larochelle, H., Descoteaux, M., 2015. Cleaning up the mess: Tractography outlier removal using hierarchical QuickBundles clustering. *Proceedings of International Society of Magnetic Resonance in Medicine (ISMRM)*.

- Côté, M.A., Girard, G., Boré, A., Garyfallidis, E., Houde, J.C., Descoteaux, M., 2013. Tractometer: Towards validation of tractography pipelines. *Medical Image Analysis* 17, 844–857. doi:10.1016/j.media.2013.03.009.
- Daducci, A., Dal Palu, A., Lemkaddem, A., Thiran, J.P., 2015. COM-MIT: Convex Optimization Modeling for Microstructure Informed Tractography. *IEEE Transactions on Medical Imaging* 34, 246–257. doi:10.1109/TMI.2014.2352414.
- De Benedictis, A., Petit, L., Descoteaux, M., Marras, C.E., Barbareschi, M., Corsini, F., Dallabona, M., Chioffi, F., Sarubbo, S., 2016. New insights in the homotopic and heterotopic connectivity of the frontal portion of the human corpus callosum revealed by microdissection and diffusion tractography. *Human Brain Mapping* 37, 4718–4735. doi:10.1002/hbm.23339.
- Descoteaux, M., 2015. High Angular Resolution Diffusion Imaging (HARDI), in: *Wiley Encyclopedia of Electrical and Electronics Engineering*, pp. 1–25. doi:10.1002/047134608X.W8258.
- Descoteaux, M., Angelino, E., Fitzgibbons, S., Deriche, R., 2007. Regularized, fast, and robust analytical Q-ball imaging. *Magnetic Resonance in Medicine* 58, 497–510. doi:10.1002/mrm.21277.
- Essayed, W.I., Zhang, F., Unadkat, P., Cosgrove, G.R., Golby, A.J., O'Donnell, L.J., 2017. White matter tractography for neurosurgical planning: A topography-based review of the current state of the art. *NeuroImage: Clinical* 15, 659–672. doi:10.1016/j.nicl.2017.06.011.
- Fey, M., Lenssen, J.E., 2019. Fast Graph Representation Learning with PyTorch Geometric, in: *ICLR Workshop on Representation Learning on Graphs and Manifolds*.
- Fillard, P., Descoteaux, M., Goh, A., Gouttard, S., Jeurissen, B., Malcolm, J., Ramirez-Manzanares, A., Reisert, M., Sakaie, K., Tensaouti, F., Yo, T., Mangin, J.F., Poupon, C., 2011. Quantitative evaluation of 10 tractography algorithms on a realistic diffusion MR phantom. *NeuroImage* 56, 220–234. doi:10.1016/j.neuroimage.2011.01.032.
- Fischl, B., 2012. FreeSurfer. *NeuroImage* 62, 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, 313–327. doi:10.1016/j.neuroimage.2010.07.033.
- Frigo, M., Deslauriers-Gauthier, S., Parker, D., Aziz Ould Ismail, A., John Kim, J., Verma, R., Deriche, R., 2020. Diffusion MRI tractography filtering techniques change the topology of structural connectomes. *Journal of Neural Engineering* 17, 065002. doi:10.1088/1741-2552/abc29b.
- 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. *Frontiers in Neuroinformatics* 8, 1+.
- Garyfallidis, E., Brett, M., Correia, M.M., Williams, G.B., Nimmo-Smith, I., 2012. QuickBundles, a Method for Tractography Simplification. *Frontiers in Neuroscience* 6. doi:10.3389/fnins.2012.00175.
- Garyfallidis, E., Côté, M.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. doi:10.1016/j.neuroimage.2017.07.015.
- Girard, G., Caminiti, R., Battaglia-Mayer, A., St-Onge, E., Ambrosen, K.S., Eskildsen, S.F., Krug, K., Dyrby, T.B., Descoteaux, M., Thiran, J.P., Innocenti, G.M., 2020. On the cortical connectivity in the macaque brain: A comparison of diffusion tractography and histological tracing data. *NeuroImage* 221, 117201. doi:10.1016/j.neuroimage.2020.117201.
- Girard, G., Whittingstall, K., Deriche, R., Descoteaux, M., 2014. Towards quantitative connectivity analysis: Reducing tractography biases. *NeuroImage* 98, 266–278. doi:10.1016/j.neuroimage.2014.04.074.
- 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.
- Graves, A., Schmidhuber, J., 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18, 602–610.
- Gupta, V., Thomopoulos, S.I., Corbin, C.K., Rashid, F., Thompson, P.M., 2018. FIBERNET 2.0: An automatic neural network based tool for clustering white matter fibers in the brain, in: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 708–711. doi:10.1109/ISBI.2018.8363672.
- Gupta, V., Thomopoulos, S.I., Rashid, F.M., Thompson, P.M., 2017. FiberNET: An Ensemble Deep Learning Framework for Clustering White Matter Fibers, in: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*, pp. 548–555. doi:10.1007/978-3-319-66182-7\_63.
- Hau, J., Sarubbo, S., Houde, J., Corsini, F., Girard, G., Deledalle, C., Crivello, F., Zago, L., Mellet, E., Jobard, G., Joliot, M., Mazoyer, B., Tzourio-Mazoyer, N., Descoteaux, M., Petit, L., 2017. Revisiting the human uncinate fasciculus, its subcomponents and asymmetries with stem-based tractography and microdissection validation. *Brain Structure and Function* , 1–18doi:10.1007/s00429-016-1298-6.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778.
- Henderson, F., Abdullah, K.G., Verma, R., Brem, S., 2020. Tractography and the connectome in neurosurgical treatment of gliomas: The premise, the progress, and the potential. *Neurosurgical Focus* 48, E6. doi:10.3171/2019.11.FOCUS19785.

- Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. *Neural Computation* 9, 1735–1780.
- Huang, Z., Xu, W., Yu, K., 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991* [arXiv:1508.01991](https://arxiv.org/abs/1508.01991).
- Jenkinson, M., Beckmann, C.F., Behrens, T.E.J., Woolrich, M.W., Smith, S.M., 2012. FSL. *NeuroImage* 62, 782–790. doi:[http://dx.doi.org/10.1016/j.neuroimage.2011.09.015](https://doi.org/10.1016/j.neuroimage.2011.09.015).
- Jeurissen, B., Descoteaux, M., Mori, S., Leemans, A., 2017. Diffusion MRI fiber tractography of the brain. *NMR in Biomedicine* 32, e3785. doi:10.1002/nbm.3785.
- Jörgens, D., Descoteaux, M., Moreno, R., 2021. Challenges for Tractogram Filtering, in: *Anisotropy Across Fields and Scales*, pp. 149–168. doi:10.1007/978-3-030-56215-1\_7.
- Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems* 25, 1097–1105.
- Lee, M.H., O'Hara, N., Sonoda, M., Kuroda, N., Juhasz, C., Asano, E., Dong, M., Jeong, J.W., 2020. Novel Deep Learning Network Analysis of Electrical Stimulation Mapping-Driven Diffusion MRI Tractography to Improve Preoperative Evaluation of Pediatric Epilepsy. *IEEE Transactions on Biomedical Engineering* 67, 3151–3162. doi:10.1109/TBME.2020.2977531.
- Legarreta, J.H., Petit, L., Rheault, F., Theaud, G., Lemaire, C., Descoteaux, M., Jodoin, P.M., 2021. Filtering in Tractography using Autoencoders (FINTA). *Medical Image Analysis* , 102126doi:10.1016/j.media.2021.102126.
- Lemkaddem, A., Skiödebrand, D., Dal Palú, A., Thiran, J.P., Daducci, A., 2014. Global tractography with embedded anatomical priors for quantitative connectivity analysis. *Frontiers in Neurology* 5, 232. doi:10.3389/fneur.2014.00232.
- Maffei, C., Jovicich, J., De Benedictis, A., Corsini, F., Barbareschi, M., Chioffi, F., Sarubbo, S., 2018. Topography of the human acoustic radiation as revealed by ex vivo fibers micro-dissection and in vivo diffusion-based tractography. *Brain Structure and Function* 223, 449–459. doi:10.1007/s00429-017-1471-6.
- 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., He, R., 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., Auria, A., Esteban, O., Lemkaddem, A., Thiran, J.P., Cetingul, H.E., Odry, B.L., Mailhe, B., Nadar, M.S., Pizzagalli, F., Prasad, G., Villalon-Reina, J.E., Galvis, J., Thompson, P.M., Requejo, F.D.S., Laguna, P.L., Lacerda, L.M., Barrett, R., Dell'Acqua, F., Catani, M., Petit, L., Caruyer, E., Daducci, A., Dyrby, T.B., Holland-Letz, T., Hilgetag, C.C., Stieltjes, B., Descoteaux, M., 2017. The challenge of mapping the human connectome based on diffusion tractography. *Nature Communications* 8, 1349. doi:10.1038/s41467-017-01285-x.
- Masci, J., Rodolà, E., Boscaini, D., Bronstein, M.M., Li, H., 2016. Geometric deep learning, in: *SIGGRAPH ASIA 2016 Courses*, pp. 1–50. doi:10.1145/2988458.2988485.
- Mazoyer, B., Mellet, E., Percey, G., Zago, L., Crivello, F., Jobard, G., Delcroix, N., Vigneau, M., Leroux, G., Petit, L., Joliot, M., Tzourio-Mazoyer, N., 2016. BIL&GIN: A neuroimaging, cognitive, behavioral, and genetic database for the study of human brain lateralization. *NeuroImage* 124, 1225–1231. doi:10.1016/j.neuroimage.2015.02.071.
- Mori, S., Crain, B.J., Chacko, V.P., Van Zijl, P., 1999. Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging. *Annals of Neurology* 45, 265–269. doi:10.1002/1531-8249(199902)45:2<265::AID-ANA21>3.0.CO;2-3.
- Nath, V., Schilling, K.G., Parvathaneni, P., Huo, Y., Blaber, J.A., Hainline, A.E., Barakovic, M., Romascano, D., Rafael-Patino, J., Frigo, M., Girard, G., Thiran, J.P., Daducci, A., Rowe, M., Rodrigues, P., Prčkovska, V., Aydogan, D.B., Sun, W., Shi, Y., Parker, W.A., Ould Ismail, A.A., Verma, R., Cabeen, R.P., Toga, A.W., Newton, A.T., Wasserthal, J., Neher, P., Maier-Hein, K., Savini, G., Palesi, F., Kaden, E., Wu, Y., He, J., Feng, Y., Paquette, M., Rheault, F., Sidhu, J., Lebel, C., Leemans, A., Descoteaux, M., Dyrby, T.B., Kang, H., Landman, B.A., 2020. Tractography reproducibility challenge with empirical data (TraCED): The 2017 ISMRM diffusion study group challenge. *Journal of Magnetic Resonance Imaging* 51, 234–249. doi:10.1002/jmri.26794.
- Neher, P.F., Descoteaux, M., Houde, J.C., Stieltjes, B., Maier-Hein, K.H., 2015. Strengths and weaknesses of state of the art fiber tractography pipelines – A comprehensive in-vivo and phantom evaluation study using Tractometer. *Medical Image Analysis* 26, 287–305. doi:10.1016/j.media.2015.10.011.
- Neher, P.F., Stieltjes, B., Maier-Hein, K.H., 2018. Anchor-Constrained Plausibility (ACP): A Novel Concept for Assessing Tractography and Reducing False-Positives, in: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pp. 20–27. doi:10.1007/978-3-030-00931-1\_3.
- Nie, X., Shi, Y., 2019. Topographic Filtering of Tractograms as Vector Field Flows, in: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pp. 564–572. doi:10.1007/978-3-030-32248-9\_63.
- Ocampo-Pineda, M., Schiavi, S., Rheault, F., Girard, G., Petit, L., Descoteaux, M., Daducci, A., 2021. Hierarchical Microstructure Informed Tractography. *Brain Connectivity* doi:10.1089/brain.2020.0907.
- O'Donnell, L.J., Wells, W.M., Golby, A.J., Westin, C.F., 2012. Unbiased Groupwise Registration of White Matter Tractography, in: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012*, pp. 123–130. doi:10.1007/978-3-642-33454-2\_16.

- O'Donnell, L.J., Westin, C.F., 2007. Automatic Tractography Segmentation Using a High-Dimensional White Matter Atlas. *IEEE Transactions on Medical Imaging* 26, 1562–1575. doi:10.1109/TMI.2007.906785.
- Olivetti, E., Avesani, P., 2011. Supervised segmentation of fiber tracts, in: *Proceedings of the First International Conference on Similarity-based Pattern Recognition*, pp. 261–274. doi:10.1007/978-3-642-24471-1\_19.
- Pang, G., Shen, C., Cao, L., Hengel, A.V.D., 2021. Deep learning for anomaly detection: A review. *ACM Computing Surveys (CSUR)* 54, 1–38.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., 2019. PyTorch: An imperative style, high-performance deep learning library, in: *Advances in Neural Information Processing Systems*, pp. 8024–8035.
- Pestilli, F., Yeatman, J.D., Rokem, A., Kay, K.N., Wandell, B.A., 2014. Evaluation and statistical inference for human connectomes. *Nature Methods* 11, 1058–1063. doi:10.1038/nmeth.3098.
- Petit, L., Ali, K.M., Rheault, F., Boré, A., Cremona, S., Corsini, F., De Benedictis, A., Descoteaux, M., Sarubbo, S., 2022. The structural connectivity of the human angular gyrus as revealed by microdissection and diffusion tractography. *Brain Structure and Function* doi:10.1007/s00429-022-02551-5.
- Pierpaoli, C., Jezzard, P., Basser, P.J., Barnett, A., Di Chiro, G., 1996. Diffusion tensor MR imaging of the human brain. *Radiology* 201, 637–648. doi:10.1148/radiology.201.3.8939209.
- Presseau, C., Jodoin, P.M., Houde, J.C., Descoteaux, M., 2015. A new compression format for fiber tracking datasets. *NeuroImage* 109, 73–83. doi:10.1016/j.neuroimage.2014.12.058.
- Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 652–660. arXiv:1612.00593.
- Raffelt, D., Tournier, J.D., Rose, S., Ridgway, G.R., Henderson, R., Crozier, S., Salvado, O., Connelly, A., 2012. Apparent Fibre Density: A novel measure for the analysis of diffusion-weighted magnetic resonance images. *NeuroImage* 59, 3976–3994. doi:10.1016/j.neuroimage.2011.10.045.
- Reddy, C.P., Rathi, Y., 2016. Joint Multi-Fiber NODDI Parameter Estimation and Tractography Using the Unscented Information Filter. *Frontiers in Neuroscience* 10. doi:10.3389/fnins.2016.00166.
- Rheault, F., De Benedictis, A., Daducci, A., Maffei, C., Tax, C.M.W., Romascano, D., Caverzasi, E., Morency, F.C., Corrivetti, F., Pestilli, F., Girard, G., Theaud, G., Zemmoura, I., Hau, J., Glavin, K., Jordan, K.M., Pomiecko, K., Chamberland, M., Barakovic, M., Goyette, N., Poulin, P., Chenot, Q., Panesar, S.S., Sarubbo, S., Petit, L., Descoteaux, M., 2020a. Tractostorm: The what, why, and how of tractography dissection reproducibility. *Human Brain Mapping* 41, 1859–1874. doi:10.1002/hbm.24917.
- Rheault, F., Poulin, P., Valcourt Caron, A., St-Onge, E., Descoteaux, M., 2020b. Common misconceptions, hidden biases and modern challenges of dMRI tractography. *Journal of Neural Engineering* 17, 011001. doi:10.1088/1741-2552/ab6aad.
- Rheault, F., St-Onge, E., Sidhu, J., Maier-Hein, K., Tzourio-Mazoyer, N., Petit, L., Descoteaux, M., 2019. Bundle-specific tractography with incorporated anatomical and orientational priors. *NeuroImage* 186, 382–398. doi:10.1016/j.neuroimage.2018.11.018.
- Schiavi, S., Ocampo-Pineda, M., Barakovic, M., Petit, L., Descoteaux, M., Thiran, J.P., Daducci, A., 2020. A new method for accurate in vivo mapping of human brain connections using microstructural and anatomical information. *Science Advances* 6, eaba8245. doi:10.1126/sciadv.aba8245.
- Schilling, K.G., Nath, V., Hansen, C., Parvathaneni, P., Blaber, J., Gao, Y., Neher, P., Aydogan, D.B., Shi, Y., Ocampo-Pineda, M., Schiavi, S., Daducci, A., Girard, G., Barakovic, M., Rafael-Patino, J., Romascano, D., Rensonnet, G., Pizzolato, M., Bates, A., Fisch, E., Thiran, J.P., Canales-Rodríguez, E.J., Huang, C., Zhu, H., Zhong, L., Cabeen, R., Toga, A.W., Rheault, F., Theaud, G., Houde, J.C., Sidhu, J., Chamberland, M., Westin, C.F., Dyrby, T.B., Verma, R., Rathi, Y., Irfanoglu, M.O., Thomas, C., Pierpaoli, C., Descoteaux, M., Anderson, A.W., Landman, B.A., 2019. Limits to anatomical accuracy of diffusion tractography using modern approaches. *NeuroImage* 185, 1–11. doi:10.1016/j.neuroimage.2018.10.029.
- Schilling, K.G., Petit, L., Rheault, F., Remedios, S., Pierpaoli, C., Anderson, A.W., Landman, B.A., Descoteaux, M., 2020. Brain connections derived from diffusion MRI tractography can be highly anatomically accurate—if we know where white matter pathways start, where they end, and where they do not go. *Brain Structure and Function* 225, 2387–2402. doi:10.1007/s00429-020-02129-z.
- Smith, R., Raffelt, D., Tournier, J.D., Connelly, A., 2020. Quantitative streamlines tractography: Methods and inter-subject normalisation doi:10.31219/osf.io/c67kn.
- Smith, R.E., Tournier, J.D., Calamante, F., Connelly, A., 2012. Anatomically-constrained tractography: Improved diffusion MRI streamlines tractography through effective use of anatomical information. *NeuroImage* 62, 1924–1938. doi:10.1016/j.neuroimage.2012.06.005.
- Smith, R.E., Tournier, J.D., Calamante, F., Connelly, A., 2015a. The effects of SIFT on the reproducibility and biological accuracy of the structural connectome. *NeuroImage* 104, 253–265. doi:10.1016/j.neuroimage.2014.10.004.
- Smith, R.E., Tournier, J.D., Calamante, F., Connelly, A., 2015b. SIFT2: Enabling dense quantitative assessment of brain white matter connectivity using streamlines tractography. *NeuroImage* 119, 338–351. doi:10.1016/j.neuroimage.2015.06.092.
- Smith, R.E., Tournier, J.D., Calamante, F., Connelly, A., 2013. SIFT: Spherical-deconvolution informed filtering of tractograms. *NeuroImage* 67, 298–312.

- Takemura, H., Caiafa, C.F., Wandell, B.A., Pestilli, F., 2016. Ensemble Tractography. *PLoS computational biology* 12.
- 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. *Proceedings of the National Academy of Sciences of the United States of America* 111, 16574–16579. doi:10.1073/pnas.1405672111.
- Tournier, J.D., Calamante, F., Connelly, A., 2007. Robust determination of the fibre orientation distribution in diffusion MRI: Non-negativity constrained super-resolved spherical deconvolution. *NeuroImage* 35, 1459–1472. doi:10.1016/j.neuroimage.2007.02.016.
- Tournier, J.D., Smith, R., Raffelt, D., Tabbara, R., Dhollander, T., Pietsch, M., Christiaens, D., Jeurissen, B., Yeh, C.H., Connelly, A., 2019. MRtrix3: A fast, flexible and open software framework for medical image processing and visualisation. *NeuroImage* 202, 116137. doi:10.1016/j.neuroimage.2019.116137.
- Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9.
- 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.
- Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A., Bottou, L., 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research* 11.
- Wang, J., Aydogan, D.B., Varma, R., Toga, A.W., Shi, Y., 2018. Modeling topographic regularity in structural brain connectivity with application to tractogram filtering. *NeuroImage* 183, 87–98. doi:10.1016/j.neuroimage.2018.07.068.
- Wang, R., Benner, T., Sorensen, A.G., Wedeen, V.J., 2007. Diffusion toolkit: A software package for diffusion imaging data processing and tractography, in: *Proceedings of International Society of Magnetic Resonance in Medicine (ISMRM)*.
- Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M., 2019. Dynamic Graph CNN for Learning on Point Clouds. *ACM Transactions on Graphics* 38, 146.
- 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., Hirjak, D., Maier-Hein, K.H., 2019. Combined tract segmentation and orientation mapping for bundle-specific tractography. *Medical Image Analysis* 58, 101559. doi:10.1016/j.media.2019.101559.
- Xia, Y., Shi, Y., 2020. Groupwise track filtering via iterative message passing and pruning. *NeuroImage* 221, 117147. doi:10.1016/j.neuroimage.2020.117147.
- Xu, H., Dong, M., Lee, M.H., O'Hara, N., Asano, E., Jeong, J.W., 2019. Objective Detection of Eloquent Axonal Pathways to Minimize Postoperative Deficits in Pediatric Epilepsy Surgery Using Diffusion Tractography and Convolutional Neural Networks. *IEEE Transactions on Medical Imaging* 38, 1910–1922. doi:10.1109/TMI.2019.2902073.
- Yang, J.Y.M., Yeh, C.H., Poupon, C., Calamante, F., 2021. Diffusion MRI tractography for neurosurgery: The basics, current state, technical reliability and challenges. *Physics in Medicine & Biology* doi:10.1088/1361-6560/ac0d90.
- Yeh, C.H., Jones, D.K., Liang, X., Descoteaux, M., Connelly, A., 2020. Mapping Structural Connectivity Using Diffusion MRI: Challenges and Opportunities. *Journal of Magnetic Resonance Imaging*, jMRI.27188doi:10.1002/jmri.27188.
- Yeh, C.H., Smith, R.E., Liang, X., Calamante, F., Connelly, A., 2016. Correction for diffusion MRI fibre tracking biases: The consequences for structural connectomic metrics. *NeuroImage* 142, 150–162. doi:10.1016/j.neuroimage.2016.05.047.
- Yeh, F.C., Panesar, S., Barrios, J., Fernandes, D., Abhinav, K., Meola, A., Fernandez-Miranda, J.C., 2019. Automatic Removal of False Connections in Diffusion MRI Tractography Using Topology-Informed Pruning (TIP). *Neurotherapeutics* 16, 52–58. doi:10.1007/s13311-018-0663-y.
- 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. *Frontiers in Neuroinformatics* 5. doi:10.3389/fninf.2011.00023.
- Zalesky, A., Fornito, A., Cocchi, L., Gollo, L.L., van den Heuvel, M.P., Breakspear, M., 2016. Connectome sensitivity or specificity: Which is more important? *NeuroImage* 142, 407–420. doi:10.1016/j.neuroimage.2016.06.035.
- Zhang, F., Cetin Karayumak, S., Hoffmann, N., Rathi, Y., Golby, A.J., O'Donnell, L.J., 2020. Deep white matter analysis (DeepWMA): Fast and consistent tractography segmentation. *Medical Image Analysis* 65, 101761. doi:10.1016/j.media.2020.101761.
- Zhang, F., Daducci, A., He, Y., Schiavi, S., Seguin, C., Smith, R.E., Yeh, C.H., Zhao, T., O'Donnell, L.J., 2022. Quantitative mapping of the brain's structural connectivity using diffusion MRI tractography: A review. *NeuroImage* 249, 118870. doi:10.1016/j.neuroimage.2021.118870.
- 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.



## Appendix A. Additional Experiments with Verifyber

### Appendix A.1. Verifyber Test on BIL&GIN data

*BILGIN-EP*. BIL&GIN (Mazoyer et al., 2016) data source. Exclusive labeling (Petit et al., 2022). Single subject, with tractogram of 1.5M streamlines obtained via PF-ACT tractography (Girard et al., 2020).

Table A.7: Generalization test on BILGIN-EP. Models have been trained on one of the 5-fold split of HCP-EP. We do not report standard deviation, as BILGIN-EP is a single-subject dataset. The results confirm Verifyber as best performing model. It maintains very high precision, while it loose some points in terms of recall. This is probably due to the different distribution of plausible/non-plausible.

Method	Accuracy	Precision	Recall	DSC
bLSTM	90.7	94.4	93.1	93.7
PN	91.2	95.1	93.1	94.1
DGCNN	91.4	95.5	92.8	94.1
<b>VF</b>	<b>92.5</b>	<b>96.2</b>	<b>93.7</b>	<b>94.9</b>

### Appendix A.2. Training Verifyber with Signal-based Supervision

In this study, we adopted fiber *weights* obtained with SIFT2 (Smith et al., 2015b) as target for Verifyber, which was trained for regression. The results suggested that VF can learn weights regression decently well, but cannot capture the density-based filtering criteria. In fact, with tractograms having similar fiber density map, VF was able to predict weights, but as soon as new tractograms had different fiber density our methods failed. See Figures A.16, A.17, A.18

Table A.8: Regression results on SIFT2 weights for HCP-EP. Both Verifyber and PointNet (Qi et al., 2017) were trained on the training subjects of HCP-EP, for which we computed the SIFT2 weights. The table shows the test results averaged over 4 test subjects in terms of Mean Absolute Error(MAE) — the same metric optimized during training —, and Spearman’s rank correlation coefficient. The former indicate the absolute regression error, while the latter indicate how much the ranking of fibers obtained for SIFT2 weights is preserved after weights prediction.

Model	CSD	MAE	Spear. Corr.
Verifyber	DiPy	0.173 ( $\pm 0.011$ )	0.766 ( $\pm 0.039$ )
PointNet	DiPy	0.175 ( $\pm 0.014$ )	0.757 ( $\pm 0.033$ )

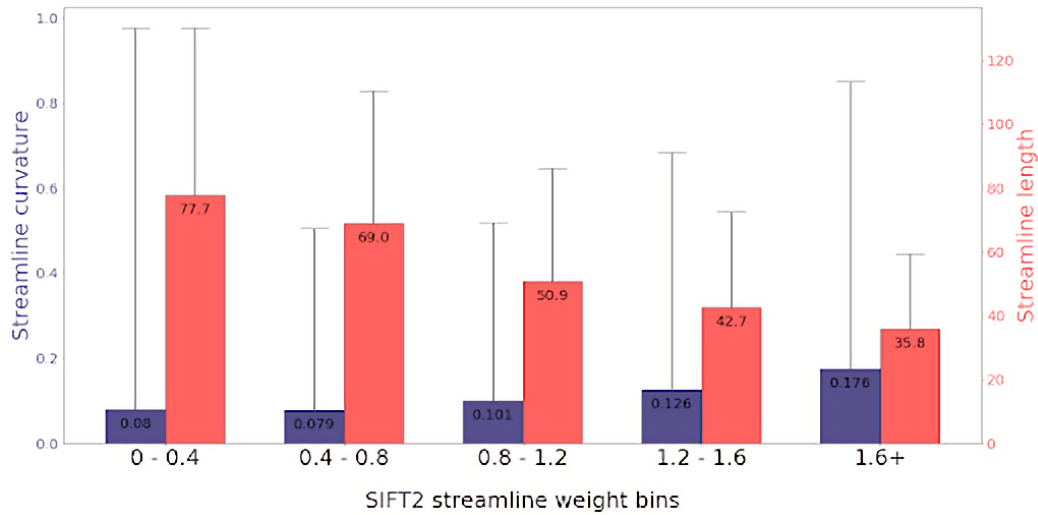


Figure A.16: SIFT2 weights per streamline categories in a single tractogram. Relationship between streamline anatomy and SIFT2 streamline weights. Each weight interval contains at least 10% of fibers. The trend of length and curvature shows that high-weight is assigned to very curved and short fibers. Viceversa, low-weight is given to straight and long fibers. This behavior is almost opposite compared to the anatomical labeling of Extractor (Petit et al., 2022). Note also a very high standard deviation.

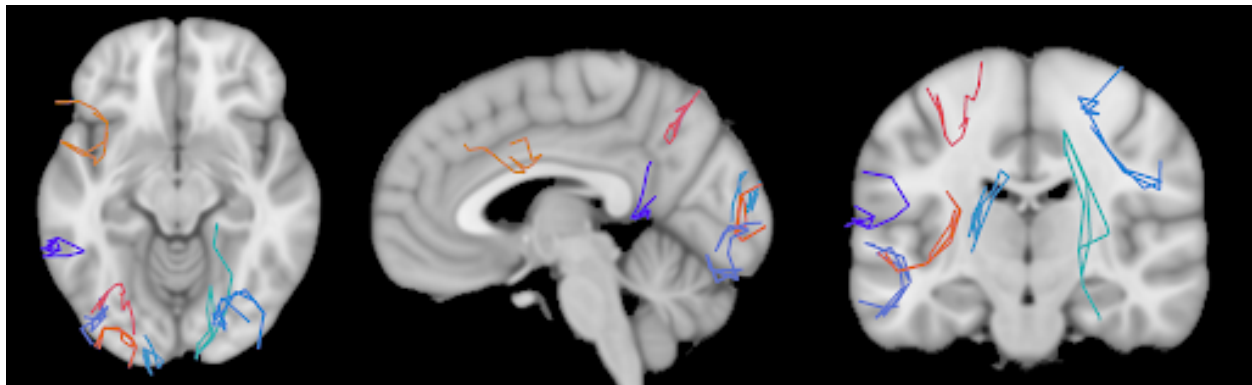


Figure A.17: Visual example of non-plausible fibers according to SIFT2. 8 streamlines with weights > 3, lengths > 92 mm, and curvatures > 2 (MNI space, downsampled to 16 points). The figure shows that SIFT2 weights are not a good indicator for the anatomical plausibility of a streamline.

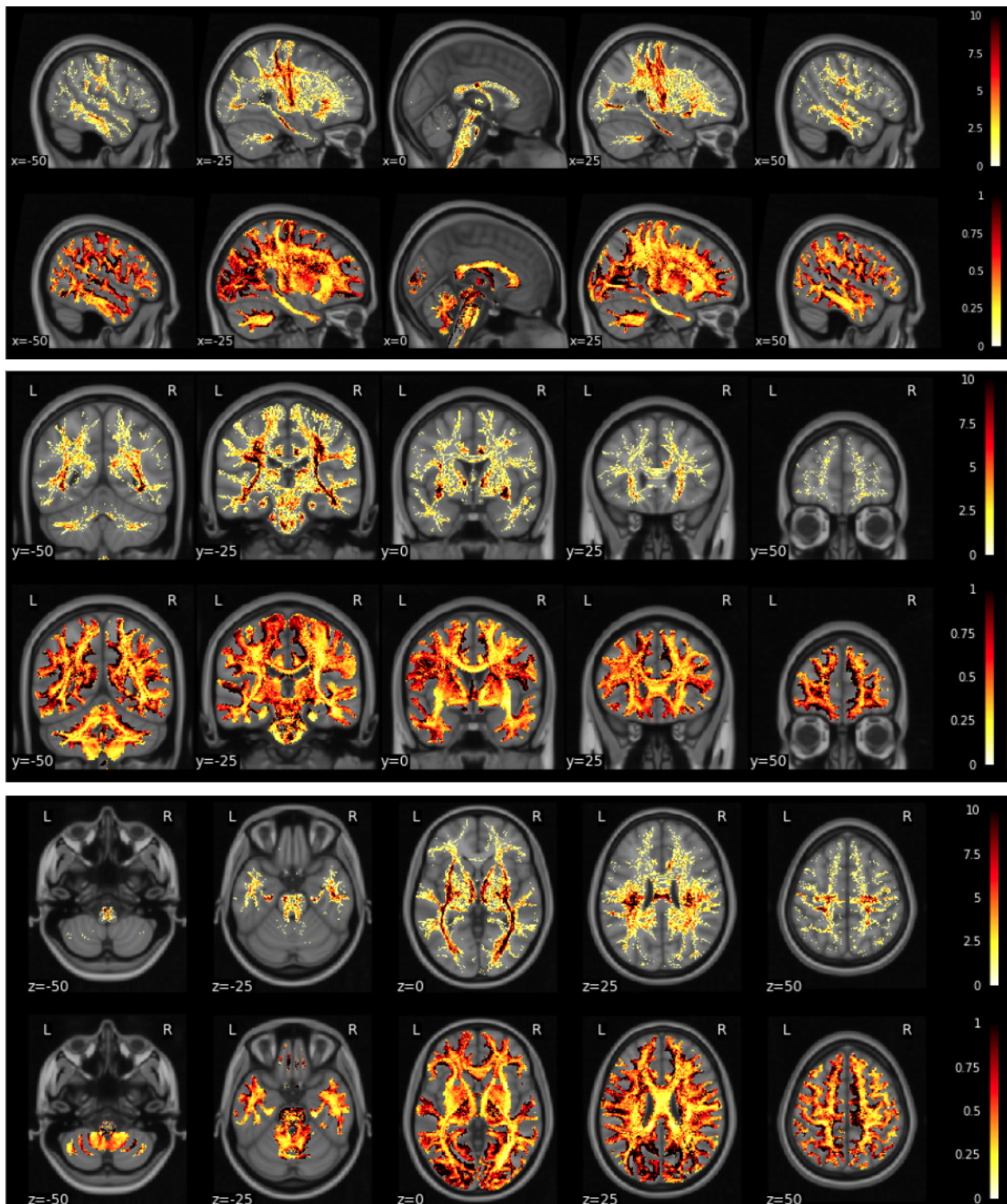


Figure A.18: Spatial distribution of SIFT2 weights vs. fiber density. SIFT weights distribution (top rows) compared to streamline density map (bottom rows). Observing the two types of maps, it is easy to see the high degree of correlation.

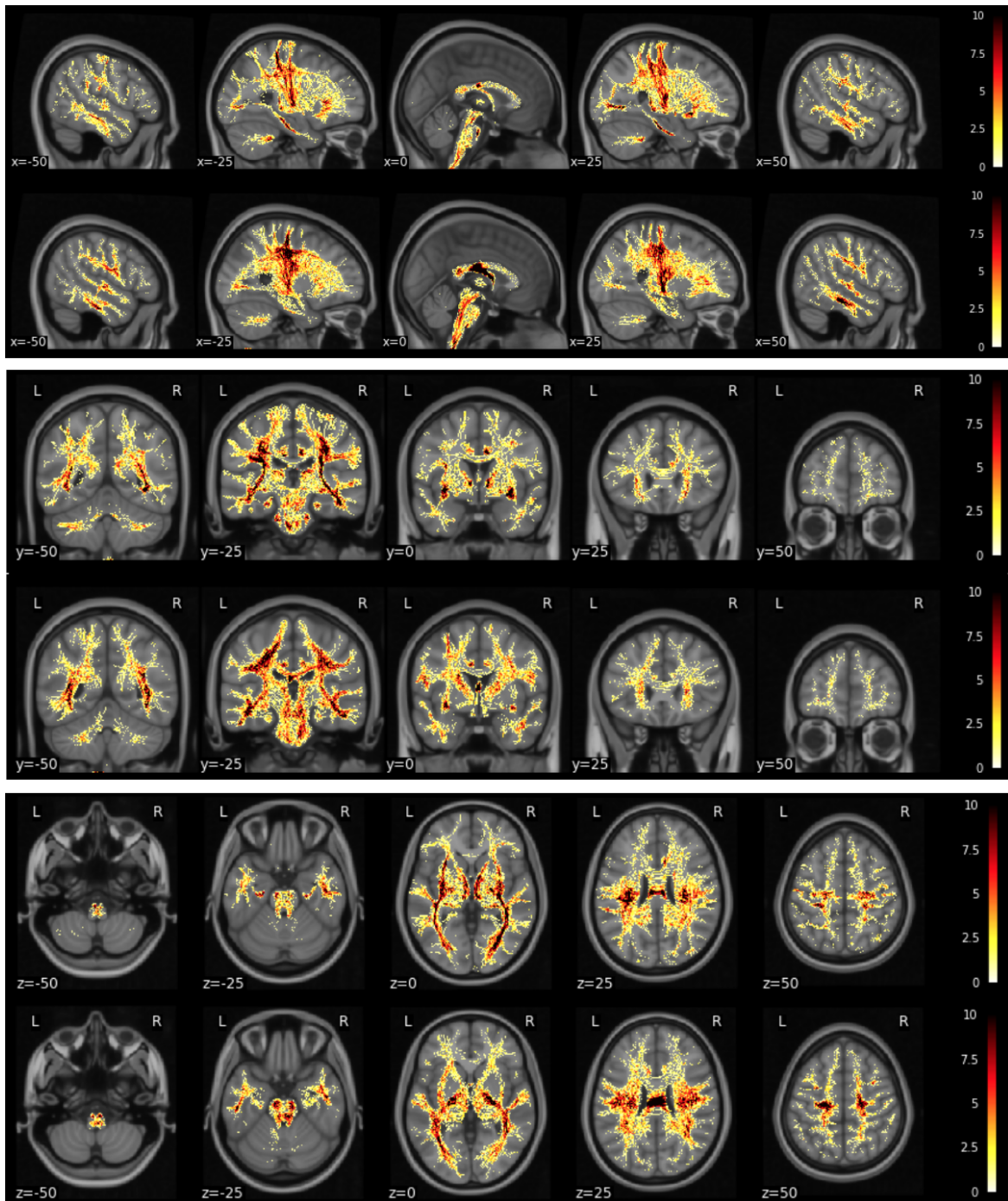


Figure A.19: Fiber density map obtained after Ensemble Tractography (Takemura et al., 2016) (top rows) and Particle Filtering tractography (Girard et al., 2014) (bottom rows). This figure shows that we cannot expect generalization of VF<sup>SIFT2</sup> across the two kind of tractograms.

## Appendix B. Investigation of FINTA

Additional results obtained with the reproduced FINTA (Legarreta et al., 2021). See Figures B.20, B.21, B.22.

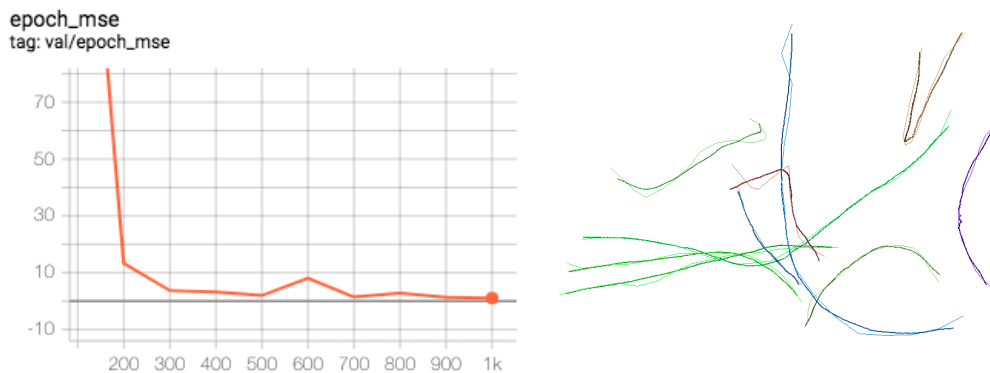


Figure B.20: Reproduced FINTA autoencoder. The plot on the left refer to a training performed on the dataset HCP-IZ. In this experiment, according to the setting used by the Authors of FINTA, we split the streamlines of the HCP-IZ tractogram into 80/20 train/test. The train set was then split in 4 buckets, 3 used for train and 1 for validation, which is performed once every 100 epochs. The training is stopped after 1000 epochs as it has converged to an MSE value in validation  $\sim 1\text{mm}$ . The image on the right depicts the qualitative reconstruction using the trained autoencoder. Thin fibers are the original, while fat are the reconstructed. It can be observed that the reconstruction is smooth and reasonably correct.



Figure B.21: Non-plausible fibers (HCP-EP) reconstructed using FINTA autoencoder. As can be noted original streamlines (thin) have a clear non-plausible trajectory with sharp curves. The denoising action performed by the autoencoder reconstruct smoother trajectories (fat), which we may observe to be more similar to plausible than to non-plausible fibers. Such denoising behaviour is expected when adopting an autoencoder model. However, as the output reconstructions reflect the learned embedding, such a behaviour can be a source of error for the task of tractogram filtering.

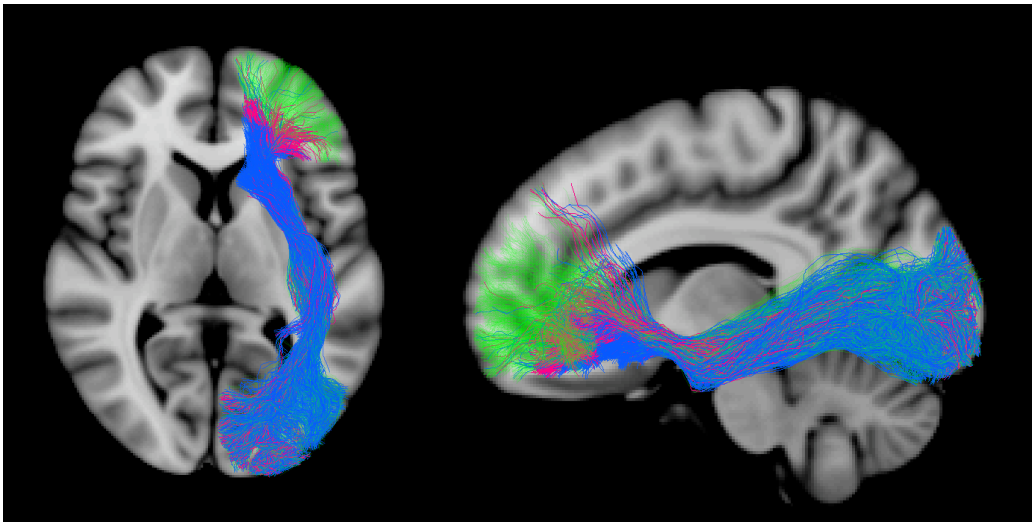


Figure B.22: IFOF left from HCP-IW Wasserthal et al. (2018) after classification of streamlines as plausible ( $p$ ) and non-plausible ( $np$ ) using FINTA trained on HCP-IZ Zhang et al. (2018). Considering the same classification performed with Verifyer (see Figure 16 of the paper) as a reference labeling, we investigate the possible false positive prediction of FINTA i.e., fibers that are  $np$  according to Verifyer, but that FINTA classifies as  $p$ . In the figure we indicate such fibers in fuchsia, while we color with green and blue fibers classified by both methods as  $p$  and  $np$  respectively.