UNIVERSITÀ DEGLI STUDI
DI TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE
ICT International Doctoral School

Toward the “Deep Learning” of Brain White Matter Structures

Doctoral thesis of

Pietro Astolfi

Advisor

Dr. Paolo Avesani

Co-Advisor

Dr. Diego Sona

2021

Abstract

In the brain, neuronal cells located in different functional regions communicate through a dense structural network of axons known as the white matter (WM) tissue. Bundles of axons that share similar pathways characterize the WM anatomy, which can be investigated in-vivo thanks to the recent advances of magnetic resonance (MR) techniques.

Diffusion MR imaging combined with tractography pipelines allows for a virtual reconstruction of the whole WM anatomy of in-vivo brains, namely the tractogram. It consists of millions of WM fibers as 3D polylines, each approximating thousands of axons. From the analysis of a tractogram, neuroanatomists can characterize well-known white matter structures and detect anatomically non-plausible fibers, which are artifacts of the tractography and often constitute a large portion of it. The accurate characterization of tractograms is pivotal for several clinical and neuroscientific applications. However, such characterization is a complex and time-consuming process that is difficult to be automatized as it requires properly encoding well-known anatomical priors.

In this thesis, we propose to investigate the encoding of anatomical priors with a supervised deep learning framework. The ultimate goal is to reduce the presence of artifactual fibers to enable a more accurate automatic process of WM characterization. We devise the problem by distinguishing between volumetric and non-volumetric representations of white matter structures. In the first case, we learn the segmentation of the WM regions that represent relevant anatomical waypoints not yet classified by WM atlases. We investigate using Convolutional Neural Networks (CNNs) to exploit the volumetric representation of such priors. In the second case, the goal is to learn from the 3D polyline representation of fibers where the typical CNN models are not suitable. We introduce the novelty of using Geometric Deep Learning (GDL) models designed to process data having an irregular representation. The working assumption is that the geometrical properties of fibers are informative for the detection of tractogram artifacts.

As a first contribution, we present StemSeg that extends the use of CNNs to detect the WM portion representing the waypoints of all the fibers for a specific bundle. This anatomical landmark, called stem, can be critical for extracting that bundle. We provide the results of an empirical analysis focused on the Inferior Fronto-Occipital Fasciculus (IFOF). The effective segmentation of the stem improves the final segmentation of the IFOF, outperforming with a significant gap the reference state of the art.

As a second and major contribution, we present Verifyber, a supervised tractogram filtering approach based on GDL, distinguishing between anatomically plausible and non-plausible fibers. The proposed model is designed to learn anatomical features directly from the fiber represented as a 3D points sequence. The extended empirical analysis on healthy and clinical subjects reveals multiple benefits of Verifyber: high filtering accuracy, low inference time, flexibility to different plausibility definitions, and good generalization.

Overall, this thesis constitutes a step toward characterizing white matter using deep learning. It provides effective ways of encoding anatomical priors and an original deep learning model designed for fiber.

Keywords [Brain, Neuroimaging, Tractography, Deep Learning, Geometric Deep Learning]

Contents

List of Tables

List of Figures

Publications

Amorosino, G., Peruzzo, D., Astolfi, P., Redaelli, D., Avesani, P., Arrigoni, F., and Olivetti, E. (2020). Automatic Tissue Segmentation with Deep Learning in Patients with Congenital or Acquired Distortion of Brain Anatomy. In *Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-Oncology*, pages 13–22. Springer.

Astolfi, P., De Benedictis, A., Sarubbo, S., Bertó, G., Olivetti, E., Sona, D., and Avesani, P. (2020a). A Stem-Based Dissection of Inferior Fronto-Occipital Fasciculus with A Deep Learning Model. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 267–270. IEEE.

Astolfi, P., Verhagen, R., Petit, L., Olivetti, E., Masci, J., Boscaini, D., and Avesani, P. (2020b). Tractogram Filtering of Anatomically Non-plausible Fibers with Geometric Deep Learning. In Martel, A. L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M. A., Zhou, S. K., Racoceanu, D., and Joskowicz, L., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 291–301, Cham. Springer International Publishing.

Bertò, G., Bullock, D., Astolfi, P., Hayashi, S., Zigiotto, L., Annicchiarico, L., Corsini, F., De Benedictis, A., Sarubbo, S., and Pestilli, F. (2021). Classifyber, a robust streamline-based linear classifier for white matter bundle segmentation. *NeuroImage*, 224:117402.

Olivetti, E., Gori, P., Astolfi, P., Bertó, G., and Avesani, P. (2020). Nonlinear Alignment of Whole Tractograms with the Linear Assignment Problem. In *International Workshop on Biomedical Image Registration*, pages 3–11. Springer.

Svoboda, J., Astolfi, P., Boscaini, D., Masci, J., and Bronstein, M. (2020). Clustered Dynamic Graph CNN for Biometric 3D Hand Shape Recognition. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE.

Chapter 1

Introduction

The white matter structure of the human brain The human brain can perform plenty of tasks, like processing sensory signals, reasoning, memorizing, and regulating the human body. The execution of such tasks happens across different regions of the brain, usually located in the cerebral cortex. Neuronal cells in the cortical regions exchange information between each other through electrochemical spikes traveling along *axonal pathways*. Most of the axonal pathways traverse the internal part of the brain to connect neurons of different regions, forming a dense connective tissue called *white matter* (WM). Within the white matter, axons are organized in *bundles* of similar pathways that share the same connection scope (Meynert, 1885). The spatial disposition of these bundles constitute the so-called *WM structure* or *anatomy* or *structural connectivity*. Unfortunately, the comprehensive organization of WM structural connectivity is still unknown, and the scientific debate is discussing further WM characterizations. For example, neuroanatomists frequently discover or refine landmark anatomical regions where bundles pass through or cross/intersect with other bundles (Hau et al., 2017). Characterizing more accurately the WM is beneficial for many clinical applications like surgical intervention, brain diseases, neuroplasticity understanding, and neuroscientific studies (Jeurissen et al., 2017; De Benedictis et al., 2016; Yang et al., 2021; Yeh et al., 2020; Kruper et al., 2021).

The anatomy of the white matter has been historically studied via invasive techniques, such as histology, direct stimulation, or post-mortem dissection. Only in the last 20 years, the advancement of neuroimaging techniques allowed reliable non-invasive analysis. The development of diffusion magnetic resonance imaging (dMRI) (Le Bihan et al., 1986) has enabled the in-vivo inspection of the white matter anatomy. dMRI estimates a map (3d image) of the local directions of white matter pathways by measuring the diffusivity of water molecules, which diffuse along the myelin sheath that surrounds the axons (Basser et al., 1994). Nowadays, thanks to complex computational pipelines built upon dMRI data (Basser et al., 2000), we can retrieve a virtual approximation of the whole white matter connectivity, namely the *tractogram*, which is a collection of millions of virtual white matter *fibers* each representing thousands of axonal pathways. The WM fibers, also referred to as *streamlines*, are virtually modeled as *3D polylines*, i.e., unoriented sequences of points in the 3D space, having varying length and trajectory. The grouping of anatomically similar fibers constitutes *bundles.*

Deep learning Along with neuroimaging advancements, the last decades have witnessed the widespread use of machine learning (ML) approaches as computational tools to autonomously complete tasks driven by data experience. For example, *deep learning* (DL) (LeCun et al., 2015; Schmidhuber, 2015) has been a breakthrough for many applications, such as computer vision (Krizhevsky et al., 2012) and natural language processing (Bahdanau et al., 2015). The main reason for the DL success resides in the ability of *deep neural networks* (NNs) to learn new *representations* of the input signal (Bengio et al., 2013), that are relevant for the target task at hand, e.g., image classification and segmentation. This property allows scientists to avoid the manual engineering of discriminative data representations, i.e., *feature vectors*, which is necessary for traditional ML models as a pre-condition to obtain good performance. The recent leap in high-performance parallel computing allows the training of *complex* DL models having millions of learnable *parameters*, which may encode thousands of features. In the presence of enough data, learning richer feature representations often leads to better performance than using traditional handcraft feature vectors. On the contrary, when too few training data are available, the quality of the learned representation is not *general* enough to successfully perform the target tasks on unseen data.

The amount of data required by DL models exponentially increases with the model complexity/size, i.e., the number of learnable parameters, becoming quickly impossible to be collected and manually annotated. To solve this issue known as *curse of dimensionality* (Bach, 2017) and learn meaningful representations with less data, a NN must exploit the structural regularities present in the data itself. This is possible by encoding such regularities within the NN architecture as *structural inductive bias* (Mitchell, 1980). For example, *Convolutional Neural Networks* (CNNs) (LeCun and Bengio, 1995), which show good performance on grid-based data like images, use local convolutional and pooling filters that are invariant/equivariant to shift and local deformation of the input. However, CNNs are not ideal with irregular, non-grid data structures, and other types of architectures may be needed. *Geometric Deep Learning* (GDL) (Bronstein et al., 2017) was born to provide suitable architectures for domains like graphs (Kipf and Welling, 2017) and point-clouds (Qi et al., 2017; Wang et al., 2019), where there is no grid structure, and the NN inductive bias must ensure permutation invariance/equivariance on nodes and points respectively.

1.1 Motivations

Open problems of tractograms Tractograms are data derivatives obtained as the outcome of a complex pipeline of data processing. After the diffusion MRI preprocessing (Glasser et al., 2013; Fischl, 2012; Jenkinson et al., 2012; Tournier et al., 2019) and diffusivity *model reconstruction* (Pierpaoli et al., 1996; Tournier et al., 2007; Descoteaux, 2015), the *tractography* or *tracking* (Mori et al., 1999; Basser et al., 2000; Jeurissen et al., 2017) computes the fiber pathways. The tracking process is an iterative procedure that requires the definition of many parameters, like the policy of seeding (where to start tracking), the strategy of stepping (how to track next point of a fiber), the stop criterion (where/when to stop tracking), and the constraints on curvature and length of fibers. Slightly different parameter choices may produce quite different tractograms (Thomas et al., 2014), and the evaluation of their accuracy is not straightforward (Neher et al., 2015).

In recent years, the tractography community has given great attention to the evaluation of the tractograms accuracy. Multiple international contests, e.g., (Fillard et al., 2011; Maier-Hein et al., 2017; Schilling et al., 2019b; Nath et al., 2020), have proposed the use of physical phantoms, simulated diffusion data, post-mortem dissection, and histological tracing as ground-truth reference. While the debate on the definition of a proper ground-truth is still open (Rheault et al., 2020a) — and far from ending —, state-of-the-art tractography pipelines can produce tractograms comprehensive of the well-known WM pathways but at the cost of generating a significant amount of *false positive* fibers (Maier-Hein et al., 2017; Schilling et al., 2019a). Usually, they draw twisted, unrealistic fiber trajectories or fibers that traverse/terminate in the wrong anatomical regions of the WM. Such fibers are not compliant with the underlying anatomy, and they are artifacts due to the intrinsic ambiguity of the diffusion signal (Jeurissen et al., 2017). Different dispositions of axonal pathways, e.g., *kissing and crossing fibers*, might correspond to the same diffusion signal.

To mitigate the false positives problem, there is the need to enrich the characterization of anatomical principles in tractography and tractograms. This can be done by imposing *anatomical priors* during the tracking or directly on the reconstructed tractogram. For example, there exist attempts to restrict the tracking area to bundle-specific regions (Rheault et al., 2019), or to impose the passage/termination of fibers only in certain tissues (Smith et al., 2012). However, these are not the only ways of encoding anatomical priors, and other approaches define anatomical priors as fiber features (Petit et al., 2019) like fiber geometry. In general, the anatomical prior definition can involve either (i) specific regions of the white matter (*ROI-based prior*), or (ii) properties inherent of the fiber trajectory (*fiber-based prior*).

ROI-based prior. Anatomical regions are typically used by neuroanatomists to syntactically characterize and extrapolate bundles. In ex-vivo dissection, they locate key white matter regions and scrape the surroundings following the route traced by axonal pathways until dissecting an entire bundle. Similarly, in virtual bundle segmentation, experts use volumetric masks of the region of interests (ROIs) already available in brain parcellation atlases to obtain a first very broad segmentation. Then, they manually draw additional ROIs to refine the segmentation and avoid *false positive* streamlines.

In the recent literature, there have been many tentatives to automatize ROIs for bundle segmentation. From the one side, there is a series of methods that seek to emulate manual bundle segmentation by using *segmentation rules* expressed as logical combinations of ROIs, e.g., White Matter Query Language (WMQL) (Wassermann et al., 2013, 2016) and Automated Fiber Quantification (AFQ) (Yeatman et al., 2012). Rules are usually defined on template brains, where ROI atlases are ready-to-use, and then are applied to new subjects via volumetric brain co-registration. On the other side, there are data-driven approaches like TractSeg (Wasserthal et al., 2018), that cast the segmentation of bundles as a single ROI segmentation, i.e., the *bundle volumetric mask*. These approaches exploit the exceptional power of CNN in segmenting images to segment bundle masks directly from fiber orientation maps.

However, there are some pitfalls in both approaches that may reduce the accuracy of the final bundle segmentation. In particular, the accuracy is directly related to the quality of the ROIs. In the rule-based approaches, ROIs from existing atlases are rather coarse due to the standardization process (Figley et al., 2017) and characterize only partially the white matter anatomy (Hansen et al., 2021). Moreover, the co-registration step is complex and frequently registers poorly the WM tissue. In CNN-based approaches, despite co-registration is not needed, the predicted bundle masks hardly capture the detailed structure of bundles. Studies like (Bertò et al., 2021) have empirically shown the tendency of CNN-based approaches, e.g., TractSeg, to miss details in the most complex bundles — generally in the termination regions.

Fiber-based prior Fiber-based characterization of WM priors can be considered as an alternative strategy that goes beyond the possible weaknesses of ROIs, which may not remove/avoid some false positive fibers. In this characterization, fibers can be directly characterized based on their anatomical correspondence as *anatomically plausible* or *non-plausible*. Ideally, such distinction would allow the complete *filtering* of tractograms from false positives, identified as anatomically non-plausible fibers. However, since the whole white matter structural connectivity is still not known, the definition of a unique plausibility criterion is not possible (Jörgens et al., 2021).

Existing *tractogram filtering* methods encode different plausibility criteria based on the *signal explainability* of fibers and/or the *fiber consistency*. In all cases, they adopt an *unsupervised* approach for filtering. Signal-based filtering, e.g., (Smith et al., 2013, 2015a; Pestilli et al., 2014; Daducci et al., 2015), removes streamlines when their spatial distribution do not match the distribution derived by the amount of underlying diffusion signal. Differently, consistency-based filtering methods cast the filtering task as an outlier detection task. They detect false positive fibers by looking at the excessive deviation of fiber trajectories from the local distributions, which are normally estimated from known bundles, either population-wide (O'Donnell and Westin, 2007; Wang et al., 2018; Xia and Shi, 2020) or intra-individual (Yeh et al., 2019; Legarreta et al., 2021). Lastly, there exist *mixed approaches*, e.g., (Aydogan and Shi, 2015; Neher et al., 2018; Nie and Shi, 2019; Schiavi et al., 2020; Ocampo-Pineda et al., 2021), that try to merge the two criteria by focusing on outlier removal in well-known bundles, while regulating the remaining fibers using the signal prior.

A shared drawback of current approaches is the lack of explicit anatomical priors to encode fiber plausibility. The only attempt to use anatomical constraints is in recent mixed approaches but limited to specific bundles. Without anatomical grounding, unsupervised fiber-based approaches may not guarantee to detect anatomically non-plausible fibers effectively. For example, fibers not explainable by the underlying signal can still represent an anatomically plausible pathway (Smith et al., 2020a; Frigo et al., 2020; Zalesky et al., 2020b). Similarly, fibers considered as outliers might be plausible for the specific anatomy of an individual.

Anatomical prior in supervised learning Many recent studies have confirmed the benefit of using anatomical priors to obtain more accurate tractography representations (Schilling et al., 2020; Girard et al., 2020; Bertò et al., 2021; Schiavi et al., 2020). However, the majority of existing approaches either integrate anatomical priors with *intensional strategies* [1] or consider data-driven strategies. Intensional strategies are based on heuristic rules/constraints that are too general to precisely describe different individuals or use proxy objective functions unsupervised for the anatomy, e.g., signal- and consistency-based approaches. An alternative solution is given by *data-driven* methods like (Wasserthal et al., 2018). Such methods adopt an *extensional* definition of anatomy, where each ROI is assigned with an anatomical label. This allows them to train deep learning models in a *supervised* manner, guaranteeing anatomically sound solutions.

Using DL models for supervised learning allows flexibility in the input data representation. However, depending on the data structure, certain DL architectures might be more suitable than others. In the case of ROI- and fiber-based WM characterizations, the corresponding virtual representations are usually voxel- and streamline-based, respectively. Voxel-based representations, like the volumetric image of an ROI, have a grid structure and thus may be better exploited by CNNs. On the contrary, the polyline structure of streamlines is not a regular grid, and a CNN cannot properly process it due to the varying number of points and the lack of orientation. Hence, existing streamline-based DL approaches based on CNNs (Gupta et al., 2017; Xu et al., 2019; Zhang et al., 2020; Legarreta et al., 2021) resort to heuristics or embeddings to standardize the input streamline representation. However, such standardization is not needed with geometric deep learning models such as (Qi et al., 2017; Wang et al., 2019), as they are designed to handle non-regular structures

[1] The term *intensional* is borrowed from logic, where the meaning is: "specifying the necessary and sufficient properties that an object needs to have in order to be counted as a referent of the term"

with variable size and no parsing order.

1.2 Aim of the thesis

We propose to further improve the characterization of white matter priors to overcome the current limitations of the state of the art. The final goal is to reduce the impact of false positives in tractograms by providing an effective and automatic way of encoding anatomical priors in automatic methods. In particular, we hypothesize that *supervised deep learning* frameworks might produce anatomically sound solutions yet flexible to different types of anatomy characterizations, i.e., ROI-based and fiber-based.

In the context of ROI-based white matter characterization, we propose a novel approach that combines CNN- and rule-based methods. Given the volumetric representation of ROIs, we devise using CNNs to segment those regions plausible for certain bundles, which are not present in existing ROI atlases. CNNs have already proved to be effective in segmenting large ROIs such as bundle masks (Wasserthal et al., 2018), but can they be as effective when targeting small but regular eloquent ROIs? An additional benefit of CNNs would be the segmentation of such ROIs directly in the subject anatomy, avoiding complex (and error-prone) steps of co-registration. Subsequently, the integration of the segmented fiducial regions within bundle segmentation rules can be used to evaluate their impact. Therefore, the question is: will such ROIs improve the segmentation accuracy by reducing the presence of undesired fibers, i.e., *false positives*?

In the case of fiber-based WM characterization, the challenge of proposing a supervised deep learning approach is hard due to the 3D polyline structure of streamlines. Can we successfully apply DL directly on streamlines (and not only on voxels, as in the ROI-based case)? We argue against the use of CNNs as it is currently done in the literature since the grid-targeted inductive bias of CNNs may be suboptimal when considering the unoriented sequential structure of fibers. Alternatively, we propose to investigate the white matter using Geometric Deep Learning. Are GDL models really more suitable for the streamline representation? Can such models learn the anatomical features of streamlines? These questions are yet to be answered in the literature. However, given the existing relation between fiber geometry and anatomical plausibility (Petit et al., 2019), we speculate that the supervised training of a GDL model to distinguish between anatomically plausible and non-plausible fibers might enable the learning of geometrical features underlying fibers' anatomy. The remaining question is: what kind of GDL model can be more effective for artifactual fiber detection?

1.3 Novel contributions

The overall contribution of this thesis is the investigation of how different deep learning models may help in the ROI- and fiber-based characterization of the white matter anatomy.

We present two distinct methods for the two types of characterization, namely StemSeg and Verifyber, respectively.

StemSeg As the first contribution, we present *StemSeg*, a data-driven approach that uses a CNN to individuate fiducial white matter regions that are eloquent for a specific bundle, namely *stem* of the bundle. According to several recent neuroanatomical studies (Hau et al., 2016, 2017; Sarubbo et al., 2013, 2019; De Benedictis et al., 2012, 2014, 2016) the anatomical stem of some association bundles like the Inferior Fronto-Occipital Fascicle (IFOF) can be used to delineate all the pathways composing the bundle. It also emerges that a *stem-based* virtual segmentation, where the anatomical stem is represented by a manually drawn *stem-ROI*, is more accurate compared to standard rule-based segmentation and significantly reduces the presence of false positives fibers. In our StemSeg, we propose to automatize the segmentation of the stem-ROI in order to enable the subsequent stem-based bundle segmentation. We design a two-steps method. First, we train a U-Net (Ronneberger et al., 2015) to segment the *stem-ROI* on a fiber orientation image, like the colored Fractional Anisotropy (CFA) map. Then, we integrate the predicted stem as a waypoint region in a WMQL segmentation rule (Wassermann et al., 2016) that also includes the well-known termination regions derived from a WM atlas (Rolls et al., 2015). We apply the method to the case of the IFOF, where this segmentation rule allows extracting the target (left or right) IFOF from a tractogram.

For the empirical analysis of StemSeg, we considered the open diffusion data of the Human Connectome Project (HCP). After generating the tractogram for more than 150 subjects, we have collaborated with two expert neuroanatomists to obtain a reference set of stem-ROIs and two smaller validation sets (one from each expert) containing the manually segmented IFOF bundles. We make the dataset of stem-ROIs available to the public (De Benedictis et al., 2019). We consider this dataset a valuable contribution to the community, as it is the first large collection of IFOF stem-ROIs.

The quantitative evaluation of StemSeg in the first step of stem-ROI segmentation reveals satisfactory results. The segmented stem-ROIs have the correct shape and are located in correspondence to the IFOF bottleneck. This result validates our hypothesis of adopting a CNN model to segment small fiducial ROIs. The subsequent deployment of the predicted stem-ROIs for the IFOF segmentation produces very accurate bundles. The volumetric overlapping with the reference IFOFs is higher than 93% and robust to the different validation sets. Such results greatly outperform the state-of-the-art TractSeg (Wasserthal et al., 2018), with a ∼30% gap. Hence, the choice of a stem-based approach in contrast to the direct bundle mask segmentation has proved successful. Additionally, we provide qualitative evidence of the StemSeg robustness in clinical quality data with a low magnetic field scanner by segmenting stem-ROIs and IFOFs on patients affected by cancer.

Verifyber As the second and most relevant contribution, we present *Verifyber* a novel Geometric Deep Learning model for tractogram filtering. Verifyber is designed as a supervised classifier to discriminate between anatomically plausible and non-plausible fibers. The intuitive idea is to exploit the increasing effort of the community in proposing anatomical labelings of tractograms, e.g., (Petit et al., 2019; Zhang et al., 2018). As the WM anatomy is not fully known, such labelings are in continuous evolution and propose different elicitation of anatomical knowledge, i.e., follow different criteria of anatomical plausibility. We distinguish between *exclusive* labeling policies like (Petit et al., 2019), which conservatively prioritize the definition of anatomical *non-plausibility* (false positives characterization) and *inclusive* policies like (Zhang et al., 2018), which prioritize the definition *plausible* fibers (well-known bundles). Although these labelings do not share a unified WM characterization, they both suggest a direct relation between the fiber geometry and its plausibility. In our proposal, we consider both these types of labeling, one at a time, to train a novel GDL model conceived for learning anatomical features directly from the fiber representation as a 3D points sequence. The proposed model is the first to be fully invariant to the fiber orientation while being sensitive to the sequentiality of points in the trajectory. Once trained, Verifyber can filter fibers in a new tractogram very fast, e.g., 1M fibers in less than 1 min.

To evaluate Verifyber, we present an extensive set of experiments on HCP diffusion data. In the first group of experiments, we investigate the performance of our GDL architecture compared to other alternative models, like Bidirectional LSTM (bLSTM) (Graves and Schmidhuber, 2005), PointNet (PN) (Qi et al., 2017), and Dynamic Graph CNN (DGCNN) (Wang et al., 2019). We measure the filtering accuracy with 5-fold cross-validation over a set of 20 tractograms labeled with *exclusive* policy (Petit et al., 2019). In the test set, we show filtering accuracy higher than 95% and substantiate the superiority of Verifyber in learning longer and complex fiber trajectories. Then, in the second group of experiments, we test the robustness and adaptability of Verifyber to different labeling, tractography pipelines, and data source. Additionally, we compare our supervised approach with a competing unsupervised approach like FINTA (Legarreta et al., 2021). The overall result is that Verifyber is easily adaptable to evolving labeling, and regardless of the labeling adopted, is robust to different tractography scenarios, which instead is not the case of FINTA. Finally, a last qualitative investigation on clinical patients shows promising filtering results that can be very helpful to support doctors.

1.4 Structure of the thesis

- In Chapter 2 we present the basic principles of the white matter anatomy, and we describe how they can be investigated using current in-vivo imaging techniques (Chapter 2.1.2). Then, in Chapter 2.2 we introduce the fundamentals of deep learning, reserving a deeper explanation to Convolutional Neural Networks and Geometric

Deep Learning, which are particularly of interests for this thesis.

- In Chapter 3 we provide a targeted description of the related works differentiating between ROI-based WM characterization and tractogram filtering.
- Then, in Chapter 4 we highlight current lacks and pitfalls and argue for possible solutions to them in the form of research questions (RQ 1, RQ 2).
- Chapter 5 refers to the *StemSeg* contribution. First, we introduce and present the method; then, we show our empirical analysis. Finally, we discuss the results.
- Chapter 6 presents *Verfyber*. After the introduction and the description of the method, we discuss our methodological choices, positioning Verifyber with respect to alternative models (see Chapter 6.3). Next, we report a large set of experiments distinguishing between *model-related* and *task-related*, which we discuss in the subsequent section.
- In Chapter 7 we draw the conclusions of the thesis, arguing the contribution this work carries to the community and the possible impacts it can have for the future.
- Appendix A reports additional specifications for datasets
- Appendix B contains multiple supplementary investigations, which can help to clarify doubts to the reader.

Chapter 2

Background

This Chapter describes the contextual information necessary to understand the studies presented in this thesis. As this thesis lies in the between neuroscience and computer science, we will present a twofold background: brain white matter anatomy (Section 2.1), and deep learning (Section 2.2)—the specific target of the thesis bounds the level of details to which these areas are presented. Many concepts might be obvious if the reader is an expert in those fields. In that case, this Chapter can be skipped and used only as a reference.

2.1 Brain white matter anatomy

The brain is the most fascinating organ in the human body, responsible for all the actions, feelings, and reasonings a person can do. It is an incredibly complex object, and for this reason, still partially undiscovered. While the main emphasis is on understanding brain functioning, e.g., reasoning and memorization, a mandatory precondition is comprehending the brain's structural connectivity. In particular, understanding brain connectivity concerns the anatomy/structure of brain white matter (WM), which is being investigated in the last decade. The next sections illustrate what we know about the WM and how it is possible to investigate it.

2.1.1 White matter structural connectivity

Neurons are the basis of brain anatomy and activity. They are electrically excitable (spiking) cells communicating with each other through synapses. More precisely, as illustrated in Figure 2.1, a neuron is composed of a cell body (soma) containing the nucleus, surrounded by many short-range connections, dendrites, and extended by a single long-range connection, the axon, which has synaptic terminations. To guarantee a proper circulation of electrical spikes, axons inside the brain are insulated with sheaths of myelin, as shown in Figure 2.1, a white-colored compound consisting of water (∼40%), lipids (∼40-50%), and proteins (∼10-20%).

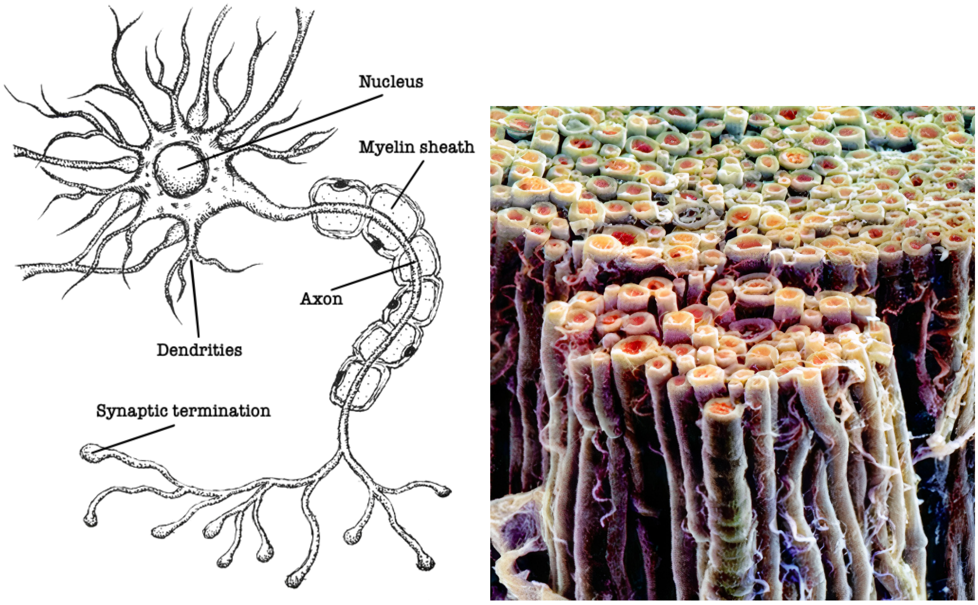


Figure 2.1: Illustration of a neuron cell on the left. SEM photo of myelinated axons. The myelin sheath is colored with whiter color, while the axon is in orange. Image source: Doctor C's On Line Histology, 2015

The disposition of neurons in the brain determines the two main issues of the brain: *grey matter* and *white matter*. The neurons' soma and dendrites are mostly displaced in the external layer of the brain, namely the cortex, and compose the grey matter, while the neurons' axons fill the internal part of the brain, which, due to their white color, is called white matter. A simple illustration of this situation is shown in Figure 2.2, where can also be noted the peculiar shape of the cortex. Such a shape is known as *brain folding*, and it uniquely characterizes each individual. When the cortex folds outward is called *gyrus*, vice versa when it folds inward, it is called *sulcus*. See Figure 2.2 for a schematic example.

Some gyri and sulci are well-known and can be recognized in all individuals (Gray, 1918). A map of the major gyri and sulci is reported in Figure 2.3. From the map, we may recognize that some of the sulci define the border between the different brain lobes: frontal, parietal, temporal, and occipital (colored in Figure 2.3). For example, the central sulcus and the Sylvian fissure separate the frontal lobe from the parietal lobe and the temporal lobe, respectively. More in general, gyri and sulci are used as landmarks by neuroanatomists, who can build different cortex *parcellations* based on them. Such parcellations subdivide the cortex into multiple anatomical regions, usually associated with different brain functions. The different regions of the cortex communicate, like nodes in a graph, thanks to the white matter structural connections between them, i.e., fascicles of neuronal axons.

A white matter structural unit is an axon connecting one neuron to another. The pathway traveled by an axon characterizes its scope. Usually, a certain pathway is traversed

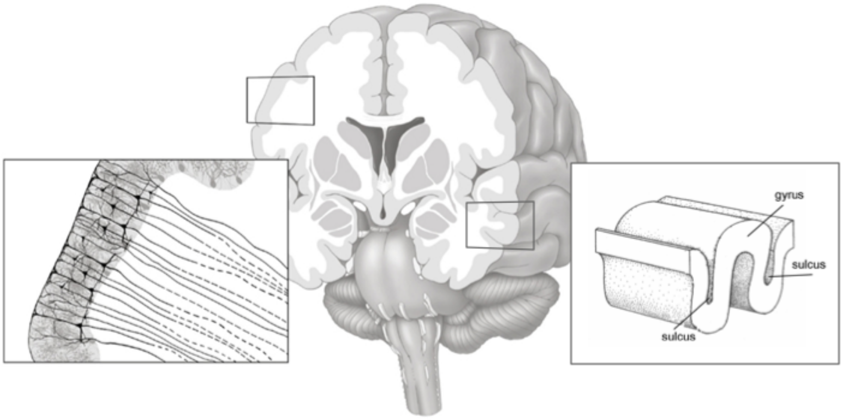


Figure 2.2: Coronal view of the brain. In contrast grey matter and white matter. Left zoom shows the disposition of neuronal cells with soma within the grey matter and axons within the white matter. Right zoom shows a simplistic cortical folding to illustrate the difference between giry and sulci. Adapted from (Bonilha et al., 2015)

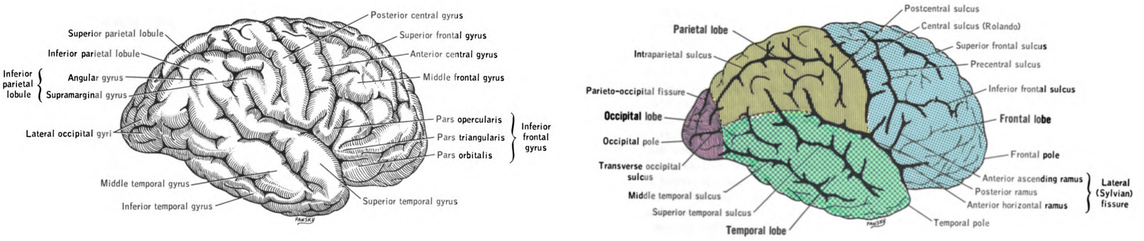


Figure 2.3: Gyri (left) and Sulci (right) based parcellation of the cortex. Colors highlight the different brain lobes: light blue: frontal, yellow: parietal, red: occipital, green: temporal. Adapted from (House and Pansky, 1960)

by a group of axons sharing the same scope. Such groups are called *bundles* of axons and can be categorized based on the regions of the grey matter they connect (Meynert, 1885):

- *Association* bundles connect cortical regions (gyri) belonging to the same hemisphere (see Figure 2.4 left). They have variable shapes and lengths, from short U-shaped bundles connecting adjacent gyri to long straight bundles connecting frontal and occipital lobes. Examples of some of the major association bundles are the Inferior Fronto-Occipital Fasciculus (IFOF), the Arcuate Fasciculus (AF), the Cingulum (CG), and the Uncinate Fasciculus (UF).

- *Projection* bundles connect the spinal cord, the cerebellum, and the deep grey matter regions of the brain stem, with cortical regions. As in the case of association bundles, also projection bundles do not cross the hemispheres. Among the most well-known projection bundles are the Cortico Spinal Tract (CST), cortico-thalamic radiation bundles, and cortico-pontine bundles.

- *Commissural* bundles connect left and right hemispheres. This category of bundles is dominated by the brain's largest bundle, the Corpus Callosum (CC), which binds left and right regions along all the fronto-occipital directions. Due to its wide extensions, CC can be subdivided into multiple components, usually seven: CC1 - CC7, differing from the most frontal to the most occipital location, respectively.

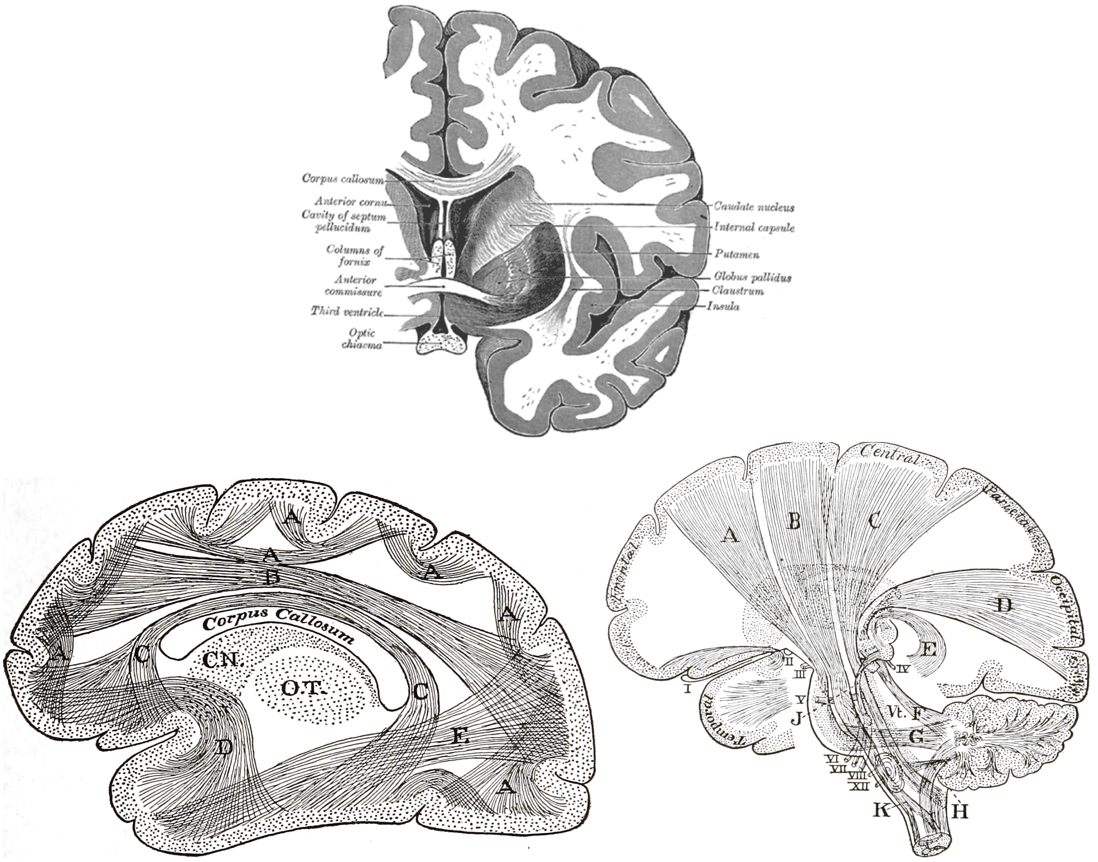


Figure 2.4: Illustrations of association (left), projection (center), and commissural (right) bundles. Adapted from gray (Gray, 1918)

The white matters bundles described above, and sketched in Figure 2.4, can also be identified on real ex-vivo brains thanks to *Klingler dissection* (Klingler et al., 1935; Zemmoura et al., 2016). In the procedure proposed by Klingler et al., the brain is soaked with a solution containing formalin, and then it is dried and cleaned from blood vessels. At this point, the white matter axonal pathways are still not visible. However, after a long freezing period (weeks) followed by fast de-freezing, they become visible due to the formalin solution dilation. The bundles are manually dissected by scraping the brain with rudimental wood utensils. The example of a Klingler dissection is shown in Figure 2.5.

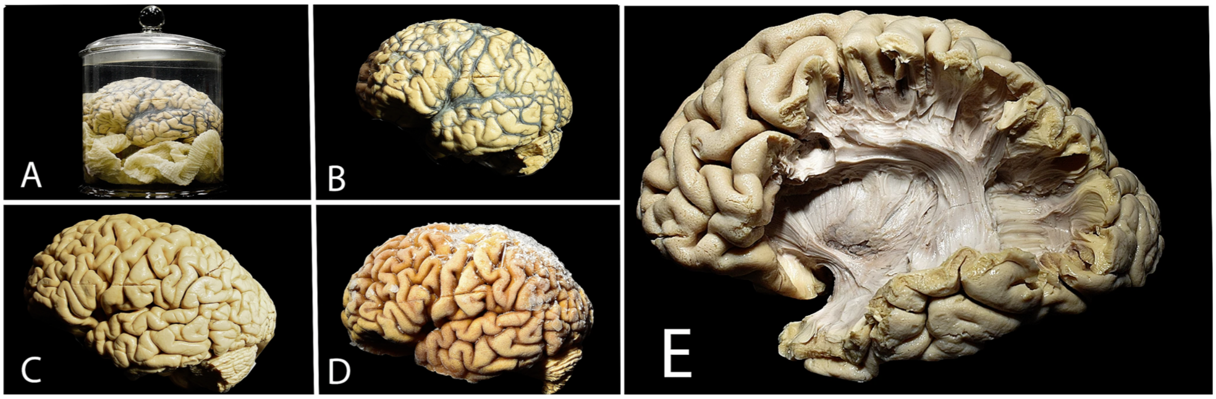


Figure 2.5: Ex-vivo dissection of the IFOF bundle. A → B: formalin. B → C: resection of blood vessels. C → D: refrigerate. D → E: shock de-freeze and dissection. The white matter pathways are now visible. Adapted from (Dziedzic et al., 2021)

2.1.2 Virtual representation of white matter structures

Megnetic Resonance Imaging The in-vivo study of the human brain, in particular of the white matter, requires the use of non-invasive imaging techniques. The last decades have seen the spread of techniques based on megnetic resonance due to a significant hardware development accompanied by an equally important acquisition protocol refinement. Magnetic Resonance Imaging (MRI) exploits a big magnetic field, usually of 1.5T, 3T, or 7T, to highlight different types of contrasts (Bitar et al., 2006) in the form of a volumetric image. The standard brain MRI acquisition, namely *structural* MRI (sMRI), is sensitive to the contrast between grey matter and white matter (and Cortico-Spinal fluid (CSF)). Looking at Figure 2.6 it is easy to notice this contrast, as well as the absence of information about the white matter structure, which is represented with a constant light gray color. Different contrasts may be obtained in sMRI by tweaking the acquisition parameters. For example *T1-weighted* (T1w) images (see Figure 2.6) enhance only the presence of fat in tissues, while *T2-weighted* highlights fat and water presence. In this thesis, we will always refer to T1w as a structural MRI image.

Diffusion MRI To retrieve information about the WM structures we need a *diffusion* MRI (dMRI) protocol (Le Bihan et al., 1986; Basser et al., 1994; Jones, 2010). dMRI uses multiple magnetic field gradients, i.e., oriented variations of the magnetic field, to estimate the diffusivity direction of water molecules that are present in the myelin sheath of white matter axons (Minati and Węglarz, 2007). In this way, dMRI produces a map of the brain white matter, namely Diffusion Weighted Image (DWI), composed of multiple channels, representing the intensity of the water diffusivity in a certain gradient direction. As this type of measure is indirect, i.e., it does not trace diffusion in every single axon, we can only obtain a discretized approximation of the diffusion at the millimetric voxel resolution (Pierpaoli et al., 2001). Specifically, the outcome is a sequence of volumetric

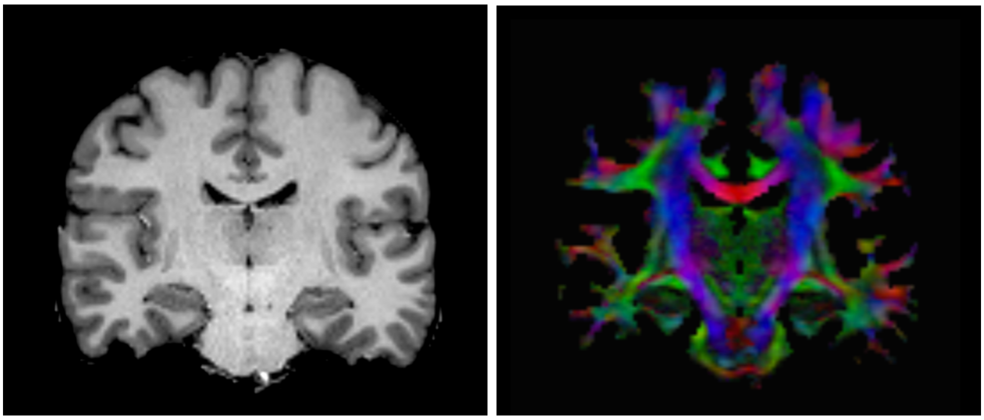


Figure 2.6: Coronal view of structural T1w image (left) and diffusion-based colored FA (right). The T1w highlights the contrast between grey (dark grey) and white matter (light gray). The CFA encodes the local diffusion anisotropy in the three main direction: left-right (red), front-occipital (green), and inferior-superior (blue).

images, also known as diffusivity map, with shape (X, Y, Z, D). Each 3D coordinate (x, y, z) indicates the voxel location, and d indicates the gradient direction. Moreover, gradients are usually acquired at multiple field intensity and timiing, the so-called *b-values*, resulting in a further increasing of the number of diffusivity maps (one for each b-value). Different hardware and protocols may acquire a different number of directions D and b-values, leading to completely different diffusion maps. Also, a diffusion acquisition requires a non-negligible amount of time, from a few to tens of minutes, during which multiple signal distortions like small head movement, field variations, etc., happen (Zhuang et al., 2006; Pierpaoli, 2010).

Luckily, the acquisition protocols and distortion correction pre-processing have improved over the recent years due the introduction of modern software/hardware acceleration techniques, but also to some world- or nation-wide initiatives like the Human Connectome Project (HCP) (Van Essen et al., 2013) and the UK BioBank (Miller et al., 2016). These projects provide public access to MRI data in quantity, thousands of subjects, while guaranteeing high acquisition standards, i.e., providing high-quality data. We will exploit such public resources for the empirical analysis of Chapter 5 and 6.

Model fitting The map of diffusivity within the brain WM is not a ready-to-use object, as the diffusivity information encoded by the several gradients acquired needs to be summarized/aggregated. One way to distillate relevant diffusion information is to fit a mathematical diffusivity model within each voxel of the image. This step is also called *model reconstruction.* In the last years, different kinds of models have been proposed, from the most simple capturing only three directions for each voxel, namely Diffusion Tensor Imaging (DTI) (Basser et al., 1994; Pierpaoli et al., 1996), to more complex ones, namely Higher Angular Resolution Diffusion Imaging HARDI (Descoteaux, 2015), able to capture tens of directions in each voxel, so that a local Orientation Distribution Function (ODF) is generated.

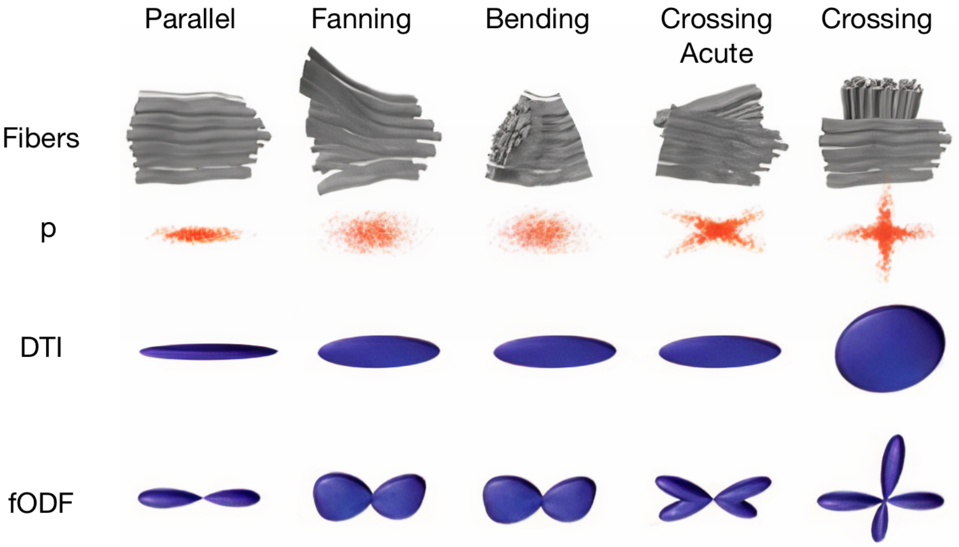


Figure 2.7: DTI versus HARDI modeling. The figure depicts different fiber configurations coupled with the expected pattern p, and the reconstructed models using DTI and HARDI (fODF). Adapted from (Alexander and Seunarine, 2010)

DTI models In DTI, a 3D *tensor* is fitted locally in each voxel, and its three main diffusivity directions are computed as principal eigenvectors, $\epsilon_1, \epsilon_2, \epsilon_3$ and eigenvalues, $\lambda_1, \lambda_2, \lambda_3$, forming an ellipsoid as in Figure 2.7. The same eigenvalues can also be used to compute a Fractional Anisotropy (FA) map,

$$\mathrm{FA} = \sqrt{\frac{(\lambda_1 - \lambda_2)^2 + (\lambda_2 - \lambda_3)^2 + (\lambda_1 - \lambda_3)^2}{2(\lambda_1^2 + \lambda_2^2 + \lambda_3^2)}},$$

which compresses the local diffusivity to a single scalar value close to 0 in case of an isotropic zone (sphere) or to 1 in case of strong anisotropy (thin ellipsoid). The FA map can be enriched by considering the anisotropy for each of the conventional diffusivity directions (x, y, z): right-left, anterior-posterior, and superior-inferior, respectively. By associating the RGB encoding to these components we obtain the *colored FA* map shown in Figure 2.6. In particular,

$$\mathrm{R} = \mathrm{FA}\cos\alpha \quad \mathrm{G} = \mathrm{FA}\cos\beta \quad \mathrm{B} = \mathrm{FA}\cos\gamma$$

, where α, β, γ are the angles between the principal eigenvector ϵ_1 and the x, y, z axes respectively. Unfortunately, in many zones of the white matter, the axonal pathways present complex configurations like crossing or kissing, and the modeling of signal based on 3D tensors cannot capture those configurations (Descoteaux, 2015).

HARDI models The choice of HARDI models mitigates this problem. The fitting of the underlying ODF can be performed using different paradigms e.g., model-based (Tournier et al., 2004, 2007; Dell'Acqua et al., 2007) versus model-free (Tuch, 2004; Wedeen et al., 2005). In particular, we mention the *Constrained Spherical Deconvolution* (Tournier et al., 2004, 2007; Dell'Acqua et al., 2007) model-based approach that produces in output a *fiber* ODF (fODF) modelled with *spherical harmonics* as the ones depicted in Figure 2.7. It is important to remark that even these advanced HARDI models are not a definitive solution to the ambiguity of the diffusion signal in certain regions of the white matter. Indeed, the task of reconstructing diffusion signal is ill-posed, as the same signal might represent different axonal pathways configurations (Mangin et al., 2013; Jeurissen et al., 2017).

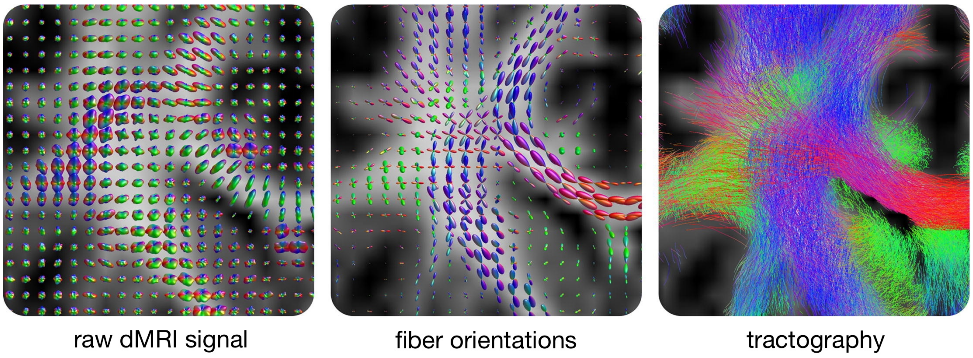


Figure 2.8: Summary of tractography pipeline. The first step is diffusion MRI signal acquisition. The second step is model reconstruction — in this case performed with CSD. The final step is the fiber tracking. Image taken from (Tournier, 2019)

Tractography A subsequent step to obtain the virtual representation of white matter axonal pathways is to estimate them from the diffusivity model. In this step, referred as *tracking* or *tractography* (Basser et al., 1994; Mori et al., 1999; Conturo et al., 1999; Basser et al., 2000), we can move from a voxel-based representation to a pathway-based representation (see Figure 2.8). Tractography algorithms track a pathway by (i) *seeding* the starting point in a certain 3D coordinate within a WM/GM region, (ii) estimating the next point at a predefined distance called *step* based on the underlying diffusivity orientations, and (iii) iterating the next-step estimate until a *stopping condition* is satisfied, e.g., gray matter is reached. The next-step decision is subject to multiple tracking *hyperparameters* such as the *maximum angular difference* between the previous and the next point, and the length of the *step-size*. The tuning of these hyperparameters leads to the generation of different pathways. Their values must be selected in order to reduce the generation of anatomically implausible pathways, usually referred to as *false positive*, while maintaining a comprehensive coverage of the well-known pathways. Moreover, different heuristics of seeding and stopping, and tracking paradigms affect the tractography output.

Based on the heuristics of seeding, stepping, and stopping, we can differentiate existing tractography algorithms. A usual distinction distinguishes among the following dichotomies:

- *Deterministic vs. probabilistic.* In *deterministic* algorithms, the next-step decision is performed by always choosing the principal diffusion direction. This means that if we seed two times in the same 3D point, we obtain two equal pathways. When the underlying diffusion model encodes the basic three directions, e.g., DTI, deterministic algorithms can track only the easiest pathways (Mori et al., 1999; Basser et al., 2000), i.e., pathways encoded by the principal directions of voxels, while in the case of an underlying HARDI model, the algorithms can also track some of the trajectories passing through confusion regions, e.g., crossing (Descoteaux et al., 2009; Tournier et al., 2012). However, given the intrinsic ambiguity of the diffusion signal, choosing the same next step given similar local conditions may lead to missing many existing pathways. For this reason, *probabilistic* algorithms choose the next step by randomly sampling from a pool of possible directions that are within the range of the maximum angular difference (Descoteaux et al., 2009; Tournier et al., 2010, 2012). As a result, such algorithms produce a richer set of pathways, including some *false positives* due to the stochastic procedure combined with the underlying signal noise.

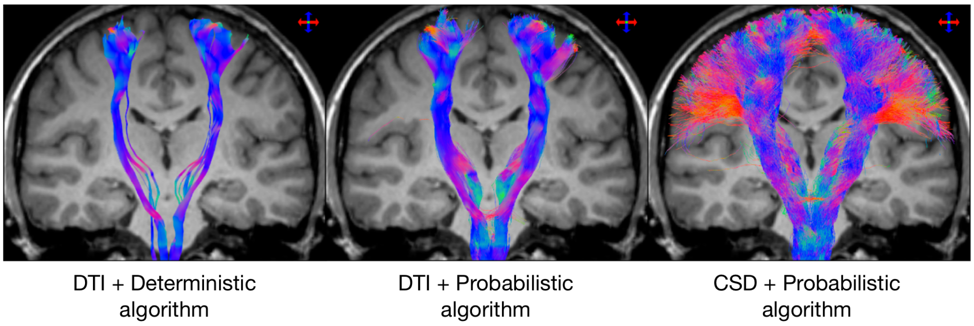

DTI + Deterministic algorithm
DTI + Probabilistic algorithm
CSD + Probabilistic algorithm

Figure 2.9: Example of deterministic and probabilistic tracking of the Cortico-Spinal Tract (CST). Image taken from Calamante (2019)

- *Local vs. global.* This dichotomy opposes algorithms that decide the next tracking step based only on the current situation, like in a Markov chain, to algorithms that consider the impact of the next step for the entire pathway tracked until the current point or to the neighbor pathways already tracked (Jbabdi et al., 2007; Smith et al., 2012; Girard et al., 2014; Lemkaddem et al., 2014; Dong et al., 2017b; Aydogan and Shi, 2018). The simplistic approach of *local* algorithms is sensitive to local noise present in the diffusivity model, which then may lead to tracking *false positive* pathways that do not respect the underlying anatomy of the white matter. On the other hand, *Global* approaches try to overcome the limitations of local methods but are not able to solve the issue of false positives. The reason is that also global approaches inherit the limits of locally reconstructed diffusion models.

- *Whole-brain vs. bundle-specific.* The tractography procedure does not impose to track pathways covering the entire volume of the white matter. Instead, one might be interested in tracking only a portion of it, or a specific bundle (Yendiki et al., 2011; Thomas et al., 2014; Chamberland et al., 2017; Reisert et al., 2018; Rheault et al., 2019; Wasserthal et al., 2019; Maffei et al., 2021). In this case, the seeding and stopping criteria can be delimited to certain white matter regions according to the target bundle. Additional bundle-specific constraints concerning fiducial waypoints, the bundle mask, or regions of avoidance can limit even more the tracking. Such restrictive constraints cannot be applied in the case of whole-brain tractography, where, however, one might adopt global constraints. For example, a common global constraint imposes the seeding and stopping only in the interface between white and grey matter, where we expect neurons' axons to start or end. Another possibility is to pre-determine the different tissues in the brain and use the tissue map to overall constraint the tracking as in Anatomically-constrained tractography (ACT) (Smith et al., 2012). In terms of computational costs, focusing only on specific white matter bundles consistently reduces the total amount of computation needed, and consequently, the duration of the process. Instead, whole-brain tractography is rather time-consuming.

Tractogram, fibers, bundles The final result of a whole-brain tractography is a *tractogram* (see Figure 2.10), an explicit representation of the brain connectivity structure within the white matter. It is composed of a collection of virtual pathways, usually hundreds of thousands or millions, representing the main axonal pathways. Such virtual pathways are represented as 3D *polylines* by sequences of 3D points, also called *fibers* or *streamlines*. However, in this thesis, we will indistinctly call them in one or in the other way. The estimated *fibers* are characterized by different trajectories and lengths. In a typical brain, the lengths range between 2cm of U-shaped fibers connecting adjacent gyri to 20/25cm of fibers connecting the external border of anterior and occipital lobes. The trajectory of a fiber, which reflects with a certain approximation several real axonal pathways, define its anatomical meaning. Fibers having similar anatomical meanings constitute virtual *bundles*, which are the virtual homologous of white matter bundles described in the previous section. An explicative visualization of the analogy between real and virtual bundles is shown in Figure 2.11.

Brain co-registration Another concept we need to introduce is the co-registration between brain images. Indeed, different subjects and, more in general different MR acquisitions usually lie in different *spaces*. The result is that the different brain images are not aligned with each other. The mismatch can be stronger when the two brain images come from different subjects or smaller when the two images are two separate acquisitions of the same brain. To compare different brain images, we need to co-register them in a common

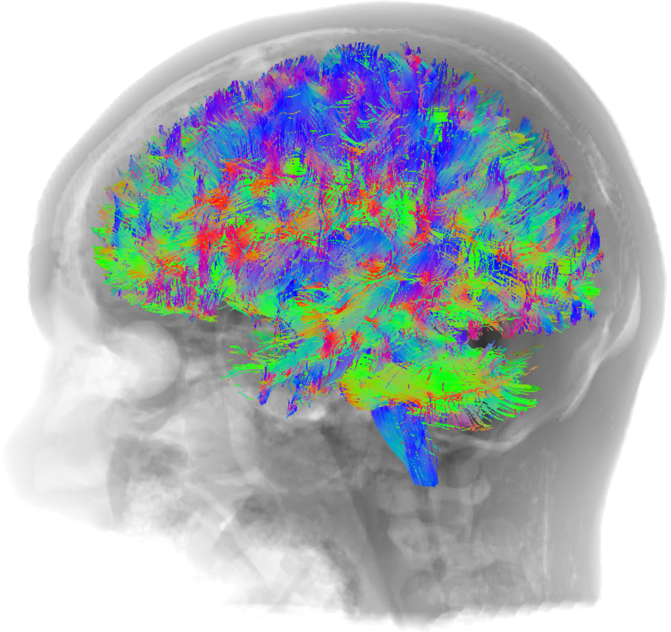


Figure 2.10: Whole-brain tractogram. Virtual representation of the white matter structural connectivity composed of millions of fibers. Fibers are colored according to the local orientation: red for left-right, green for front-occipital, and blue for inferior-superior

space. This operation can be be done through *rigid*, *linear*, or *non-linear* warping. Given a *target* or *reference* brain image and a *moving* brain image, an optimization algorithm computes the desired transformation that maps the *moving* to the *reference*. In the case of rigid and linear transformations, the mapping is global and thus is represented by an affine matrix. The difference between the two is that rigid involves only translation, rotation, and reflection, while linear can also modify an image's scale and shear. More advanced, even though more complex to be optimized, are non-linear warps as they compute a transformation, namely *displacement field*, which acts differently on each voxel. Example of existing tools to compute image co-registration are Advanced Normalization Tools (ANTs) (Avants et al., 2008, 2009) and FMRIB Software Library (FSL) (Jenkinson et al., 2012).

Tractogram registration Similar to what happens with brain images, one may need to normalize tractograms or bundles of different subjects to the same space, e.g., standard brains. In this case there are two registration approaches: *image-based* and *streamlines-*

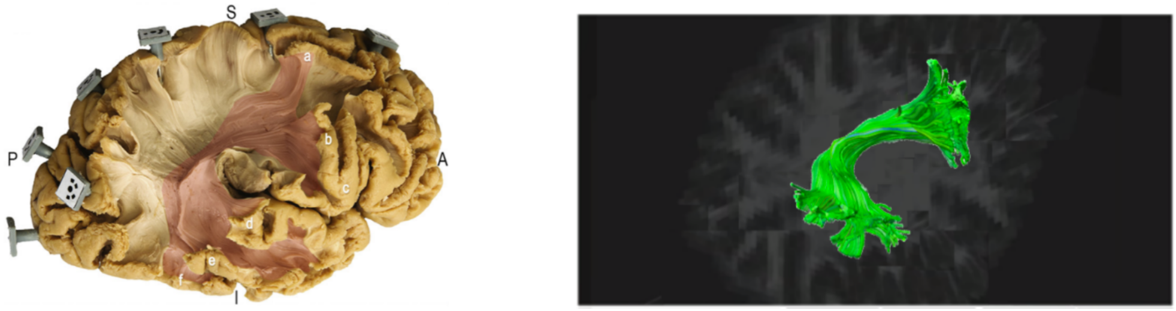


Figure 2.11: Real (left) and virtual (right) depiction of AF. Ex-vivo photo taken from (De Benedictis et al., 2018).

based. In image-based, the co-registration is computed, as usual, at the voxel level using brain images and then applied to the points composing each fiber. In *streamlines-based*, the fibers of the two subjects are used to compute either a linear transformation via overall fiber alignment (O'Donnell et al., 2012; Garyfallidis et al., 2015) or a non-linear transformation via multiple local bundle alignments (Wassermann et al., 2011; Olivetti et al., 2016; Sharmin et al., 2016; Olivetti et al., 2020). While computationally more intensive, non-linear streamline-based methods have been demonstrated to be more accurate compared to non-linear image-based methods (Olivetti et al., 2020).

Cortical atlases Co-registering brains to a common space enable the definition of cortical atlases as the average of several cortical parcellations of different subjects. Neuroanatomists draw such parcellations by virtually encoding the anatomical knowledge concerning the main sulci and gyri via volumetric parcels of structural images (see Figure 2.3). Several atlases of cortical parcellations have been proposed. Basic parcellations only regards gyri and sulci (Tzourio-Mazoyer et al., 2002; Desikan et al., 2006), while more advanced atlases either refine such parcellations into more fine-grained parcels (Destrieux et al., 2010; Rolls et al., 2015) or integrate multi-modal information such as functional signal (Glasser et al., 2016; Eickhoff et al., 2018). Typically, such atlases are defined on template brains, obtained from the averaging of a population. One example is the widely adopted standard of the Montreal Neurosciences Institute - International Consortium for Brain Mapping ((MNI-ICBM-152) (Fonov et al., 2011), for which it has been computed T1w, T2w, and Proton Density (PD) contrasts starting from 152 subjects.

Bundle atlases / White matter atlases The possibility to investigate in-vivo white matter structures through tractography has also led to the creation of a standard collection of virtual bundles, namely *bundle atlases.* Early attempts were made by (Catani and Thiebaut de Schotten, 2012; Wakana et al., 2007; Oishi et al., 2008; Mori et al., 2005) using basic DTI tracking, while more recently, new atlases based on advanced tracking pipeline and large population of subjects were presented (Yeh et al., 2018; Zhang et al., 2018). In (Yeh et al., 2018) the averaging is made at the level of DWI, which are non-linearly co-registered to same standard space, MNI-ICBM-152 (Fonov et al., 2011), during the process of model reconstruction (Yeh and Tseng, 2011). Differently, in (Zhang et al., 2018) the averaging is performed once the tractogram is computed for each of the considered (100) subjects. In this case, the co-registration is only linear, and it is computed considering the tractograms rather than volumetric images (O'Donnell et al., 2012). The two different computation pipelines have led to visually different atlases: in (Yeh et al., 2018) bundles are smoother and less voluminous, while in (Zhang et al., 2018) intra-bundle trajectory variance is higher and the fanning more spread.

2.2 Deep learning

In the context of this thesis, Deep Learning (DL) (LeCun et al., 2015) represents a core component, and this section summarizes some of the foundations needed to understand our research choices. In particular, we first describe the basic principles and assumptions that are behind the success of Convolutional Neural Networks (CNNs) (LeCun and Bengio, 1995) (Section 2.2.1), in order to introduce by analogies and contrasts Geometric Deep Learning (GDL) (Masci et al., 2016; Bronstein et al., 2017, 2021), a recent family of methods which extends CNNs to new domains different from images.

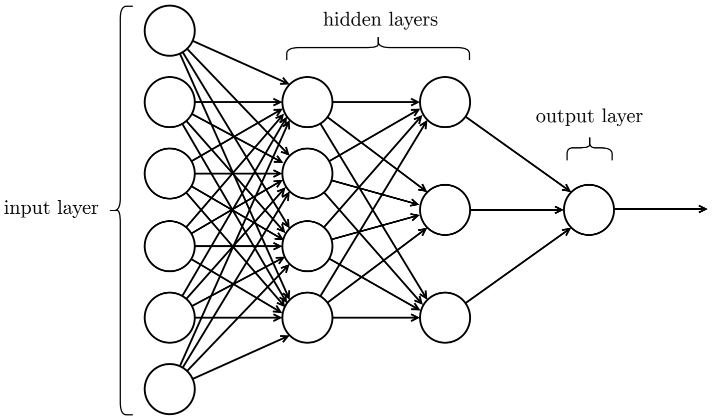


Figure 2.12: MultiLayer Perceptron (MLP) (Rosenblatt, 1961) architecture.

Deep Learning is a family of machine learning models that relies on artificial neural networks (ANNs or NNs) (LeCun et al., 2015). In the last decade, DL has established the state of the art for several applications such as computer vision (Srivastava et al., 2015; He et al., 2016; Ronneberger et al., 2015), machine translation (Bahdanau et al., 2015), and natural language processing (Devlin et al., 2019; Brown et al., 2020). However, the origins of NNs date back to the previous century when McCulloch and Pitts (McCulloch and Pitts, 1943) proposed the artificial neuron model, which was then used as a building block by Rosenblatt (Rosenblatt, 1958) to define the first single-layer neural network, namely *Perceptron*. The number of neurons in a NN layer defines its width, while the number of layers defines its depth. The stacking of several *Perceptron* layers in depth constitutes an example of a *deep* NN known as multi-layer Perceptron (MLP) (Rosenblatt, 1961).

Multilayer Perceptron. The Multilayer Perceptron model (Rosenblatt, 1961) is the simplest kind of deep NN architecture. It is composed of an *input* layer, several intermediate layers called *hidden* layers, and an *output* layer, see Figure 2.12. All these layers have the same structure, typically referred as *fully-connected* or *dense*, which involves a parametrized linear function $f(\cdot)$ followed by a non-linear activation $\sigma(\cdot)$. We may formulate a layer at depth $l \in \{1, ..., L\}$ as follows:

$$\begin{aligned} f^{(l)}(\mathbf{z}_{l-1}) &= \mathbf{W}_l \mathbf{z}_{l-1} + \mathbf{b}_l \\ \mathbf{z}_l &= \sigma(f^{(l)}(\mathbf{z}_{l-1})), \end{aligned} \tag{2.1}$$

where $\mathbf{z}_{l-1}$ is the output vector of the previous layer while $\mathbf{W}$ and $\mathbf{b}$ are learnable parameters with $\mathbf{W}$ being a matrix of weights and $\mathbf{b}$ the bias vector. For the sake of simplicity we can indicate a layer as a single function $h_{\boldsymbol{\theta}}^{(l)}$, parametrized by the set of learnable parameters $\boldsymbol{\theta}^{(l)}$. It follows the MLP formulation:

$$\mathbf{y} = h_{\boldsymbol{\theta}}^{(L)} \circ h_{\boldsymbol{\theta}}^{(L-1)} \circ \cdots \circ h_{\boldsymbol{\theta}}^{(1)} \tag{2.2}$$

The training of a NN aims to learn the set of parameters that best approximate a given data distribution. In practice, this is formulated as an optimization problem, where the parameters are adjusted by iteratively minimizing a differentiable error function, also known as the *loss* function. The training procedure is based on the *Hebbian learning rule* (Hebb, 1949), which consists of an iterative update of the network parameters based on the minimization of the current prediction error. The direction or *gradient* toward which each parameter must be updated can be computed with the partial derivative of the error with respect to the parameter itself. The computation of the derivatives is done in the *backward* direction: starting from the output layer until the input layer, in a procedure called *error backpropagation* (Werbos, 1988; Rumelhart et al., 1986). Once the gradients are computed, an *optimization* algorithm, usually based on Stochastic Gradient Descent (SGD) (Robbins and Monro, 1951), is in charge of computing the updated set of parameters also considering a scaling factor, namely *learning rate*, which regulates the amount of update to be done at each training iteration. We may formulate the learning rule as follows:

$$\begin{aligned} \mathbf{W}_{t+1} &= \mathbf{W}_t + \Delta \mathbf{W} \\ \Delta \mathbf{W} &= -\eta \frac{\partial \mathcal{L}}{\partial \mathbf{W}} \end{aligned} \tag{2.3}$$

where, $\mathbf{W}$ is the weight matrix, $\mathcal{L}$ is the loss function, and η is the learning rate.

From the definition of the *loss* function, we can have different training settings. The two most usual are *supervised* and *unsupervised.* In the first, the loss exploits external supervision, i.e., data annotations or *labels*, to quantify the correctness of a prediction. In the second setting, instead, there is no availability of labeling the data. Thus the loss function must encode some kind of proxy tasks where there is no need of labels, e.g., input reconstruction.

Representation learning The success of NNs resides in their ability to learn hierarchical representations directly from data. In contrast to traditional machine learning models, which need to pre-define a set of features representing the input data, NNs can learn new features guided by the optimization of the task at hand. The intuitive idea is each layer learns a mapping that project its input data into a new *latent space*, where data are distributed according to a new set of features more convenient for the objective task. The stacking of multiple layers allows the learning of hierarchical representations.

While representation learning is the most powerful aspect of NNs, the generation of meaningful representations highly depends on the architectural choices, e.g., the type and disposition of layers. Indeed, the architecture of a NN conveys an *inductive bias* for the type of data regularities that the NN can learn. Different data structures present different regularities and thus require different NN architectures to be properly processed. For this reason a large variety of NNs is being proposed in the literature (Schmidhuber, 2015), which may be categorized using a few fundamental class of architectures: *Recurrent NNs* (RNNs), *Convolutional NNs* (CNNs), *PointNets*, *Graph NNs* (GNNs), and *Transformers.* Among these, we set our attention on CNNs, PointNets, and GNNs in the next paragraphs.

2.2.1 Convolutional neural networks

The rise of Convolutional Neural Networks (CNNs) has origins to emulate human vision. Fukushima (Fukushima, 1980) and later Lecun et al. (LeCun et al., 1989) proposed the firsts pioneer CNNs architecture highly inspired by the vision circuit model of the human brain proposed by (Hubel and Wiesel, 1959). While not effective at that time, these works provided groundbreaking intuitions that are now at the base of every modern CNN architecture. The idea is to exploit the grid structure underlying the images to generate a hierarchical sequence of latent representations that capture certain visual patterns independently of where these are located in the image grid. Learning new representations at different grid resolutions allows a CNN to individuate both simple and complex patterns like corners and faces, respectively. These properties led to a breakthrough in computer vision when in 2012 Krizhevsky et al. won the ImageNet classification challenge. Since then, there has been continuous advancement by introducing milestone architectures to tackle different computer vision tasks. Residual networks (Srivastava et al., 2015; He et al., 2016) has become the standard for image classification, as well as encoder-decoder architecture like U-Net (Ronneberger et al., 2015) set the state-of-the-art for many image

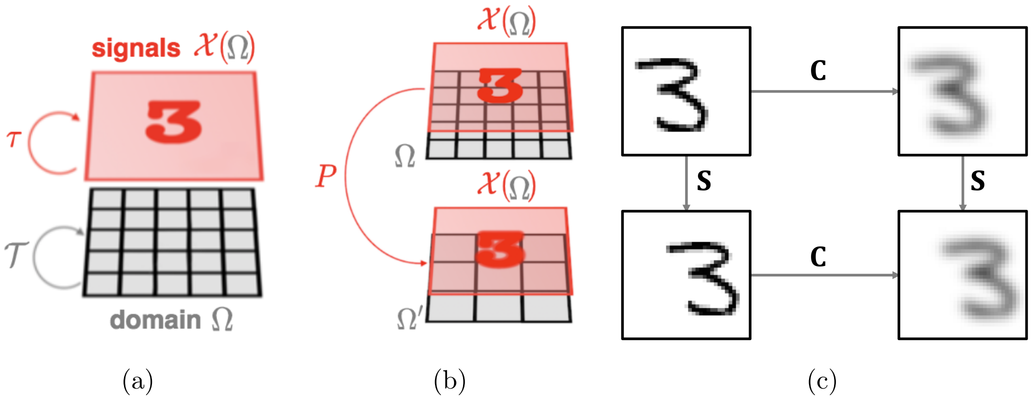


Figure 2.13: (Symmetries of the grid domain. (a) relation between the grid domain, the image signal, and the symmetries. (b) Example of pooling (P) to coarsen the grid. (c) Translation equivariance of the image convolution (C), S stands for shift. Adapted from (Bronstein et al., 2021)

segmentation tasks including neuroimaging applications (Dong et al., 2017a; Wasserthal et al., 2018).

A better explanation of the success of CNNs may be obtained by adopting a geometric perspective as proposed by (Bronstein et al., 2021). In this geometrical view, given a *domain* characterized by a certain structure like a *grid* and a class of *signals* defined on it as images, we want to explicit which are the *geometric priors* that a NN must encode as part of its inductive bias in order to learn a task on these signals. In particular, we describe three types of geometric priors playing a key role in existing NNs: *symmetries*, *local deformations*, and *scale separation*.

Considering the specific case of CNNs, which excels on task based on the image signal, we may refer, as domain, to the *Euclidean* plane discretized with a 2D grid $\Omega = \mathbb{Z}_H \times \mathbb{Z}_W$, and, as signal, to the grayscale images defined as functions on that grid: $x \in \mathcal{X} : \Omega \to \mathbb{R}$. Moreover, we can formalize the usual image-based tasks, e.g., image classification and segmentation, defining a generic target function $y : \mathcal{X}(\Omega) \to \mathcal{Y}$ that we aim to approximate with our NN.

Symmetries With this formulation, we may identify a group of *symmetries* $\mathcal{T}(\Omega)$ that characterizes the grid domain Ω with respect to y. We may define a *symmetry* of an object or a system as a transformation $\tau \in \mathcal{T}$ that preserves a certain property of it. Symmetries are always invertible transformations, and their composition generates a new symmetry. The symmetries of a domain are also inherited by the function defined on the domain itself, $\mathcal{X}(\Omega)$ in this case, and thus impose a geometric structure on the target function y. However, different target functions, i.e., tasks, may be affected differently by the domain symmetries. In some tasks the output of y is *invariant* with respect to $\tau \in \mathcal{T}(\Omega)$:

$$y(\tau(x)) = y(x) \tag{2.4}$$

Alternatively, there are other tasks in which the output of y has the same structure of the input, and thus must be affected by τ in the same way of the input:

$$y(\tau(x)) = \tau(y(x)) \tag{2.5}$$

In this case y is said to be *equivariant* with respect to $\tau \in \mathcal{T}$.

Stationarity Considering the 2D grid domain Ω, the group of *translations* are symmetries for its structure, and are applied via the *shift operator* to images $x \in \mathcal{X}(\Omega)$: $\tau_s(x) = x - s$, where $s \in \Omega$ is a 2D vector. See Figure 2.13. In the case of image classification the class of an image is invariant to shifting, while in case of segmentation the output mask must shift equally i.e. equivariant to the input. Hence, to learn efficiently such tasks we search for a NN that is *translation invariant* or *equivariant* depending on the task.

Local deformations and scale separation While symmetries are *global* transformations to which we aim to provide an *exact* invariance or equivariance, in practice it often happens that there are *local* symmetries, namely *deformations*, to which we can only provide an *approximate* invariance. Local deformation fields ℓ can alter the signal at different degrees of entity, from imperceptible deformations to highly visible distortions. We can indicate the distortion as the smoothness of the deformation $\nabla\ell$. Considering invariance (2.4) and equivariance (2.5) equations respectively, we obtain the following:

$$||y(\ell(x)) - y(x)|| \approx ||\nabla\ell|| \tag{2.6}$$

$$||y(\ell(x)) - y(x)|| \leq ||\nabla\ell||||x|| \tag{2.7}$$

When one of these approximate equality or inequality is verified, we say the y is *locally stable*. Such a property is crucial if we consider as local deformation the *coarsening* ℓ_p of an image, which aggregates adjacent pixels. In fact, thanks to (2.6) we can state that a hierarchy of grid resolutions like $\Omega, \Omega', \Omega'', \ldots$, maintains at each scale the important information of x, see Figure 2.13. Also, assuming the local stability of y, we can factorize it using different resolutions:

$$y \approx y' \circ \ell'_p,\ \ell'_p : \mathcal{X}(\Omega) \rightarrow \mathcal{X}(\Omega'), \quad y' : \mathcal{X}(\Omega') \rightarrow \mathcal{Y} \tag{2.8}$$

In practice, this means that we can separate across scales long-range interactions between features that model complex objects like faces.

Grid convolution Let us define a discrete convolutional filter in case of a 2D grayscale image $\mathbf{x} = x(\Omega)$ as a small matrix $c(\boldsymbol{\theta})$ of learnable parameters having size (H^c, W^c). The convolution on $\mathbf{x}$ is then referred as $c(\boldsymbol{\theta})\mathbf{x} = \mathbf{x} \star \boldsymbol{\theta}$. For each point of the image, (u_1, u_2), the convolution operator first computes an element-wise multiplication between the filter centered on that point and the point neighborhood, then it reduce the obtained matrix via summation (see Figure 2.17):

$$(\mathbf{x} \star \boldsymbol{\theta})_{ij} = \sum_{i=0}^{H^c-1} \sum_{j=0}^{W^c-1} \boldsymbol{\theta}_{ij} \mathbf{x}_{u_1+i, u_2+j} \tag{2.9}$$

We may observe that the choice of a small convolutional filter is in agreement with the local deformation prior. Indeed, according to (2.9) and (2.7), the convolution operation is *equivariant* with respect to local deformations including local translations, and thus *locally stable*. Also, as the convolutional filter is slid along all the coordinates $(u_1, u_2) \in \mathbf{x}$ following a sequential arbitrary order, we derive other two desired properties. The first is that the sliding filter becomes equivariant to global shift i.e. to image translations, while the second is that the sliding allows the use of a fixed (small) amount of parameters regardless of the size of the image.

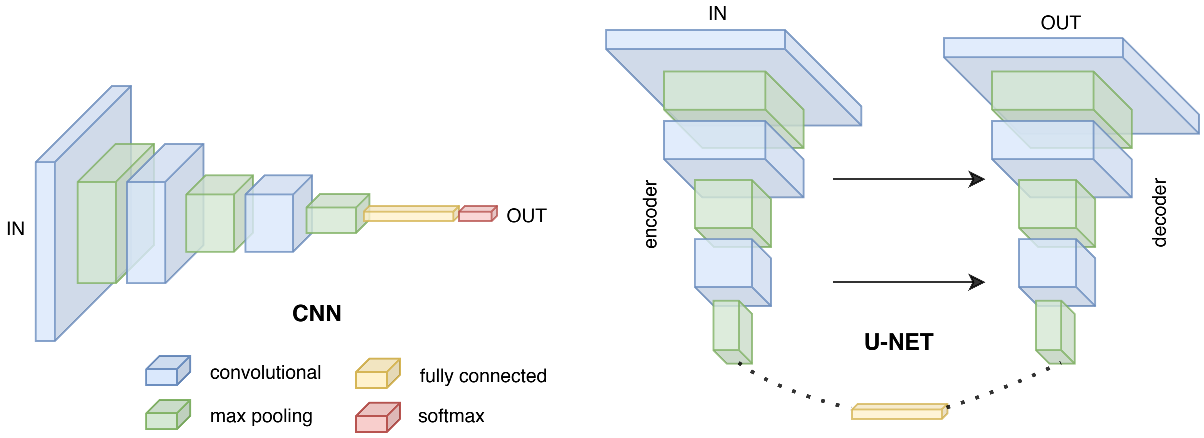


Figure 2.14: Schematic example of a CNN architecture (left) and U-Net (Ronneberger et al., 2015) (right).

CNN architecture A *convolutional layer*, $g(\cdot)$, is composed of a convolution $c(\boldsymbol{\theta})$ followed by a non-linear activation $\sigma(\cdot)$:

$$g_{\boldsymbol{\theta}}^{(l)}(\mathbf{x}) = \sigma(c(\boldsymbol{\theta})\mathbf{x}) \tag{2.10}$$

In a typical CNN architecture (see Figure 2.14(left)) for classification, several convolutional layers are stacked in-depth to compose the *encoder* part of the network. However, while these layers guarantee *local stability* to the encoder, they struggle to capture long-range dependencies. For this reason, convolutional layers are usually interleaved with *local pooling* layers $p(\cdot)$, which perform coarsening using aggregators like *max* or *mean.* In this way, the CNN may also exploit the *scale separation* prior. Then, to produce a single latent feature vector of the input, a CNN may either flatten the feature maps or employ a *global pooling* layer p_G that performs a single reduction again using max or mean. Lastly, the CNN uses a *classification head*, typically an MLP $h_{\boldsymbol{\theta}}$, which takes as input the single descriptor generated by the *encoder* and outputs a classification prediction. We may formulate the whole architecture as follows:

$$\begin{aligned} h_{\boldsymbol{\theta}}^{enc} &= p_G^{(L-1)} \circ g_{\boldsymbol{\theta}}^{(L-1)} \circ p^{(L-2)} \circ g_{\boldsymbol{\theta}}^{(L-2)} \circ \cdots \circ p^{(1)} \circ g_{\boldsymbol{\theta}}^{(1)} \\ \mathbf{y} &= h_{\boldsymbol{\theta}}^{(L)} \circ h_{\boldsymbol{\theta}}^{enc} \end{aligned} \tag{2.11}$$

A schematic example of a CNN architecture is shown in Figure 2.14(left).

U-Net A very common CNN architecture for image segmentation to which we will refer later in this thesis is U-Net (Ronneberger et al., 2015). The "U" in the name indicates the peculiar shape of this architecture, shown in Figure 2.14 (right). Differently from classification networks, in segmentation architectures, the encoder is followed by a specular network, namely *decoder*, that instead of encoding and downsampling, decodes (via convolution) and upsamples until it generates an output having the same size as the input image. In the case of binary segmentation, the output image is a mask having 1s in the pixel where the target object is expected to be and 0s elsewhere. In the U-Net architecture, the *decoder* layers are conditioned directly, through *skip connections*, by the layer of the encoder that is specular in the "U". The idea behind the skip connections is that features extracted while encoding/downsampling the image are richer than features extracted while decoding/upsampling it. Hence, integrating the two different features map done via summation generates higher-quality feature maps. Interestingly, skip connections have been demonstrated to be particularly important for medical imaging applications (Drozdzal et al., 2016).

2.2.2 Geometric deep learning

We have shown that the geometric prior on the *Euclidean* grid domain allows CNN to work exceptionally well on a grid-based signal. However, in the real world, there are many other domains and signals not characterized by a grid, but for which we would like to extend CNN-like models. We refer to the family of models that does not rely on a Euclidean grid domain as *Geometric Deep Learning* (GDL) (Masci et al., 2016; Bronstein et al., 2017, 2021).

Graphs and point clouds In particular, for the aim of this thesis, we are interested in introducing two *non-Euclidean* domains: *graphs* and *point clouds.* A graph $\mathcal{G}$ is a structure

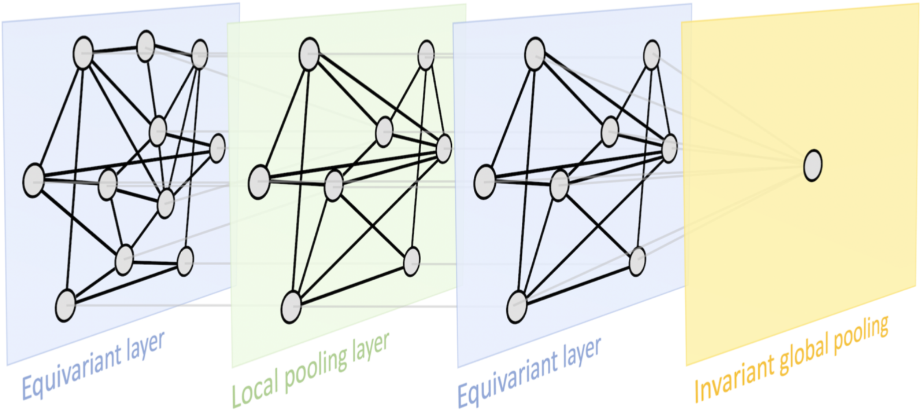


Figure 2.15: Typical architecture of a GDL neural network. The figure illustrate a graph neural network where the input graph is embedded through an equivariant layer, then locally aggregated with a permutation invariant pooling, and again embedded with a second graph convolution layer. Finally, applying global max pooling generates a single descriptor for the input graph, which can be classified via a standard MLP. Image taken from (Bronstein et al., 2021)

where nodes $\mathcal{V}$ are connected each other with edges $\mathcal{E}$: $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. A point cloud then may be seen as a special graph without edges, $\mathcal{E} = \emptyset$, also referred to as *set*. Graphs and point clouds are highly flexible structures, and indeed they may be used to model different real-world scenarios. Based on the scenario, the nodes can have a completely different semantic such as persons in social graphs, regions in brain connectome, or 3D points in meshes. Thus we can generically indicate them as d-dimensional vectors $\mathbf{x} \in \mathcal{V}$. Following the same notation used for grid we obtain: $\Omega = \mathcal{G}$, and $\mathcal{X} = (\mathcal{G}, \mathbb{R}^d)$.

Cardinality and graph isomorphism Both graphs and point clouds domain when considered in the usual tasks of object classification or segmentation (node classification) are characterized by the same group of symmetry, the *permutation* group. This type of symmetry preserves the *cardinality* $|\mathcal{X}|$ as well as the connectivity of a graph i.e,. produces a new graph *isomorphic* to the input one. The permutation group applies on $\mathcal{X}$ by permuting in all possible orderings the points/nodes in it. If we indicate the permutation transformation as τ_p, we can define *permutation invariance* and *equivariance* as in (2.4) and (2.5) respectively, by simply replacing τ with τ_p.

GDL architectures To have CNN-like architectures also for graphs and point clouds, we may recall the prototypical architecture of CNN, which alternate parametrized equivariant (convolutional) layers with local or global poolings. We want to individuate similar types of layers to compose GDL architectures, see Figure 2.15. However, the grid convolution as in (2.9) requires locality and stationarity symmetries that are not present by constructions

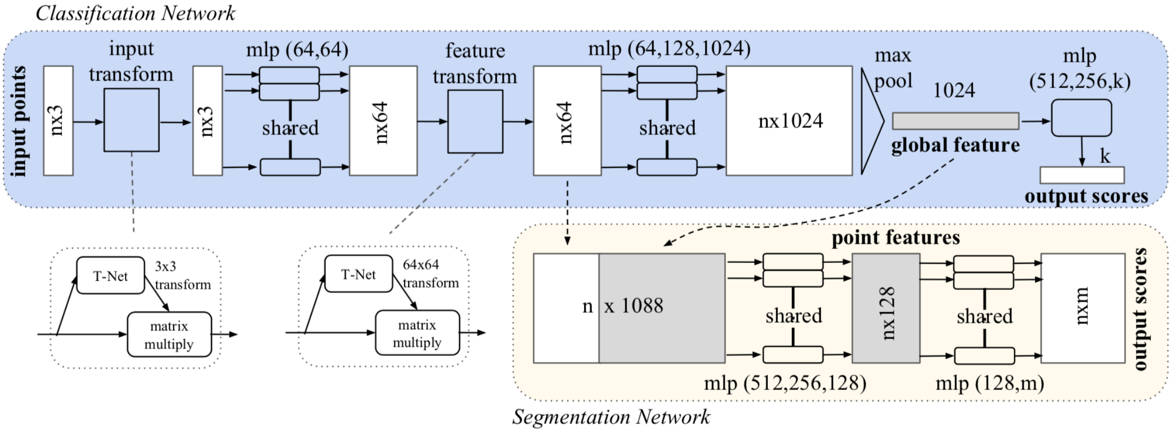


Figure 2.16: PointNet architecture. Image taken from (Qi et al., 2017)

in non-Euclidean domains. Also, grid-based local poolings are not always applicable as we can no more rely on fixed-size neighborhoods. Instead, global poolings might be a viable solution for both point clouds and graphs, but only at the condition of using permutation invariant operations, e.g., max or mean.

Point-wise linear function In point clouds the relations between points are not explicit i.e., $\mathcal{E} = \emptyset$, and thus we can encode them independently. One way to do that is with a point-wise learnable linear function f like the one of dense layers in Equation (2.1). We can show the *permutation equivariance* of f by considering equation (2.5); by substituting y with the linear function f, $\tau(\cdot)$ with the permutation function $\tau_p(\cdot)$, and the single point $\mathbf{x} \in R^d$ with the entire point cloud $\mathcal{X}$ we obtain:

$$f(\tau_p(\mathcal{X})) = \tau_p(f(\mathcal{X})) \tag{2.12}$$

Then, considering the shared point-wise application of f, we can write:

$$\begin{aligned} \mathbf{W}\tau_p(\mathbf{x}_i) + \mathbf{b} &= \tau_p(\mathbf{W}\mathbf{x}_i + \mathbf{b}),\ \forall \mathbf{x}_i \in \mathcal{X} \\ \tau_p(\mathbf{x}_i) &= \mathbf{x}_i,\ \forall \tau_p \in \mathcal{T} \\ \mathbf{W}\mathbf{x}_i &= \mathbf{W}\mathbf{x}_i \end{aligned} \tag{2.13}$$

Pointnet A famous architecture for point cloud, which will be recalled later in the thesis, is PointNet (Qi et al., 2017) (see Figure 2.16). PointNet is a flexible yet powerful model that can be used for point cloud classification and segmentation with little variations. The classification variant is composed of an *encoder* which maps the point cloud $\mathcal{X}$ to a single descriptor, namely *global feature*, and a *classification head.* The encoder and the classification head are MLPs to guarantee permutation equivariance and invariance, respectively. The encoder is the same in the segmentation variant, while the segmentation head is built to be equivariant to the input permutations. To do so, the *global feature* is stacked $|\mathcal{X}|$ times and concatenated to a latent representation of $\mathcal{X}$ obtained after the first layers of the encoder. The newly obtained representation of the point cloud is classified point-wise with an MLP. It is important to notice that this network never downsamples or aggregates points in the input point cloud except for the single global pooling at the end of the encoder. Hence, it never exploits the local context of the points.

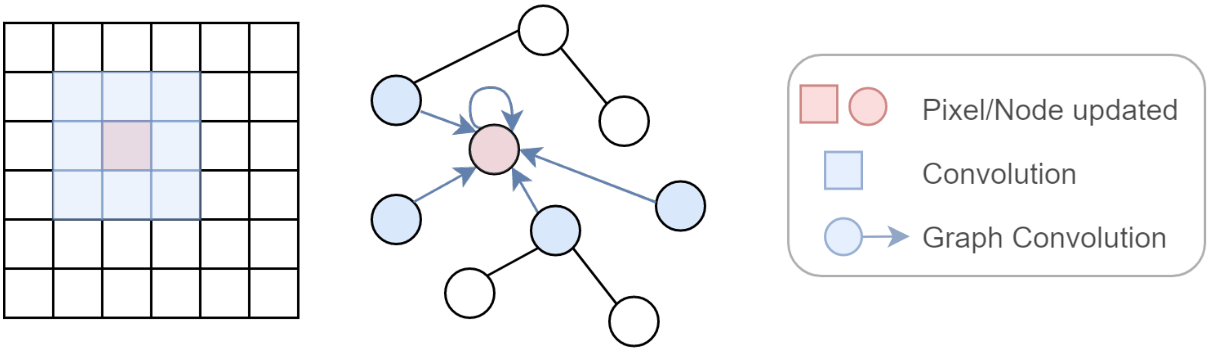


Figure 2.17: Comparison between grid and graph convolution. On the left an example of how a grid convolutional filter is applied on the pixel structure. The pink pixel is updated using the information in the contextual pixels. On the center, the analogous situation with graph convolution. Nodes at 1-hop distance from the pink node are used to aggregate/learn contextual information. Image credits: Simone Scardapane

Graph convolution When considering graphs with $\mathcal{E} \neq \emptyset$, we have *locality* given by 1-hop edges. Differently from locality in grid, here we do not have a fixed number of local neighbors but rather a varying number depending on the existing edges. We may define the set of 1-hop neighbors of a node $\mathbf{x}_i$ as $\mathcal{N}_i = \{j : (i, j) \in \mathcal{E}\}$. Again, as in the case of point clouds, we want to apply a shared function node-wise, this time indicated with ϕ, to obtain permutation equivariance on the entire graph $\mathcal{G}$. This means that ϕ is locally applied to each tuple $(\mathbf{x}_i, \mathcal{N}_i)$ – in a parallelism with the grid convolution (Section 2.2.1), ϕ represents the convolutional filter, which in this case is known as *propagation* or *message passing*. See Figure 2.17. To guarantee global permutation equivariance, ϕ must not affect the structure $\mathcal{G}$. Hence, it has to be *permutation invariant*. A *graph convolution* ϕ can be obtained with an *aggregator* operator $\bigoplus$ such as max or average, applied over a learnable linear transformation of $\mathbf{x}_j \in \mathcal{N}_i$ weighted by the importance of x_j with respect to x_i:

$$\mathbf{z}_i = \phi\left(\mathbf{x}_i, \bigoplus_{j \in \mathcal{N}_i} e_{ij} f(\mathbf{x}_j)\right) \tag{2.14}$$

In the simplest formulation of graph convolutions e_{ij} is a given constant (Kipf and Welling, 2017), while in more advanced formulations like graph *attentional* convolutions (Veličković et al., 2018) and dynamic graph CNN (Wang et al., 2019) the value of e_{ij} is learned.

Chapter 3

Related Works

Tractogram analysis is pivotal for several real-world applications, e.g., neuroanatomy understanding, neurosurgical planning, brain connectomics, and neuroscientific group studies. However, reconstructing highly accurate tractograms is hard, especially due to the multiple pre-processing steps and tracking parameters involved in a tractography pipeline. One of the main problems affecting tractograms is the abundant presence of *false positive* fibers, which are *anatomically non-plausible*. The current literature is facing this problem by imposing *anatomical priors* either during the tracking or after it, i.e., directly on tractograms. In this thesis, we focus on the latter. In particular, we inspect solutions that *segment* WM regions of interest (ROIs) to avoid false positives in well-known bundles (see Section 3.1), and solutions using fiber-specific plausibility criteria to *filter* entire tractograms (see Section 3.2).

Tractogram applications In the real world, there are plenty of scenarios where tractograms play a key role. For such scenarios, the quality of tractograms considerably affects the outcome of the task, and for some clinical cases, it might even be life-saving. Some exemplary applications are:

- *Neuroanatomy understanding*, e.g., (Maffei et al., 2018; Jeurissen et al., 2017; Hau et al., 2016, 2017; De Benedictis et al., 2016), where studies are aimed to confirm or explore in details the anatomy of certain white matter structures. For examples (Hau et al., 2016, 2017) individuate new termination regions for the inferior front-occipital fasciculus (IFOF) and the uncinate fasciculus (UF).

- *Neurosurgical planning and intervention*, where surgeons first plan the intervention strategy and then operate with the help of a neuro-planning and -navigation software (Essayed et al., 2017; Henderson et al., 2020; Yang et al., 2021; Clark and Byrnes, 2010; Orringer et al., 2012). The structural images, the tractography, and the volumetric mask of anatomical structures such as bundles or fiducial ROIs are provided to the software, which can be used to plan the surgery trajectory and subsequently to support the operation. The more accurate the provided information is, the higher is the possibility for neurosurgeons to limit post-operative injuries or cognitive impairments.

- *Brain connectomics*, where the focus is to understand the interactions between different cortical regions. Such interactions are modelled as *edges* of a graph whose *nodes* are the regions. The corresponding graph adjacency matrix, also known as brain *connectome* (Sporns et al., 2005; Sporns, 2013), contains in each cell an estimate of the connectivity between two regions based on the fibers connecting them. Characterizing the connectome may help to validate functional correlations but also explore the information integration, segregation, and propagation in the brain (Yeh et al., 2020; Toga et al., 2012).
- *Neuroscientific group studies*, which often require reliable in-vivo tractography. Neuroscientists may be interested in quantitatively characterizing bundles of interests by means of *tractometry*. For example, the Fractional Anisotropy (FA) distribution along a bundle produces what is called the *tract profile* (Yeatman et al., 2012; Chandio et al., 2020; Kruper et al., 2021). Statistical group analysis based on tractometry may reveal meaningful patterns across populations as well as atypical conditions of specific subgroups of individuals. Also, in the case of longitudinal studies (Leitner et al., 2015), tractometry quantifies the development of white matter structures.

Challenges of tractography Tractography pipelines are complex and involve the choice of multiple pre-processing steps and tracking parameters (see Section 2.1.2). Quantitative studies like Bastiani et al. (2012); Thomas et al. (2014) have investigated how the choice of hyperparameters might impact the result of tracking algorithms. Especially stochastic methods for probabilistic tracking introduce a significant source of variability and a critical dependency from the choice of parameters' values. In general it has emerged that slightly different choices may produce quite different tractograms (Bastiani et al., 2012; Thomas et al., 2014), and the evaluation of their accuracy is not straightforward (Neher et al., 2015; Girard et al., 2020; Rheault et al., 2020b,a).

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), ISMRM 2015 tractography challenge (Maier-Hein et al., 2017), Traced on 2017 (Nath et al., 2020), Votem on 2018 (Schilling et al., 2019b), and DISCO 2021 (Girard et al., 2021). 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 recurring weakness of tracking algorithms is the generation of false positive fibers (Maier-Hein et al., 2017). With the evolution from DTI to HARDI tractography, most algorithms

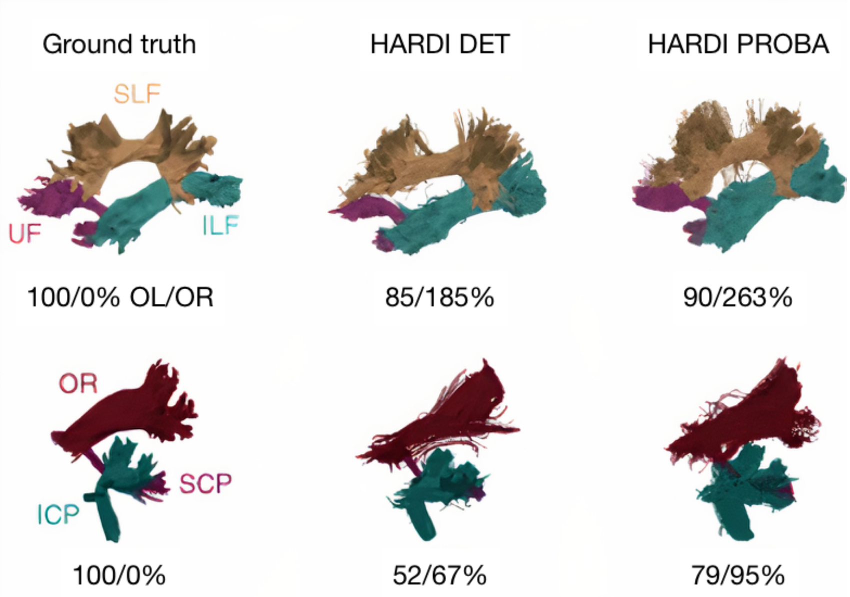


Figure 3.1: Example of false positive generation in tractography. Under each set of bundles the volumetric overlapping (OL) (volumes intersection) and overreach (OR) (volumes difference) quantify the impact of false positives and negatives. It can be observed an high OR value especially for HARDI PROBA, confirming the presence of a large amount of false positives. Image taken from (Maier-Hein et al., 2017)

have been able to produce high sensitivity tractograms, i.e., to miss less well-known pathways, at the cost of increasing the number of false positives fibers, i.e., decreasing specificity (Thomas et al., 2014), and often to over-track the easiest pathways. See Figure 3.1 The result is that the fiber density is not compatible with the physiological distribution of axons in the white matter (Raffelt et al., 2012) and that there is a large quantity of anatomically non-plausible streamlines, which for example, terminates prematurely inside the WM or connect cortical regions known to be unrelated. Balancing sensitivity and specificity of tracking is a hard problem so that there is no best tracking algorithm, but rather one can find the most suitable for a certain application (Thomas et al., 2014). For instance, (Zalesky et al., 2016) demonstrates that for connectome analysis, false positive streamlines are twice impactful as false negative ones, and consequently (Sarwar et al., 2019) suggested that deterministic is preferred over probabilistic in this application.

Anatomical prior in tractography Recent strategies of tractography are trying to mitigate the generation of false positive fibers by introducing additional anatomical constraints. The idea is to integrate such anatomical priors as part of the tracking heuristics. 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 have been constrained using topographic regularity (Aydogan and Shi, 2018) and geometric shape priors (Dong et al., 2017b). However, the brain white matter anatomy is not entirely known, and thus it is not possible to perfectly constrain a priori the whole brain tracking. Small errors in the application of whole-brain constraints like a noisy tissue segmentation in ACT (Smith et al., 2012) might inhibit the generation of entire bundles such as the IFOF or the UF, which pass through delicate regions close to the WM boundaries (Wasserthal et al., 2018).

For this reason, it turns out to be easier and effective to focus on the characterization of white matter where the anatomy is well-known (Schilling et al., 2020). This typically means to characterize WM regions of interest (ROIs) that define specific bundles (Yendiki et al., 2011; Maffei et al., 2021; Thomas et al., 2014; Chamberland et al., 2017; Reisert et al., 2018; Rheault et al., 2019). Alternatively, a more global WM characterization might be obtained at the level of tractogram by defining the property of a fiber pathway to get considered either plausible or non-plausible, i.e., artifactual. We discuss existing methods based on these two types of characterizations in the following sections 3.1 and 3.2.

3.1 ROI-based WM characterization

Anatomical regions are the easiest way of characterizing white matter. They are typically used to define the pathway of well-known bundles. For example, neuroanatomists use them to guide the ex-vivo brain dissection (Sarubbo et al., 2013; Martino et al., 2010). By localizing eloquent white matter regions, e.g., the *stem* of a bundle (Hau et al., 2016, 2017), they can dissect the pathways of interest. Analogously, given a tractogram of millions of fibers, virtual ROIs can be used to extract the (small) portion of fibers belonging to well-known bundles.

In literature this procedure is usually referred as *ROI-based bundle segmentation*, in contrast to other bundle segmentation approaches based only on the fiber pathway e.g., clustering-based (Siless et al., 2018; O'Donnell et al., 2006; O'Donnell and Westin, 2007; Guevara et al., 2012; Garyfallidis et al., 2012), correspondence-based (Garyfallidis et al., 2018; Sharmin et al., 2018), and classification-based (Gupta et al., 2017; Olivetti and Avesani, 2011; Zhang et al., 2020), which we consider out of the scope for this thesis. ROI-based approaches are widely adopted as they allow anatomically grounded, i.e., anatomy-driven, and convenient solutions that avoid dealing directly with millions of streamlines, while still producing a fiber-based representation of the bundle as output.

The typical ROI-based segmentation procedure consists of two steps: (i) ROIs are *segmented* on volumetric images of the brain, and then (ii) combined through logical operators such as `AND, OR, NOT`, to extract the fibers of interest from the tractogram. In the first step, the segmentation of ROIs can be performed manually; otherwise, ROIs can be derived by existing *template* and *parcellation* atlases. This step is crucial as it allows the encoding of anatomical knowledge into anatomical prior. We describe the existing techniques of WM ROI segmentation in the next Section. Similar to step one, also the second step of bundle extraction can be done via manual/ad-hoc combination of ROIs, or automatically, via *segmentation rules* or *protocols*, which define in a general and reproducible way the trajectory of well-known bundles. Methods that adopt a fully automatic segmentation approach involving atlas-ROIs and segmentation rules are usually referred to as *rule-based* methods.

Another approach to ROI-based WM characterization is proposed by *direct CNN-based* methods (see Section 3.1.3). In this approach, the WM characterization is faced with data-driven methods based on Convolutional Neural Networks(CNNs). The basic idea is to directly segment bundles as volumetric ROIs, namely *bundle masks*, from the diffusion images of the brain. Compared to manual and rule-based methods, which adopt two steps to obtain a bundle of fibers as output, there is a single step leading to a volumetric segmentation of the bundle. No fibers are involved during this process.

3.1.1 White matter ROI segmentation

The task of *white matter ROI (WM-ROI) segmentation* concerns the delineation of one or multiple regions within the white matter volume. Practically, this task consists in producing a binary mask of the brain, having 1s in the voxels composing the region of interest (ROI) and zeros elsewhere. In addition to the aforementioned manual versus atlas-ROIs, we may distinguish between *anatomical* and *heuristic* ROIs. The former have a well-defined anatomical meaning, delineating fiducial regions of the WM or termination regions in the interface GM/WM of gyri and sulci. Contrarily, the latter has no anatomical correspondence and serves only to facilitate the segmentation of bundles. For example, a heuristic ROI could be a 2D plane used as a primal filter to extract a broad superset of the bundle or an ROI to avoid fibers not belonging to the bundle, including artifacts.

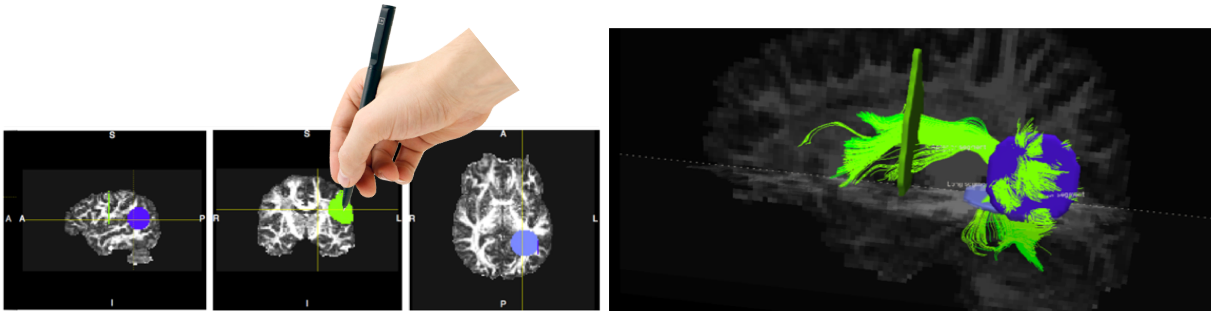


Figure 3.2: Manual ROI-based bundle segmentation. ROIs are sketched on the 2D views (sagittal, coronal, and axial) of the brain. Then, they are combined to filter the bundle of interest, in this case an Arcuate Fascicle. Image credits: Paolo Avesani

Manual ROIs In the manual ROI segmentation, schematized in Figure 3.2, the expert, with the aid of software, draws the ROIs of interest on a volumetric brain image like the

T1w or the FA of the subject at hand. In the common software, e.g., TrackVis (Wang et al., 2007), MRTrix3 (Tournier et al., 2019), and many others (Rheault et al., 2016; Norton et al., 2017; Yeh, 2021; Fischl, 2012; Jenkinson et al., 2012; Yushkevich et al., 2006), the segmentation of *anatomical* ROIs is performed on (one or more) 2D slices of the reference image. For this reason, anatomical ROIs are often planar-like, as shown in Figure 3.2. The segmented anatomical ROIs are then used as *inclusion* ROIs to extract the fibers which they intersect. If among the extracted fibers there are still artifacts or other fibers not belonging to the bundle of interest, experts can remove them by drawing ad-hoc *heuristic* ROIs, which are used as *exclusion* ROIs. Since these ROIs do not have to match the underlying anatomy, they are usually outlined with regular shapes such as spheres, disks, or entire planes, already available in the software.

The manual drawing of ROIs on subject-specific reference images is undoubtedly an ideal way of characterizing the white matter anatomy. First, because ROIs directly encode the anatomy. Second, expert supervision guarantees the extraction of anatomically plausible and accurate bundles. However, the manual approach also has some practical drawbacks. It requires high expertise, and it is very time-consuming (Jones, 2008). Also, the final segmentation depends on the expert subjective interpretation of the WM anatomy, which is often in disagreement with the interpretation of other experts (Rheault et al., 2020a; Schilling et al., 2021). As a result, the manual approach does not scale to a large or even medium population of individuals.

Template and parcellation atlas ROIs One solution to scale up ROIs segmentation to populations is the definition of atlases of ROIs on standard brains like MNI152 (Fonov et al., 2011). The idea is that ROIs might be segmented only a few (tens) times for the definition of the atlas and then derived to subject-specific anatomy through image registration. ROI atlases can be created in different ways and encode different anatomical information. For examples, in works such as (Catani et al., 2002; Catani and Thiebaut de Schotten, 2008, 2012; Mori et al., 2005; Wakana et al., 2007; Thiebaut de Schotten et al., 2011; de Groot et al., 2013; Warrington et al., 2020; Rheault et al., 2022) one or two anatomical ROIs of *“obligatory passage”* are drawn to segment each well-known bundle on a subset of subjects. Then, the segmented ROIs are averaged to obtain template ROIs. In some cases like (Wakana et al., 2007; Warrington et al., 2020) also exclusion ROIs are defined as part of the atlas. Otherwise, there are parcellation atlases, e.g., (Salat et al., 2009; Schiffler et al., 2017), which instead of defining template ROIs from manual drawing, expand cortical parcellations to cover also part of the white matter.

In general, ROI atlases are a convenient way of encoding anatomical information of WM. Once defined, their application at a large scale comes at no cost. However, the syntactical definition of comprehensive ROI atlases requires extensive knowledge (and agreement) of the WM anatomy, which is, to date, still not available. Existing atlases are limited to the most well-known major white matter bundles (Hansen et al., 2021), and they are missing some anatomical landmarks, which have been already investigated in post-mortem studies. For example, the *stems* of the IFOF and the Uncinate Fascicle (Hau et al., 2016, 2017) are not present in any of the existing atlases. Another critical aspect of ROI atlases is the poor anatomical specificity (Figley et al., 2017). The process of ROIs averaging to accommodate the anatomy of a group of individuals, and guarantee reproducibility, leads to the loss of subject-specific anatomical variabilities (Wakana et al., 2007). This phenomenon is even more accentuated by the step of registration to a new subject. Although modern (non-linear) registration algorithms, e.g., (Avants et al., 2008), employ advanced optimization techniques, it is very unlikely to perfectly map the smoothed/simplistic anatomy of a standard brain to the specific anatomy of the target subject. As a result, co-registration and generalization errors can highly affect the final segmentation, especially when involving small ROIs.

3.1.2 Segmentation rules definition

Segmentation rules are generated by combining multiple ROIs in logical propositions. In the literature there are many methods adopting a rule-based approach, e.g., Automatic Fiber Quantification (AFQ) (Yeatman et al., 2012), White Matter Query Language (WMQL) (Wassermann et al., 2013, 2016), and XTRACT (Warrington et al., 2020). These methods propose segmenting bundles based on ROIs borrowed from existing atlases. In particular, AFQ and XTRACT present segmentation rules for 20 and 42 bundles, respectively, combining multiple *waypoint* and *exclusion* ROIs. Differently, WMQL proposes a set of *open* rules, which can be adapted to different atlases of ROIs. The WMQL assumption is that there is no gold standard of anatomy, but experts may redefine rules based on new anatomical findings. An example might be again the case of the *stem* of IFOF and UF (Hau et al., 2016, 2017), whose discovery led to the revision of those bundles' termination regions. Thanks to the use of *open* rules, WMQL has become widely adopted in the community, and also in our case, we decided to use it as the reference method for the approach presented in Chapter 5.

WMQL (Wassermann et al., 2016) In WMQL the proposed *query language* is composed of predicates that act on anatomical ROIs derived from existing parcellations like (Salat et al., 2009; Desikan et al., 2006; Fischl et al., 2002), and logical operators e.g., `AND`, `OR`, which combine them to obtain segmentation rules. Specifically, Wassermann et al. identify two types of predicates, referred as: *divisional terms*, and *relative clauses*. *Divisional terms* are anatomical predicates stating if streamlines traverses or ends in a certain brain structure, such as `only(roi_name)` and `endpoints_in(roi_name)`. *Relative clauses* are instead relative position predicates indicating whether the streamlines are, for instance, `superior_of(roi_name)` or `frontal_of(roi_name)` to a brain structure. Using this formalism each bundle of interest can be defined in a strict and reproducible way.

Rule-based methods are principled for the current knowledge of anatomy. The result of the segmentation is a direct consequence of the anatomical definition provided for the bundle at hand. However, this means that a (probable) poor anatomical specificity of the employed atlas ROIs negatively impacts the accuracy of the final segmentation. For this reason, rule-based methods usually require a subsequent step of manual revision, where some additional exclusion ROIs are used to remove false positive fibers. In some cases, it can also happen that instead of false positives, a segmentation rule underestimates the true pathway by missing several streamlines as false negatives. This case usually happens when the segmentation rules contain multiple exclusion ROIs, which may wrongly fit the target brain's anatomy.

3.1.3 Direct CNN-based methods

Direct CNN-based methods adopt a data-driven approach to learn the segmentation of bundle masks in new unseen brains. A dataset of reference fiber bundles, usually obtained with *manual* and *rule-based* procedures on multiple subjects, is used to derive a new dataset composed of the corresponding bundle masks. These masks are used as ground-truth for training a segmentation CNN, which is fed with diffusion-based images such as DTI or fODF. It is important to remark that given the translation equivariance and local stability of CNNs (see Chapter 2.2.1), the input brain images need to be only rigidly co-registered. At the end of the training, the CNN is able to segment all the bundle masks of a new subject with a single inference step. The output is a binary 4D mask of shape (H, W, D, N_b), where N_b is the number of target bundles to segment.

In the recent literature various direct CNN-based methods have been presented (Li et al., 2020; Lu et al., 2021; Wasserthal et al., 2018, 2019). They mainly differ for the type of input signal adopted, e.g., DTI versus fODF peaks, and training procedure, e.g., use of self-supervised loss. Such methods have achieved exceptional performance on bundle segmentation, being the top performer until the very recent publication of Classifyber (Bertò et al., 2021). In the following, we describe in more detail TractSeg (Wasserthal et al., 2018, 2019) the pioneering and most adopted direct method, which we choose as a competitor for our experiments in Chapter 5.

TractSeg (Wasserthal et al., 2018) This method proposes to segment the volumetric mask of bundles directly from the image of the fODF peaks of the white matter. An fODF peak is a 4D image encoding the direction of the three main spherical harmonics within each voxel. The resulting shape is $(H, W, D, 9)$. As in other direct methods, the ground-truth is a 4D sequence of binary bundle masks, (H, W, D, N_b). TractSeg employs a standard U-net model (Ronneberger et al., 2015), whose training is performed by feeding randomly sampled slices of the input peaks. The evaluation is carried out on a dataset of 105 healthy subjects, split into 84/21 for train and test. The performances are measured in terms of volumetric overlapping with the reference bundle masks.

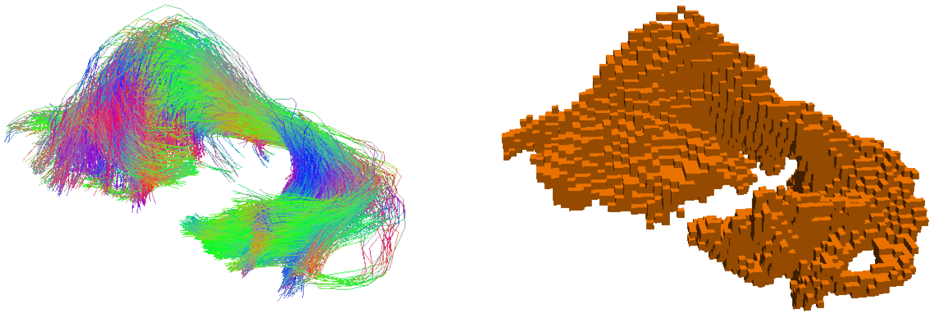


Figure 3.3: Streamline-based bundle (left) and voxel-based bundle, i.e. *bundle mask* (right).

TractSeg obtains the state of the art results in bundle segmentation. However, the produced voxel masks are rather rounded and smoothed both in the central part of the bundles but also in the termination part, where instead, the bundle is expected to have jiggled fanning. This behavior is probably due to CNN pooling layers responsible for eliminating some fine-grained features at each downsampling step. A quantification of such roundness bias is presented by means of the *fractal dimension* in (Bertò et al., 2021).

3.2 Tractogram filtering

The task of *tractogram filtering* concerns the filtering out of anatomically non-plausible fibers from tractograms, as considered false positives of the tracking procedure. This post-processing step is complementary to the anatomical constraining of tracking, which can only partially prevent the generation of false positives. Especially in the case of whole-brain tractography, the tracking constraints cannot be as accurate as they are for bundle-specific tractography (Rheault et al., 2019; Schilling et al., 2020), and thus there could be generated a higher amount of false positives. Note that also bundle-specific tracking may generate false positives if not using manually subject-specific inclusion and exclusion ROIs (Wakana et al., 2007; Schilling et al., 2020). In general, if we want to characterize the whole white matter connectivity, we cannot rely only on ROIs, but we need to act the level of single fibers to identify the ones anatomically non-plausible. Unfortunately, the current knowledge of white matter does not allow for a full definition of the notion of anatomical plausibility. Different interpretations of plausibility may lead to different filtering criteria (Jörgens et al., 2021).

In the tractogram filtering literature, we may distinguish three types of filtering solutions: signal-based (Section 3.2.1), tractography-based (Section 3.2.2), and mixed signal- and tractography-based (Section 3.2.3). Signal-based solutions formulate the filtering task as an inverse problem of signal reconstruction, while tractography-based solutions adopt filtering criteria based only on the tractography data. Mixed approaches heuristically combine the two types of filtering. All of them carry out an unsupervised strategy.

3.2.1 Signal-based filtering

In the signal-based solutions, the plausibility of fibers is estimated by computing how much the diffusion signal explains their pathways. The base assumption is that the *density* of fibers in a tractogram must be compliant with the amount of underlying diffusion signal. Practically, this means that two voxels characterized by the same diffusion intensity should underlie the same number of streamlines. Unfortunately, existing tracking methods, even when using global approaches, are not able to guarantee uniform spatial distribution of the streamlines, see Figure 3.4. To mitigate this problem, the most common signal-based filtering methods cast the filtering task as a global regularization problem by assigning a weight to each fiber. Fibers with weights lower than an arbitrary threshold are filtered out. In SIFT (Smith et al., 2013) and SIFT2 (Smith et al., 2015a,b), the weights are a proxy of the density of the fibers, higher for fibers traversing lower density white matter zones. 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 microstructural information.

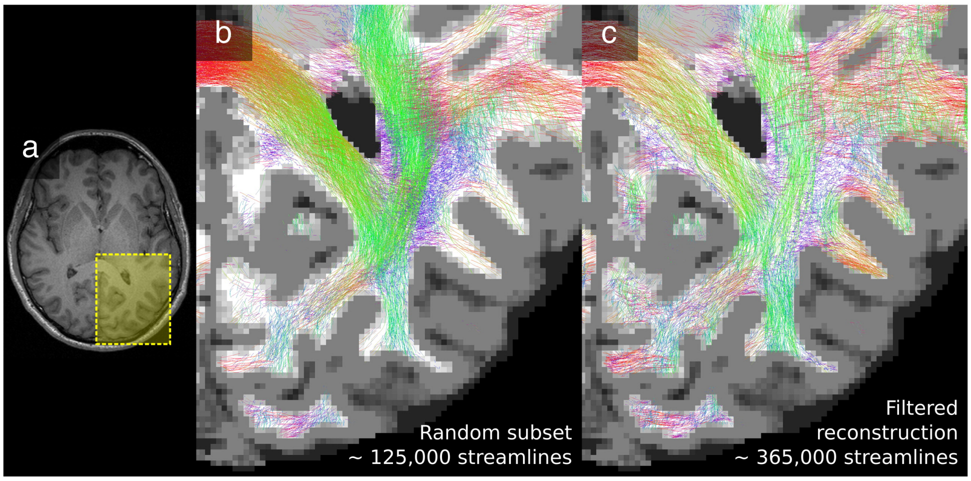


Figure 3.4: SIFT (Smith et al., 2013) regularization. By comparing b with c we observe a different spatial distribution of fibers. In c, after the filtering, we have a more uniform distribution of fibers. Image taken from (Smith et al., 2013)

In all signal-based methods, the thresholding of weights that discriminates between plausible and non-plausible fibers is managed with heuristics. Finding the optimal threshold may not be trivial and can be done only by empirical trials. Moreover, as remarked in (Smith et al., 2020a; Frigo et al., 2020; Rheault et al., 2019), the filtering operated with a regularization approach might remove fibers whose pathway is anatomically plausible. Indeed, the signal-based definition of plausibility does not capture the anatomical structure of white matter. The main motivation of signal-based filtering is instead to enable quantitative studies of tractograms such as brain connectome analysis — even though ongoing debates are arguing a controversial impact also on connectome (Zalesky et al., 2020b,a; Smith et al., 2020b).

3.2.2 Tractography-based filtering

Groub-wise filtering The alternative approaches are based only on tractography. Without considering the diffusion signal, tractography-based methods adopt filtering criteria based on the *consistency* of the tractogram structure. Their basic assumption is that the topographic regularity of tractogram structures across individuals might be a good proxy of anatomical plausibility. A fiber is considered as non-plausible when it is an *outlier* for the mean distribution over multiple subjects. Different unsupervised methods by leveraging the group-wise 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. However, consistency over individuals does not implicate anatomical plausibility. These filtering methods lack of general anatomically-informed constraints, and so they are subject to bias present in the population, mostly derived from tractography algorithms (Zhang et al., 2021).

Intra-individual filtering While a large population of tractograms provides a more robust estimate of structure regularity, thegroup consistency constraints tend to eliminate inter-individual differences. To contrast the smoothing effect due to population averaging, other unsupervised methods refer to *intra-individual* filtering criteria. For example, in (Yeh et al., 2019) the fiber density map is used as a proxy of anatomical plausibility. In following this criterion, fibers are filtered out as non-plausible when they traverse low-density areas, i.e., they are intercepted by voxels with an extremely low number of fibers. Another example is the sigma-based cleaning adopted by AFQ (Yeatman et al., 2012). A fiber is considered non-plausible when its length or trajectory is distant, respectively, four or five standard deviations from the mean of a bundle. Similar but more advanced is BundleMAP (Khatami et al., 2017) that adopts a one-class Support Vector Machine (SVM) to identify local outliers in a bundle. In general, these approaches define as non-plausible fibers that are *outlier* with respect to their local neighborhood.

FINTA (Legarreta et al., 2021) The most recent unsupervised tractography-based methods for tractogram filtering are investigating the use of deep learning techniques, such as *Filtering IN Tractography using Autoencoders* (FINTA) (Legarreta et al., 2021). Such a method proposes a convolutional autoencoder network to learn an embedded representation of fibers. The autoencoder learns the embedded representation by optimizing the reconstruction of the input fiber trajectories. An input streamline is encoded through traditional convolution and pooling layers until a single descriptor vector is obtained. This vector is also called *latent embedding*, and it is used to reconstruct the input trajectory by means of decoding layers such as upsampling and convolution. After the training procedure, the fibers are projected into a new latent space where the computation of nearest neighbors might easily detect the similarity as proximity, see Figure 3.5. The working hypothesis of FINTA is that in this new space, fibers anatomically plausible are closer to each other and far from non-plausible fibers. However, as the learning of the embedding is not driven by neuroanatomical knowledge, there is no guarantee for the subsequent fiber filtering to capture the notion of anatomical plausibility properly.

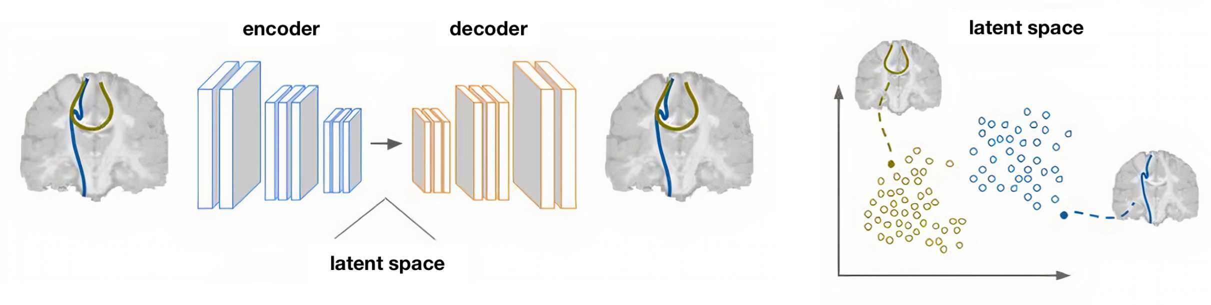


Figure 3.5: FINTA methodology. Image adapted from (Legarreta et al., 2021)

3.2.3 Mixed signal- and tractography-based filtering

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), where signal-based filtering is regularized with a priori knowledge on segmented bundles. The basic intuition is that where a priori knowledge of neuroanatomical bundles is not available, the fibers are regularized by signal-based filtering. In contrast, along the pathways of known bundles, the fibers are filtered out if they do not meet the expected topographic regularity. The result is that the envelope of well-known bundles is cleaned from outliers, while other *undefined* fibers, i.e., fibers not belonging to any well-known bundle, are filtered in agreement with the signal-based expected fiber distribution. However, this sharp distinction between fibers of bundles versus *undefined* fibers may lead to incoherent tractograms, where a streamline with an anatomically plausible pathway, but not belonging to any bundle, is declared as non-plausible because it does not satisfy the signal grounding. Viceversa, a streamline belonging to a bundle but not explained by the diffusion signal is declared plausible.

Literature recap An overall recap of existing tractogram filtering techniques points out that all the current filtering strategies, i.e., signal-based, tractography-based, and mixed, work unsupervised. In the case of signal-based strategies, the lack of supervision is compensated by the principled approach, which leads to a definition of fiber plausibility that does not coincide with anatomical plausibility. While, in the case of tractography-based and mixed approaches, the trend, especially in the more recent works (Neher et al., 2018; Nie and Shi, 2019; Schiavi et al., 2020; Ocampo-Pineda et al., 2021), is to include a priori knowledge about bundles of interest. The use of bundles prior is a workaround due to the lack of knowledge of whole-brain WM anatomy. The introduction of such a priori can be interpreted as an additional anatomical constraint, similar to what recent tractography algorithms proposed (Rheault et al., 2019).

Chapter 4

Research Questions

This thesis aims to improve the overall accuracy of tractography-based representations of the white matter. We frame this problem as a general lack of anatomical grounding in reconstructed tractograms. For this reason, we focus on the integration of additional anatomical knowledge, encoded as *anatomical priors*, to improve tractograms. We explore two types of solutions: one using *anatomical* ROIs as additional prior to characterize well-known white matter structures, and the other based on the *extensional* definition of fibers as *anatomically plausible* or *non-plausible*. In practice, we propose to integrate such priors with a *supervised deep learning* approach, where different models such as Convolutional Neural Networks, and Geometric Deep Learning models may learn the different WM characterizations, i.e., ROI-based and fiber-based. Multiple open questions arise from this working hypothesis. In RQ 1 and RQ 2 we make them explicit.

Diffusion signal is not fully informative of WM Diffusion MRI together with tractography has enabled advanced studies of the in-vivo brain white matter otherwise not possible. However, the ambiguity of the diffusion signal in certain regions makes the procedure of fiber tracking ill-posed (Jeurissen et al., 2017). The resulting tractograms are not fully compliant with the underlying white matter anatomy (Maier-Hein et al., 2017; Nath et al., 2020), lacking of sensitivity i.e. miss some pathways (*false negatives*), and/or specificity i.e. contain anatomically non-plausible pathways (*false positives*). In particular the tractogram specificity is increasingly concerning the community, as multiple studies (Petit et al., 2019; Maier-Hein et al., 2017; Nath et al., 2020) are revealing that a large portion of the fiber in a tractogram are *non-plausible*. Hence, to mitigate the presence of false positive, there is the need of integrating additional information beyond the diffusion signal, which might solve its ambiguities.

The adoption of additional a priori information require us to define:

1. Which type of information to exploit

2. How to virtually represent such information

3. How to integrate it in a computational method

Exploiting anatomical knowledge From the analysis of the literature (see Chapter 3), it emerges that the most prominent trend is to exploit anatomical knowledge as additional information to solve tractogram inaccuracies due to signal ambiguities. Anatomical knowledge is being encoded with *anatomical priors*[1] such as fiducial white matter regions or known fiber pathways. These priors are exploited both ex-ante the generation of a tractogram (Smith et al., 2012; Girard et al., 2014; Rheault et al., 2019) i.e., during tracking, and ex-post to clean it from artifacts (Schiavi et al., 2020) or extract bundles of interest (Wassermann et al., 2016; Wasserthal et al., 2018). The benefit of imposing anatomical priors in tractography has been demonstrated by quantitative studies like (Schilling et al., 2020; Girard et al., 2020), where histology or macaque tracing is used to validate the accuracy of anatomy-informed tractography. The take home message of these studies is that when computational methods are properly bounded with anatomical priors, the quality of the produced tractograms or bundles is higher, i.e., less artifacts, and more coherent with the white matter anatomy.

Voxel- and fiber-based representation of WM The practical use of anatomical priors requires proper virtual representations to encode the anatomical knowledge of interests. Certain *data representations* might be more suitable than others to encode certain anatomical priors. For example, existing methods usually represent white matter anatomy using either a *voxel-* or a *fiber-based* representation. In the first case the anatomical prior is a volumetric image like an ROI mask, while in the second case the prior is represented as 3D polylines i.e., sequences of 3D points. The use of the one or the other representation to map anatomical principles is often an argument of debate. Our position is that both the two might be useful, and their adoption depends on the type of investigation. Volumetric representations might be more convenient when the aim is to characterize specific white matter regions like the termination or the waypoint of a bundle. On the contrary, a voxel-based representation might be sub-optimal if we want to identify artifactual trajectories of pathways, as it may not capture the sequential pattern of fibers.

Anatomical priors in supervised learning How can we practically integrate the voxel- or fiber-based anatomical prior into computational methods? In the literature, this step is often carried out by means of heuristics, which end up in the definition of rules or constraints based on the current knowledge of the anatomy (Wassermann et al., 2016; Yeatman et al., 2012). However, the definition of rules is easier and effective in specific cases when *"we know where white matter pathways start, where they end, and where they do not go"* (Schilling et al., 2020). Instead, defining general rules that concern the whole

[1] Note that here the notion of anatomical prior does not refer to a bayesian prior meaning, commonly used in machine learning.

white matter, and are applicable to large populations is a much harder challenge. For this reason other methods like (Bertò et al., 2021; Wasserthal et al., 2018) prefer to adopt a data-driven approach, where data annotations are used as a mean to integrate the anatomical prior in a supervised learning setting. In fact, the use of a sufficiently large example set allows machine learning (ML) models to learn and generalize the rules underlying the label distribution. The only drawback of traditional ML approaches, e.g., (Bertò et al., 2021), is the need to represent the input with handcrafted *feature vectors*, which require a lot of domain-specific expertise and still may be suboptimal for the task at hand.

Using deep learning To avoid the handcraft design of the input data representation we endorse the use of deep learning (DL) (LeCun et al., 2015). Deep neural networks (NNs) are able to learn representations, i.e., latent feature vectors, autonomously, based on the optimization of the objective task (Bengio et al., 2013). This means that we do not need to encode a priori the anatomical information by means of sophisticated features, but rather, during NNs training we can directly learn such anatomical information in the form of latent features. However, without adopting handcrafted input vectors, we have to deal with different data structures e.g., voxel- and fiber-based, which may require different DL architecture (see Section 2.2).

Encoding the anatomical prior with voxel-based or fiber representations makes a big difference for the type of DL model we can use. In particular, the regular grid structure of voxel-based representations is optimal for Convolutional Neural Networks (CNNs) (LeCun and Bengio, 1995) (see Section 2.2.1), which can learn local features at different resolution of the grid and thus succeed in task like image segmentation. Differently, the fiber representation is not characterized by a regular structure because of the variable length and the lack of orientation. Thus, it may require different deep learning models with respect to standard CNNs. However, the current literature is lacking of satisfactory solutions to this problem. Existing fiber-based DL approaches adopts suboptimal approximations such as resampling to a fixed number of points (Gupta et al., 2017; Legarreta et al., 2021; Zhang et al., 2020) and heuristic re-orientation of fibers (Legarreta et al., 2021; Zhang et al., 2020), which allow them to use standard CNNs. Instead, we propose a novel and more principled approach based on a recent family of models, referred as Geometric Deep Learning (GDL) (Bronstein et al., 2017), see Section 2.2.2. Such models are designed to handle irregular structures like graphs (Wang et al., 2019) and point clouds (Qi et al., 2017), and thus might be more suitable to deal with the 3D polyline structure of streamlines.

ROI-based WM characterization White matter ROIs are widely adopted representations to identify fiducial anatomical regions. They are typically leveraged to define bundles as they provide a convenient/simplified way of describing fiber trajectories. ROIs can be manually drawn to encode different key anatomical points of the white matter. For example, termination and waypoints ROIs delineate the anatomical key-points of a bundle (Mori et al., 2005; Wakana et al., 2007; Warrington et al., 2020). A combination of ROIs can be used to automatically segment bundles from tractograms considering logic rules, i.e., *rule-based* bundle segmentation, such as in (Yeatman et al., 2012; Wassermann et al., 2016). Another strategy to characterizing the white matter is to consider the whole bundle as a single volumetric ROI, namely *bundle mask*, such as in (Wasserthal et al., 2018). This strategy proposes to ignore the tractogram and directly segment the whole bundle mask from the brain diffusion images using CNNs models, i.e., *direct CNN-based* methods.

In *rule-based* bundle segmentation, the quality of the ROIs is crucial to obtain accurate bundles. Ideally, ROIs are manually segmented by experts from the volumetric images of each target brain. However, this manual procedure is extremely time-consuming and cannot scale to a large number of subjects. Hence, to increase the method generalization, rule-based bundle segmentation considers ROIs from already existing atlases (Yeatman et al., 2012; Wassermann et al., 2016; Warrington et al., 2020). Unfortunately, the availability of WM ROIs is limited in atlases. Existing white matter atlases, e.g., (Wakana et al., 2007; Oishi et al., 2008; Salat et al., 2009), comprehend only some well-known regions while neglecting many other eloquent ROIs (Hansen et al., 2021). ROIs are standardized to templates missing the user's detailed information (Figley et al., 2017). Moreover, when applying atlas ROIs to specific user anatomy, the co-registration error further reduces the quality of such ROIs, leading to inaccurate bundle segmentation. For this reason, after the automatic rule-based segmentation (overall in clinic), a manual revision is mandatory.

The co-registration problem is not present in *direct CNN-based* methods since they train CNNs to segment volumetric masks in the specific subject space (Wasserthal et al., 2018). However, these methods directly segment bundles as a whole volume without considering fiber details. This voxel-based approximation often lead to a mismatch between the segmented masks and the actual fiber envelope, especially in the termination zones where the bundles' fanning draws irregular shapes (Bertò et al., 2021).

RQ 1. *Would it be possible to define an approach that exploits the information in anatomically-defined rules and the power of CNN? Can deep convolutional networks effectively segment fiducial white matter ROIs rather than entire bundle masks? If yes, is it beneficial to integrate the segmented WM ROIs into bundle segmentation rules?*

Tractogram filtering Tractogram filtering is a post-processing procedure that aims at removing anatomically non-plausible fibers from tractograms. This step is very important as tractography methods produce a large number of artifactual fibers, which are *anatomically non-plausible.* Recent tractography methods, e.g., (Rheault et al., 2019), mitigated this problem by focusing on well-known bundles and introducing ROI-based anatomical constraints. However, such solutions do not scale to the entire tractogram (Schilling et al., 2020) and produce artifacts when using template ROIs. Hence, to pursue the whole-brain anatomical plausibility, tractogram filtering is a better strategy as it analyzes fibers in their entirety. However, the limited knowledge of white matter connectivity does not allow for a gold standard distinction between anatomically plausible and non-plausible fibers. Different interpretations of plausibility may lead to different filtering criteria (Jörgens et al., 2021).

Existing tractogram filtering approaches adopt filtering criteria either following a *signal-based* plausibility of fibers (Smith et al., 2013; Pestilli et al., 2014; Daducci et al., 2015) or a *tractography-based* plausibility considering *fiber consistency* intra- and inter-individual, e.g., (Yeatman et al., 2012; Yeh et al., 2019; O'Donnell and Westin, 2007; Wang et al., 2018; Xia and Shi, 2020). However, both the filtering criteria are *unsupervised* with respect to the anatomy, resulting in suboptimal filtering. Fiber consistency criteria are subject to tractography (intra-individual) or population biases, which lead to considering non-plausible fibers specific to one individual. Instead, signal-driven filtering removes fibers whose pathway is anatomically plausible (Smith et al., 2020a; Frigo et al., 2020). For this reason, more recent approaches integrated anatomical prior of well-known bundles to extend signal-based methods (Neher et al., 2018; Nie and Shi, 2019; Schiavi et al., 2020; Ocampo-Pineda et al., 2021). But, the mixed criteria generate inconsistent tractograms, where the fiber's signal and anatomical plausibility can be conflictual.

Designing a filtering criterion suitable to capture the notion of anatomical plausibility is still an open challenge. Therefore, we propose to interpret plausibility following anatomical priors. Such priors can be used to obtain an extensional definition of fibers, i.e., labeling anatomically plausible and non-plausible fibers (Petit et al., 2019; Zhang et al., 2018), enabling supervised learning approaches.

Our proposal is to adopt a supervised deep learning approach, where available labelings guide the anatomical plausibility of the model. Given the continuous evolution of anatomical knowledge, we aim to find a model flexible to newer or different anatomical interpretations. As in the case of FINTA (Legarreta et al., 2021), we assume that the geometry of fibers is informative to understand their anatomical plausibility and that the best way to extract such geometrical information is to perform learning directly on the polyline representation. In this case, we speculate that Geometric Deep Learning models may constitute a better solution compared to standard CNNs.

RQ 2. *Can we successfully apply DL directly on streamlines? Are GDL models the most suitable to extract information from the streamline structure? How do such GDL models perform on tractogram filtering? Can they learn the anatomical features that regulate the anatomical plausibility/non-plausibility of fibers using the supervision of labeled tractograms?*

Chapter 5

Stem-based IFOF Segmentation with Deep Learning

In this Chapter, we present our first contribution *StemSeg*, which tries to answer the open problems raised in RQ1. After an initial introduction, where we present some crucial elements like the notion of anatomical *stem* and the Inferior Front-Occipital Fascicle (IFOF) (Section 5.1.1), we describe the method (Section 5.2) and the results obtained during the empirical analysis (Section 5.4). Finally, we follow-up the results by discussing their insights (Section 5.5). The reported analyses refer to a large extent to (Astolfi et al., 2020a) published at the International Symposium of Biomedical Imaging (ISBI) 2020.

Synopsis The aim of this study is to investigate how a deep learning method based on volumetric representation might be effective for the segmentation of white matter anatomical regions. As a case study, we will illustrate the improvement in the rule-based segmentation of the Inferior Frontal Occipital Fasciculus (IFOF) by combining a recent insight on white matter anatomy from ex-vivo dissection (Sarubbo et al., 2013; Hau et al., 2016) and a data-driven approach using a deep convolutional model (Ronneberger et al., 2015). A more accurate characterization of the IFOF might benefit pre-neurosurgery intervention planning and clinical studies on brain plasticity and brain disorders.

5.1 Introduction

The anatomy of a WM bundle can be characterized in-vivo at the individual level from recordings of diffusion MRI. After the first step of diffusivity model reconstruction and a subsequent step of fiber tracking, we may obtain a representation of structural brain connectivity, namely tractogram, as a collection of millions of streamlines. Each streamline is encoded with a 3D polyline. The task of bundle segmentation from a tractogram is concerned with identifying those streamlines that have a specific neuroanatomical meaning. In this study, we focus our contribution on the segmentation of the IFOF.

Current rule-based methods for bundle segmentation are either neglecting crucial neuroanatomical waypoints of a given bundle (Wassermann et al., 2013, 2016) or heavily relies on co-registration of template-ROIs (Yeatman et al., 2012; Warrington et al., 2020). When rules are not restrictive, or the co-registration is not successful, segmented bundles contain undesired streamlines, which do not reflect the anatomy of the bundle. On the contrary, very elaborated rules that combine several ROIs to avoid undesired fibers often lead to *over-segmentation*, i.e., the missing of a relevant portion of a bundle.

A recent branch of direct CNN-based bundle segmentation approaches, e.g., (Wasserthal et al., 2018), has demonstrated the efficacy of CNNs in segmenting binary masks of bundles directly from a fiber orientation map of the brain. While these methods obtain very good results on benchmarks, the produced bundle segmentation is not completely satisfactory: masks often miss the fine-grained fanning usual of WM bundles. Moreover, the produced masks are not effective if used as filtering constraints to extract bundles from a given tractogram. They can only be used to constrain the tracking area in ex-post bundle-specific tractography (Wasserthal et al., 2019).

In this work, we propose to go a step further by combining the anatomical soundness of rule-based methods with the ability of CNNs in segmenting volumetric masks, i.e., ROIs. The basic idea is to enable the automatic segmentation of waypoints ROIs, which are not comprised in existing atlases, using CNNs trained with anatomical supervision. Our method *StemSeg* segments a fiducial ROI for the IFOF, namely *stem-ROI*, directly in the native space of a subject, and use it as a waypoint constraint in a WMQL (Wassermann et al., 2013, 2016) bundle segmentation rule.

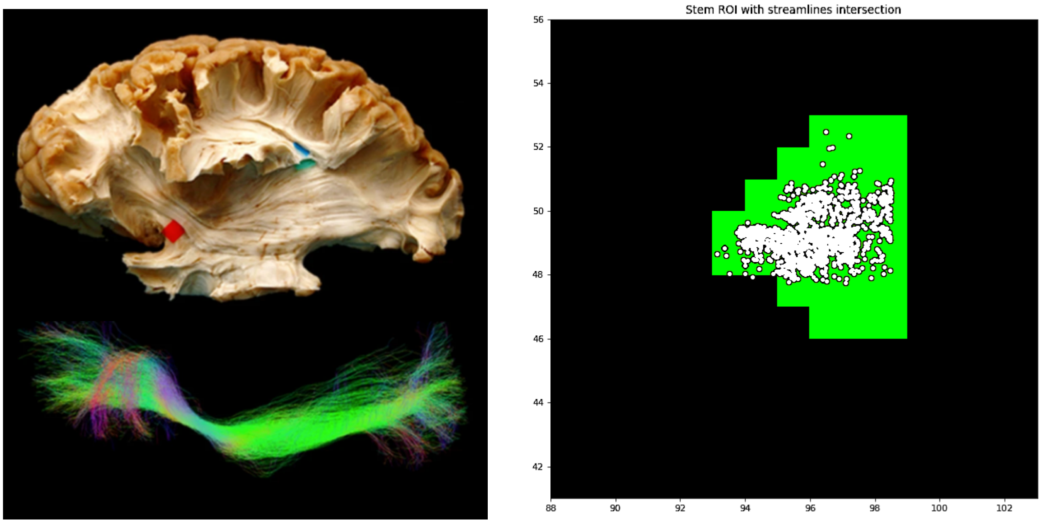
The task of stem-ROI segmentation is cast as a supervised segmentation problem of a volumetric image, and the output is a three-dimensional binary mask. Since a usual structural T1 MRI image does not provide enough information to discriminate the structure of the white matter, we refer to a colored fractional anisotropy image (CFA), a derivative from the fitting of diffusion tensor model (DTI) on DWI recordings (Basser et al., 1994). The CFA is capturing the information on local fiber orientation, and our working assumption is that the stem is characterized by streamlines with homogeneous fronto-occipital directions.

The proposed method is validated on a subset of the Human Connectome Project dataset, where expert neuroanatomists segmented the IFOF with different strategies of manual segmentation. In these tests, we provide empirical evidence that an automated stem-based IFOF segmentation can be robust for alternative strategies of manual segmentation. In addition, we prove that our IFOF segmentation significantly outperforms the most prominent method for direct bundle segmentation Tractseg (Wasserthal et al., 2018), which uses a deep learning model but neglects the anatomy of the stem. Finally, we inspect the generalization performance of StemSeg in case of simulated low resolution, as usually holds in clinical settings, as well as real patients data showing promising results in both cases.

5.1.1 IFOF and stem anatomy

The word *stem* generally indicates the central part of an object from which other parts can develop, e.g., the stem of a wine glass is the narrow vertical part supporting the container. Similarly, this term is adopted in neuroanatomy to define a very compact and homogeneous portion of a bundle, usually located in the central part of it (Sydnor et al., 2018). The typical structure of a bundle may be subdivided into three areas: the *stem*, the spray area, where pathways start to diverge from the stem, and the termination fanning, where pathways spread out to reach different cortical regions. This stem-centric view of WM bundles is particularly applicable to association bundles so that an entire taxonomy based on it has been proposed (Mandonnet et al., 2018). According to this view, the *anatomical stem* identifies univocally a bundle; only pathways belonging to that bundle should pass through it. A confirmation of stem-centric bundle definitions comes from numerous recent ex-vivo dissection studies e.g., (Kier et al., 2004; Martino et al., 2011; De Benedictis et al., 2016; Hau et al., 2017), including some focusing on the IFOF (Martino et al., 2010; De Benedictis et al., 2012, 2014; Sarubbo et al., 2013, 2019; Hau et al., 2016).

The IFOF is the longest association bundle, connecting parts of the occipital cortex, the temporo-basal area, and the superior parietal lobule to different cortical regions in the frontal lobe (Forkel et al., 2014). Based on the pathways termination, different sub-components may be individuated (Wu et al., 2016; Panesar et al., 2017). The main functional role of the IFOF is usually associated with lexico-semantic as its stimulation during awake neurosurgery causes semantic deficits (Martino et al., 2010; Turken et al., 2011; Sarubbo et al., 2013; Duffau, 2015). The recent studies (Sarubbo et al., 2013; Hau et al., 2016) on ex-vivo dissection (modified Klingler's technique) of the IFOF have provided a better characterization of its anatomy by validating the existence of the stem (see Figure 5.1a). Moreover, Hau et al. have proposed to segment a virtual ROI narrowing the anatomical stem crucial to obtain a proper IFOF segmentation from a tractogram (see Figure 5.1b). By analyzing the IFOF resulting from this stem-based segmentation, they have precisely defined some of the IFOF termination regions previously debated. The IFOF stem defined according to these studies is located in the white matter of the ventral third of the external capsule, just medial to the putamen and ventral to the claustrum (Sarubbo et al., 2013).



(a)

(b)

Figure 5.1: (a) Visual comparison of IFOF after ex-vivo dissection and bundle segmentation. The red label in the ex-vivo brain locate the anatomical stem of the IFOF. Adapted from (Sarubbo et al., 2019). (b) Stem-ROI overstimates the anatomical stem. This figure depicts a coronal section of the stem-ROI (manually segmented) and the corresponding points of intersection with streamlines of the IFOF. It can be observed that the voxel mask of the stem-ROI overestimates the anatomical stem of the IFOF.

5.2 StemSeg method

StemSeg is designed as a two steps method as depicted in Figure 5.2: the first step is in charge of the segmentation of the ROI of the stem, namely *stem-ROI*, and a second step is devoted to the segmentation of the IFOF using the *stem-ROI* as additional clause of a segmentation rule.

5.2.1 Basic notation

We define a streamline as a sequence of points $\mathbf{s} = (\mathbf{x}_1, \ldots, \mathbf{x}_L)$, where $\mathbf{x}_i \in \mathbb{R}^3$, $i \in \{1, \ldots, L\}$. A whole brain tractogram is defined as a set of streamlines, $T = \{\mathbf{s}_1, .., \mathbf{s}_N\}$, with $N \sim 10^5 - 10^6$. We denote a bundle as a subset of streamlines from the tractogram, $B = \{\mathbf{s}_1, .., \mathbf{s}_M\}$ where $\mathbf{s}_j \in T, j \in \{1, \ldots, M\}$ and $M \ll N$.

A whole brain volumetric image is denoted as $\mathbf{V}_{i,j,k} = [i] \times [j] \times [k]$. An ROI is defined as a binary mask by a function $\mathcal{V}^{ROI} : \mathbf{V}_{i,j,k} \to \{0, 1\}$. A colored FA (CFA) image is defined as $\mathcal{V}^{CFA} : \mathbf{V}_{i,j,k} \to \mathbb{C}^3$, where $\mathbb{C} = [0, 255]$ denotes the RGB encoding (see Chapter 2.1.2). A slice of a volume is defined as a two dimensional plane, where $\mathbf{V}_{i,j,\bar{k}}$, $\mathbf{V}_{i,\bar{j},k}$ and $\mathbf{V}_{\bar{i},j,k}$ are referring to axial, coronal and sagittal projections respectively.

A segmentation rule is defined as a logical expression of streamlines and ROIs, containing

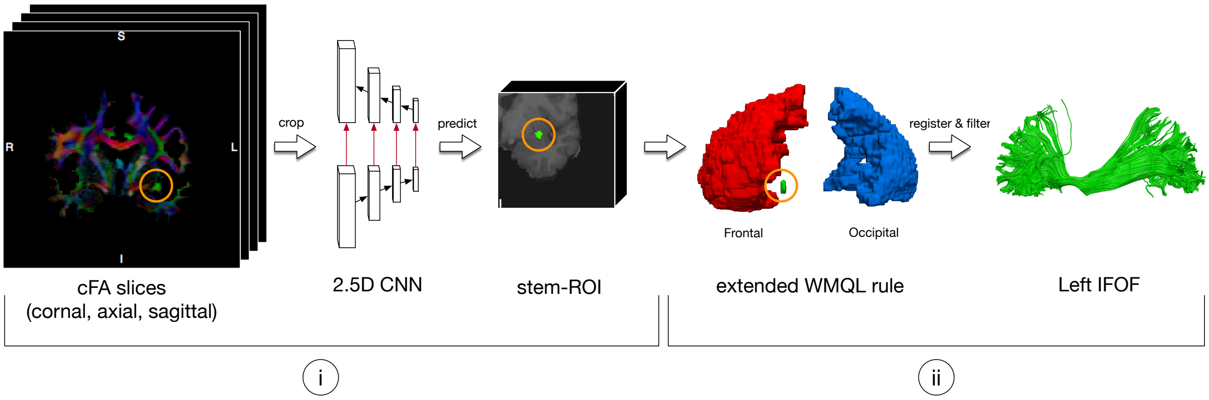


Figure 5.2: StemSeg 2-step procedure.

conjunctions and disjunctions of predicates. Referring to the WMQL syntax (Wassermann et al., 2013, 2016) (see Chapter 3.1.2), we report here two *divisional terms* that we will use in our segmentation rule: $\texttt{endpoints\_in}(\mathbf{s}_i, \mathcal{V}^{ROI})$, which is true when at least one of the termination points $\mathbf{x}_1^i$ and $\mathbf{x}_L^i$ of a streamline $\mathbf{s}_i$ is included in a binary mask $\mathcal{V}^{ROI}$, and $\texttt{waypoint\_in}(\mathbf{s}_i, \mathcal{V}^{ROI})$, which is true when at least one point of the streamlines $\mathbf{s}_i$ is included in $\mathcal{V}^{ROI}$.

5.2.2 Stem-ROI segmentation

Input and output of the CNN Even though the segmentation of the *stem-ROI* can be approached with a 3D CNN, e.g., (Çiçek et al., 2016), such a model presents several computational issues due to 3D operations, which limit the network to learn mostly local (small field of view) features. A common workaround is to reformulate the task as multiple 2D image segmentations where CNNs proved to be effective while memory-efficient. This strategy is known as *2.5D segmentation*. From the original volumetric representation $\mathbf{V}_{i,j,k}$ we randomly sample 2D slices to train a 2D CNN considering the three different projections: $\mathbf{V}_{i,j,\bar{k}}$ (axial), $\mathbf{V}_{i,\bar{j},k}$ (coronal) and $\mathbf{V}_{\bar{i},j,k}$ (sagittal). The output of the CNN is again a 2D slice containing a probability mask. The whole volumetric estimate of the stem is obtained by applying the trained CNN to all the slices of each projection sequentially, obtaining three values per voxel corresponding to the prediction for the three projections. The binary mask of the stem $\hat{\mathcal{V}}^{stem}$ is summarized by merging the three values using a binarized average, i.e., thresholding at 0.5.

CNN architecture The CNN model we used for learning the 2D slice segmentation is a U-net (Ronneberger et al., 2015), which is a famous encoder-decoder segmentation network. The encoder learns latent features by alternating convolutional layers with max pooling, which reduces the image resolution, while the decoder alternates convolutional layers with

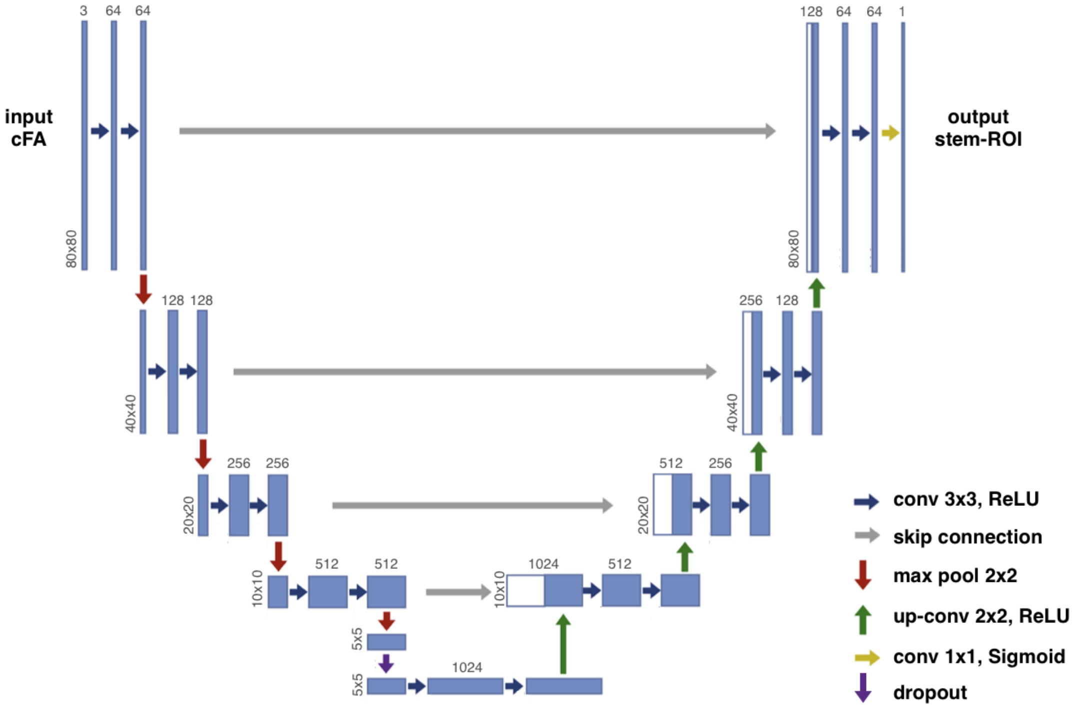


Figure 5.3: StemSeg U-net architecture.

upsampling to reconstruct a prediction mask. Furthermore, the decoder concatenates at each layer features computed in the encoder at the same resolution, giving the network the U shape. The precise network architecture of StemSeg is reported in Figure 5.3.

Choice of the input signal Which kind of MRI volume may enable an effective segmentation of the stem? DWI images are recordings of information on fiber orientation. As explained in Section 2.1.2 the fitting of a DTI model allows the estimation of the components of three main directions of diffusivity within each voxel as a colored fractional anisotropy (CFA) image: red for lateral, green for anteroposterior, and blue for vertical. Since the anatomical stem of the IFOF is a waypoint where all the fibers have a homogenous fronto-occipital orientation, we may assume that bright green color will encode such a region in a CFA image 5.6. According to this assumption, we sampled 2D slices from all the axis of $\mathcal{V}^{CFA}$.

5.2.3 Stem-based segmentation of IFOF

In the second step of StemSeg, we segment the IFOF using a WMQL rule which extends the formal definition of the IFOF with the stem-ROI.

According to recent neuroanatomy studies (Sarubbo et al., 2013; Hau et al., 2016) the formal definition of the IFOF can be encoded as a logical expression with the WQML (Wassermann et al., 2016) to select the streamlines that belong to the IFOF. The parcellation of

```
IFOF.left = (
   endpoints_in(
      frontal_inferior_lobe.left
      OR frontal_middle_lobe.left
      OR frontal_superior_orbitalis.left)
   AND endpoints_in(
          occipital_lobe.left
          OR cuneus
          OR lingual
          OR parietal_superior
          OR precuneus)
   AND waypoints_in(stem.left))
frontal_inferior_lobe.left = (
   frontal_inferior_operculus
   OR frontal_inferior_orbitalis
   OR frontal_inferior_triangularis)
occipital_lobe.left = (
   occipital_superior
   occipital_middle
   occipital_inferior)
```

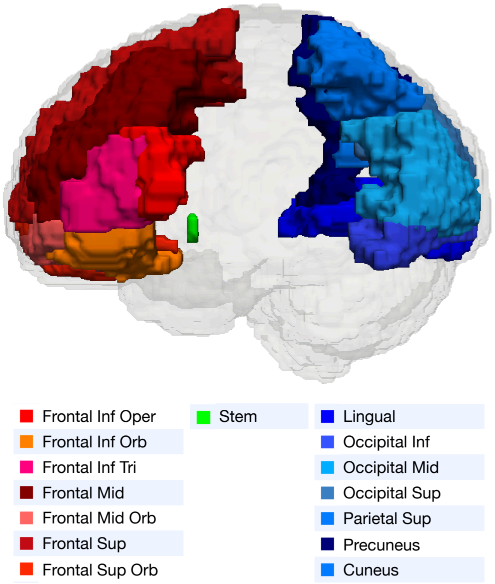


Figure 5.4: WMQL formal segmentation rule of left IFOF (black) extended with the stem (blue). The regions used refer to AAL2 atlas (Rolls et al., 2015) and are illustrated on the right.

ROIs with the termination points of streamlines can be obtained from an atlas after the normalization to a standard space. In particular, the termination points of streamlines of the IFOF should be located in several regions of the occipital and frontal lobes. The exact combination of these regions is reported (with black color) in the formal WMQL rule of Figure 5.4.

Even with an accurate ROI parcellation, the formal definition of IFOF is not enough to exclude many false positive streamlines, usually a by-product of sub-optimal tracking algorithms (see example in Figure 5.5). For this reason, we propose to refine the formal rule by imposing an additional constraint: at least one point of the streamlines of IFOF should belong to the ROI segmented as the stem of the bundle, i.e. *stem-ROI*. The conjunctive clause of `waypoint_in` is combined with the `endpoints_in` conditions (blue line in Figure 5.4). Using the stem as a waypoint condition prevents the usual manual effort of defining exclusion regions to filter out false positive streamlines. An example is reported in Figure 5.5.

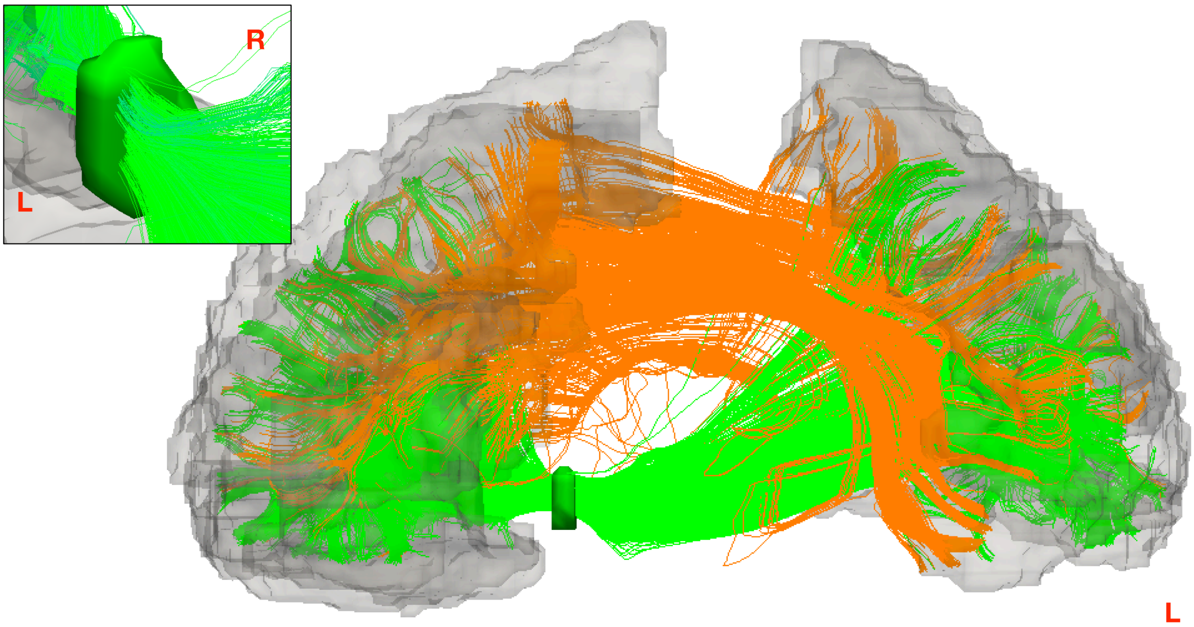


Figure 5.5: Example of segmented IFOF with stem-based segmentation (green) and the false positive streamlines (orange) not filtered out by considering only the regions of termination points. In the box the anatomy of stem.

5.3 Material

HCP diffusion We based our empirical analysis on the Human Connectome Project (HCP) (Van Essen et al., 2013) diffusion dataset. It is a public and widely adopted dataset where the acquisition and pre-processing pipeline for diffusion MRI is carefully validated (Milchenko and Marcus, 2013; Sotiropoulos et al., 2013; Glasser et al., 2013). The dataset comprises healthy subjects (all genders) aged between 24 and 35. Each subject has both the structural T1 image and the 3T Diffusion Weighted Image (DWI). The DWI has resolution 1.25mm with 270 gradients multi-shell (Andersson and Sotiropoulos, 2016), and is corrected for eddy currents (Andersson and Sotiropoulos, 2015).

HCP-low diffusion To quantitatively validate the robustness to the different image resolutions of StemSeg, we created a low-resolution version of HCP. This *HCP-low* was designed to emulate the clinical/real-world quality of diffusion images, which is often at low resolution. To generate low-resolution HCP images, we downsampled the diffusion signal of a factor of 2, passing from 1.25mm to 2.5mm. Although we are aware that the resolution decrease alone is not enough to emulate clinical quality data properly, we believe that an additional low-resolution validation dataset might still provide insights about the robustness of the proposed model.

Tracking We designed a pipeline to compute the diffusivity model both using DTI (Garyfallidis et al., 2014) and CSD (Tournier et al., 2007), after extracting 90 gradients at b=2000. From the DTI, we derived FA and CFA, while from the CSD combined with the deterministic local tracking algorithm of DiPy (Garyfallidis et al., 2014) (previous to be updated to EuDX) we generated the tractograms (uniform seeding in the voxels of white matter mask, step size 0.625mm). We sampled such data from the (temporally) last 158 acquired HCP subjects, including test-retest subjects (Buchanan et al., 2014).

ROI-based dataset Using the 158 HCP subjects, we built a dataset of labeled stem-ROIs along with the corresponding IFOF bundles. The labeling of the stem-ROIs was supervised by a neuroanatomist (S.S)., and followed the procedure used in (Hau et al., 2016). The procedure is composed of three steps: (i) streamline filtering using the termination regions derived from Hau et al. to the Automated Anatomical Labeling 2 (AAL2) atlas (Rolls et al., 2015), (ii) manual definition of ROIs to exclude false positive streamlines, until a satisfactory IFOF segmentation is achieved, (iii) manual segmentation of the stem-ROI in three consecutive coronal slices of the CFA (see Appendix A.1.1 for a visual example of the segmentation procedure). As a result, we obtained 316 stem-ROIs (step (iii)) bound to 316 segmented IFOF (step (ii)), half of them in the right hemisphere and the other half in the left hemisphere. We call this collection of labeled data, *ROI-based* dataset. We report additional stem-ROIs group statistics, such as average size and FA value, in Appendix A.1.2. The labeled stem-ROIs are available at `https://doi.org/10.25663/brainlife.pub.8`

Bundle-based dataset *ROI-based* segmentations, even when performed by experts, are not able to avoid all the false positive streamlines due to the coarse granularity of voxels. Therefore, we considered operating an additional manual IFOF segmentation using a method without the bias of ROIs, which we refer to as *bundle-based*. A second expert neuroanatomist (A.D.B.) segmented the IFOF on a subsample of 30 out of the 158 subjects using Tractome [1] (Porro-Muñoz et al., 2015). This tool gathers together streamlines based on their trajectory similarity into bundles/clusters. One streamline representative of each cluster is visualized to simplify the visual scene of the tractogram. The expert can decide which cluster to maintain and which to delete by directly picking and removing clusters. The segmentation procedure is executed iteratively, starting from a few hundred clusters to represent the whole brain tractogram. At every iteration, only the clusters of interest are maintained and exploded into new clusters more fine-grained. We report the difference between the ground-truth (GT) IFOF segmented with a *bundle-based* and *ROI-based* strategy both qualitatively in Figure 5.7a and 5.7b, and quantitatively by means of the volumetric DSC (see Section 5.4.2), obtaining 93.9 ± 3.7%. We report in Appendix A.1 a more detailed comparison of the two reference sets in terms of IFOF termination points.

[1] `https://github.com/FBK-NILab/tractome`

APSS dataset As last, we adopt a clinical dataset for the purpose of validating the generalization ability of our model. The dataset has been obtained from the Santa Chiara Hospital (APSS) in Trento (Italy). It comprises of 5 subjects affected by brain tumors. For each subject, we have available the DWI and the reconstructed tractogram, but not the reference IFOF segmentations. The DWI was acquired with a 1.5T MR scanner using 60 directions. Then, a single shell b=1000 s/mm^2 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 streamlines. We remark that using this dataset we can have a qualitative evaluation of the generalization ability of our model trained on HCP data. The main differences are data source and pre-processing (HCP vs. APSS), diffusion image resolution (2mm vs. 1.25mm), number of directions and b-values (90g b=2000 vs. 60g b=1000), and subjects condition (healthy vs. patients).

5.4 Experimental analysis

5.4.1 Stem-ROI prediction

We designed an empirical analysis to assess how effective an automatic stem-based segmentation method, like StemSeg, can be in the case of the IFOF bundle. We sampled 128 HCP subjects as training set of the CNN, with a further split into 114/14 to perform model selection. The training was iterated for 500 epochs with learning rate 1e-3, using data augmentation and batches of 47 2D slices, randomly sampled from different subjects. We treated indistinctly left and right stem-ROIs, by feeding the network with images containing only one of the two each time. To separate left and right stem-ROIs the CFA images were cropped from the original size (145x174x145) to isotropic volumes (80x80x80), knowing that by definition, the stem of the IFOF is located in the inferior, anterior, and lateral portion of each hemisphere of the brain. Note that these values refer to HCP 1.25mm input, but the cropping can be applied to every brain volume that is Anterior Commissure - Posterior Commissure (ACPC) aligned, i.e., rigidly registered with the standard MNI space. In the training procedure, we optimized the binary cross-entropy loss with weighting correction to better manage the class imbalance.

We tested the trained network over the remaining 30 subjects of the dataset, i.e., the subjects for which we have both *ROI-based* and *bundle-based* IFOF ground-truth. We evaluated the stem-ROI prediction using two measures: the mean displacement of the center of mass computed along the three axes and the mean volumetric recall ($TP/TP + FN$). The first measure aims to assess whether the CNN has a directional bias in the segmentation, while the second measure is informative for the segmentation task, where a stem underestimation can lead to the loss of many IFOF streamlines.

The obtained results in terms of the mean displacement of the center of mass along the three axes are: 0.33 mm (lateral), 0.82 mm (anteroposterior), 0.26 mm (vertical). An

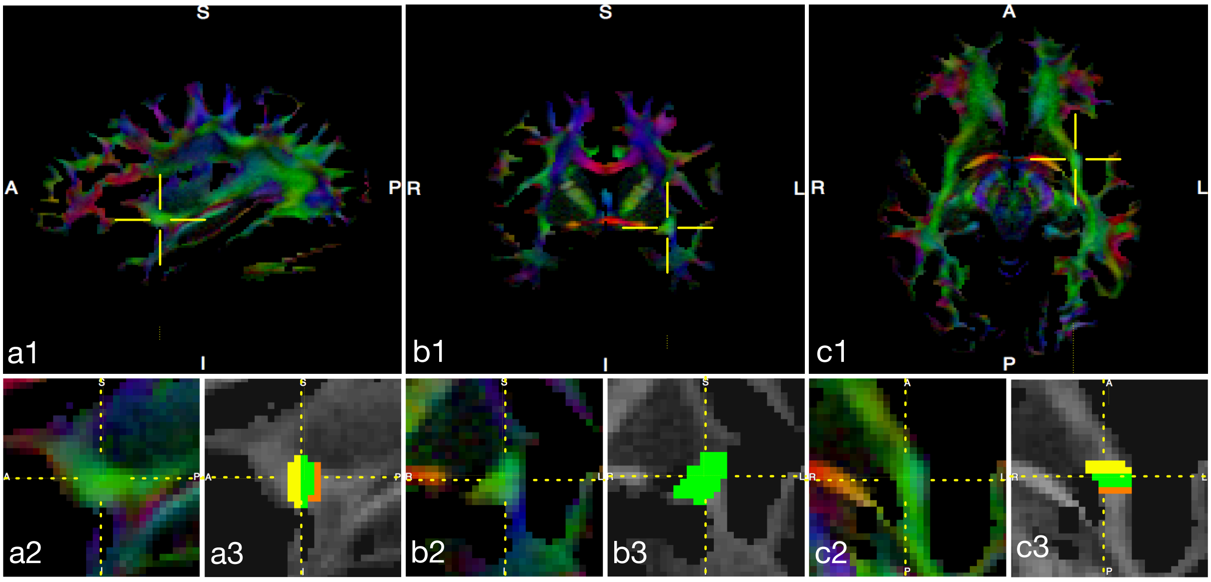


Figure 5.6: An example of stem detection in the CFA. The three panels show the sagittal, coronal and axial view of the center of mass of the stem. Each small panel reports the detailed view of the stem region (upper) and the overlap between the labeled and predicted stem (lower): true positive (green), false negative (orange), and false positive (yellow).

visual example of the center of mass displacement is shown in Figure 5.6. In the figure the lateral misplacement can be observed by looking at frame b3 and c3, the anteroposterior is in frame a3 and c3, and the vertical is in frame a3 and b3. The mean volumetric recall of the stem-ROI is 74.9±18.0.

5.4.2 IFOF segmentation

The IFOF segmentation of the 30 test subjects was performed by applying the rule in Figure 5.4 with regions taken from AAL2 atlas (Rolls et al., 2015), and co-registered to each subject through the non-linear ANTs SyN algorithm (Avants et al., 2008). According to the best practice in the literature, we evaluated the segmentation results by computing the bundle volumetric mask from the streamlines. The comparison between expert segmentation and automated segmentation was computed as overlapping of the voxel masks in terms of the Dice Similarity Coefficient (DSC):

$$\mathrm{DSC} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} = \frac{2 * \mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{5.1}$$

In addition, for the sake of completeness, we also computed the DSC on streamlines, which can quantify the actual number of misclassified streamlines. This measure is computed with the same equation of (5.1) but considering the fibers instead of voxels as units. This metric may provide a more fine-grained evaluation of the results.

A first comparison was carried out with respect to the IFOF segmented according to the *ROI-based* approach to test the accuracy of our automated stem-based method. A second comparison addressed the question of how robust is the proposed method with respect to a second ground-truth produced by a second expert with a non-ROI-based procedure, i.e., the *bundle-based* GT. The results in terms of volumetric DSC for the two GTs are reported in Table 5.1 (second row). The overlapping of streamlines for the ROI-based is 98.7 ± 0.9 and for the bundle-based 96.1 ± 2.3.

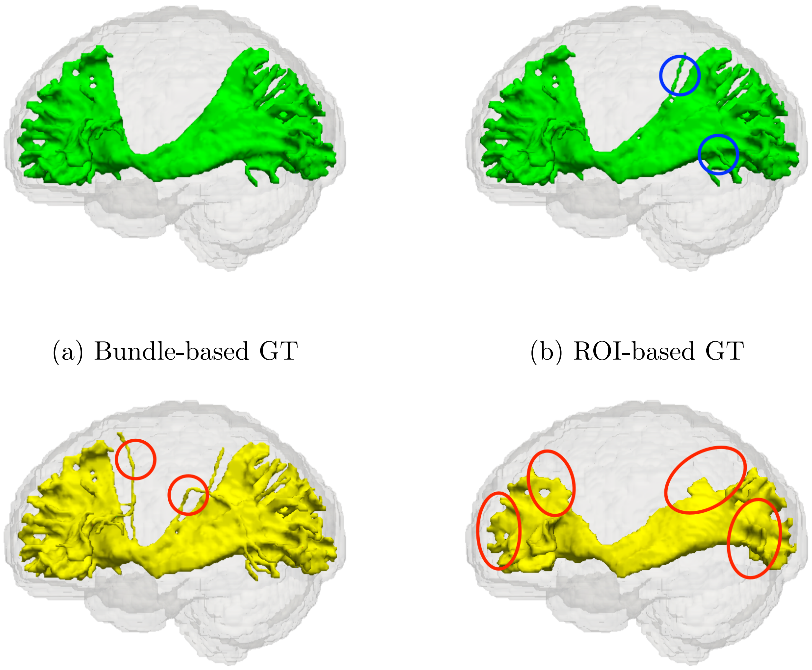
Table 5.1: Mean and standard deviation of volumetric DSC (reported in %) over 30 HCP test subjects. Both IFOF right and left are aggregated in the score computation.

Method	DWI res.	ROI-based	Bundle-based
TractSeg	1.25mm	61.0 ± 5.6	62.1 ± 5.7
StemSeg	1.25mm	**98.8 ± 1.0**	**93.8 ± 3.5**
	2.5mm	96.4 ± 4.1	93.1 ± 3.9

An additional experiment was designed to compare StemSeg with TractSeg (Wasserthal et al., 2018), the state of the art method for bundle segmentation. TractSeg casts the bundle segmentation task as a problem of volume segmentation by encoding the streamlines representation of a tract in the corresponding voxel binary mask. The segmentation is performed by training a CNN following a 2.5D approach based on U-net in analogy to what we presented here. However, unlike StemSeg, in TractSeg, the CNN is fed with a richer fiber orientation map, where the three main direction peaks extracted from the fODFs (spherical harmonics) are encoded in each voxel, resulting in a 9-dimensional volumetric image. We trained TractSeg adopting the default parameters reported in Wasserthal et al. (2018), using the same dataset split of the experiment above. For both TractSeg and StemSeg we used the same code implementation, available on the BrainLife platform (Avesani et al., 2019) at `https://doi.org/10.25663/brainlife.app.205`.

The results of the IFOF segmentation with StemSeg and TractSeg on ROI-based and bundle-based datasets are summarized in Table 5.1, and a qualitative example is reported in Figure 5.7. Note that we limit the comparison of the two methods to volumetric DSC only as the streamlines DSC cannot be computed for TractSeg. The result on volumetric DSC shows a large positive gap in favor of StemSeg, with +38% when considering the ROI-based GT, and +32% for the bundle-based GT.

Lastly, we investigated the reliability of our method in the case of clinical data quality, which is usually worse compared to a curated dataset such as HCP. We designed a qualitative study on the APSS clinical dataset, inspecting the IFOF segmentation of 5 patients affected by tumors. In this experiment, the segmentation of the stem-ROI has been possible after a rigid registration of the patients' brains to ACPC, but without any re-training of our CNN. Moreover, since no ground-truth segmentations were available for the



(a) Bundle-based GT

(b) ROI-based GT

(c) StemSeg prediction

(d) TractSeg prediction

Figure 5.7: Qualitative comparison of voxel masks of the two manual segmentation procedures (green) and the two automated segmentation methods (yellow) for a left IFOF. The blue circles highlight the anatomical differences, while the red highlights prediction errors.

APSS dataset, we designed a quantitative study by emulating a low-quality data scenario starting from the HCP acquisitions — a similar setting is also presented in (Wasserthal et al., 2018). The 3T DWI was downsampled from 1.25mm to 2.5mm and then used to compute a new *low resolution* CFA. Next, we cropped and upsampled the CFA to obtain a shape acceptable by our CNN, which was trained on *high resolution* CFA. Then, we used the predicted stem to segment the IFOF from the same tractograms used in the previous HCP-based experiments. Note that we did not compute the HCP-low segmentation score also for TractSeg as we expected no improvement on lower resolution — as Wasserthal et al. show in their paper. We presented the experiments on HCP-low only to provide a quantitative measure of the robustness of our model.

The result of the low-resolution IFOF segmentation is reported in the last row of Table 5.1. The qualitative segmentation of the patients is depicted using two different visualizations. In Figure 5.8 we show, in the top part, the detailed anatomy of an IFOF segmented in the case of a large tumor in correspondence of the occipital terminations. In the bottom part, we visualize two additional segmentations of patients having the tumors located nearby the IFOF frontal terminations. Lastly, in Figure 5.9 we provide the overall segmentations of the five APSS patients.

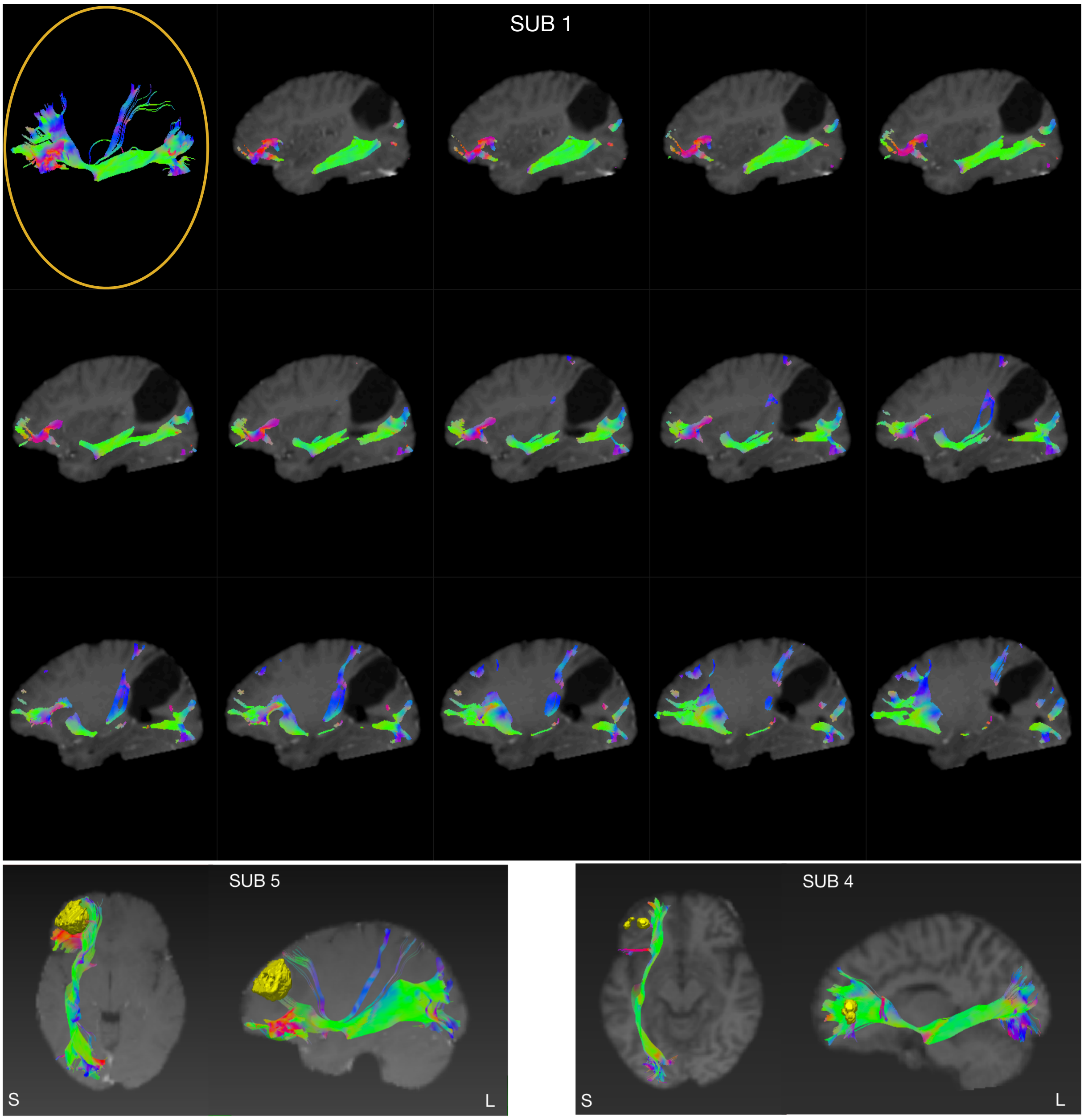


Figure 5.8: Example of IFOF segmentation in patients with tumor in correspondence of the occipital (SUB 1) or frontal (SUB 5, SUB 4) IFOF termination regions. (SUB 1) We show the entire IFOF segmented on the top left image, while the remaining images show the IFOF mapped within the different sagittal slices of the T1w (with contrast agent). The black oval hole present in the T1w is the tumor. As observable, the segmented IFOF tightly surrounds the tumor area. (SUB 5, SUB 4) A yellow 3D ROI depicts the tumor. In SUB 5, the tumor apparently pushes the *frontal superior* termination of the IFOF towards the longitudinal fissure. In SUB 4, the tumor mass hinders the reconstruction of part of the *frontal inferior* IFOF terminations, which are missing in the segmented bundle.

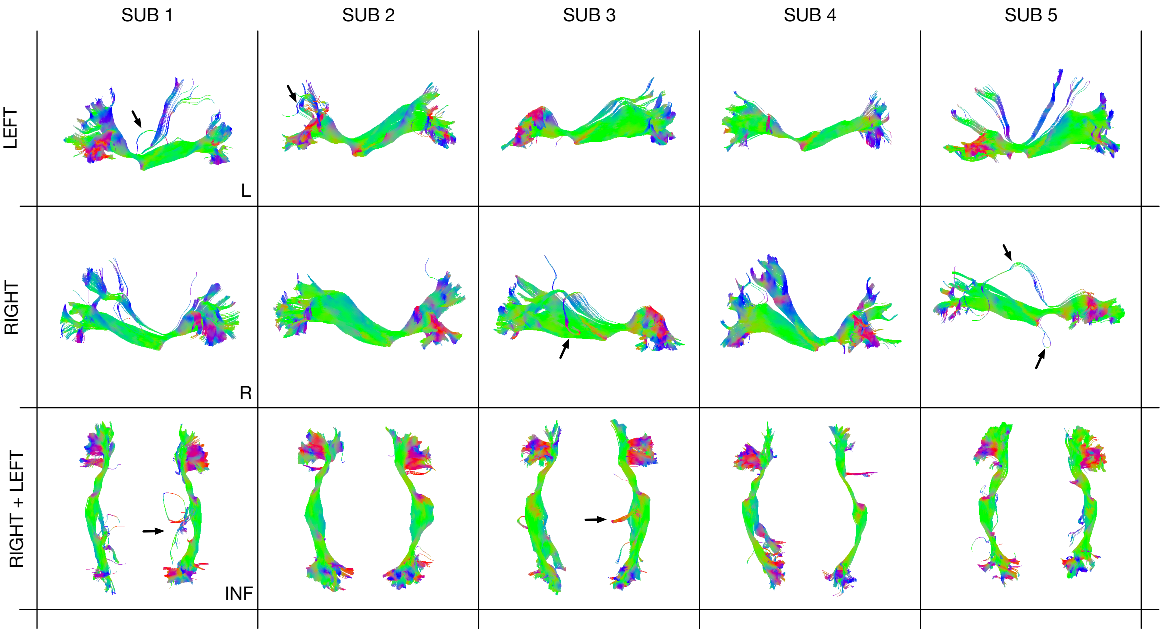


Figure 5.9: Qualitative IFOF segmentation of the 5 patients in APSS dataset. For each patient (column) we show the segmented left and right IFOFs (rows). Overall, the IFOF is successfully segmented in all the patients. However, we observe the presence of a some *false positives* fibers visually identifiable in SUB 1-L, 2-L, 3-L-R, and 5-R.

5.5 Discussion

According to the empirical results reported in Table 5.1 we may claim that the automated segmentation of stem-ROI is a viable and effective method for IFOF segmentation.

5.5.1 Using CNNs to segment small waypoints ROIs

A targeted analysis of the error in segmenting the ROI of stem reveals that it is not isotropic. The displacement of the center of mass with respect to the correct location is negligible both in vertical and lateral directions (see Figure 5.6b3). Instead, the displacement error in the anteroposterior direction is higher (see Figure 5.6a3 and 5.6c3), and it is responsible for the high variance in the recall score. Luckily, this distribution of the displacement error does not impact the accuracy of IFOF segmentation. The anteroposterior shift reflects the variance that the expert has during the manual labeling of stem-ROIs — the three manually segmented coronal slices may not be uniques. Such variance is mostly due to the similar-planar approximation that the stem-ROIs does of the anatomical stem, which instead is more extended in the anteroposterior direction, as can be noted in the ex-vivo dissection of Figure 5.1b.

In light of the previous discussion, we may consider satisfactory the results obtained on stem-ROI segmentation. The adoption of a CNN to segment the colored FA image of the brain has proved to be a correct choice. In particular, the use of a 2.5D training strategy compared to a more basic 2D strategy turned out to be decisive. In our earlier investigations with 2D strategies (see Appendix B.1), we needed heuristics to select which predicted slices to use for the generation of the final stem-ROI, and the obtained selection resulted sometimes sub-optimal. Also, the use of coronal projections only in the 2D strategy might limit the possibility for the CNN to better understand the contextual information of the stem-ROI, which is present in sagittal and axial views.

5.5.2 Automatic stem-based bundle segmentation

ROI-based segmentation tends to have more false positive streamlines while *bundle-based* segmentation allows the pruning of all outlier streamlines (see Figure 5.7b and 5.7a respectively). StemSeg is performing well (see Table 5.1), even with respect to a manual and time demanding segmentation. In particular, if we compute the DSC on fibers (instead of voxels) of the segmented bundles we obtain 96.1% of overlapping with *bundle-based* IFOF. This means that in addition to segmenting a bundle having the correct envelop i.e. bundle mask, we only missclassify a very small portion of fibers.

StemSeg significantly outperforms TractSeg as reported in Table 5.1. The large gap between the volumetric DSC measures is mainly motivated by two reasons. First, the two models produce two different output representations: streamlines for StemSeg and voxels for TractSeg. Second, CNNs tend to better perform in the segmentation of rather regular shapes like the stem-ROI compared to quite irregular shapes like bundle masks. Figure 5.7c and 5.7d, show an example of the IFOF anatomy dissected by StemSeg and TractSeg respectively. Despite using the same examples of StemSeg, TractSeg is missing the portion of superior regions, both frontal and occipital, where streamlines are fanning out. Apparently, this pattern can not be properly detected with a voxel-based representation. A similar claim is also reported in the study of (Bertò et al., 2021). The main argument is that the inductive bias of CNNs especially given by discrete pooling filters induces the learning model to be invariant to local variations present in the bundle fanning out of different individuals. The impact of such a bias is quantified in (Bertò et al., 2021) by means of volumetric *fractal dimension* (FD) of the output segmentations; a measure that quantifies the irregularity of 3D shapes. The analysis in (Bertò et al., 2021) reveals a systematic difference between the FD of the bundles segmented by TractSeg compared to the reference bundles, with the former being more fractals i.e., less irregular.

Given the results of the application of StemSeg to low quality data we may claim a very good robustness of our model. In the case of simulated low quality data we face the challenge of a lower resolution input, but without altering the (high) quality of the data source i.e. the HCP rigorous and well-tuned acquisition protocol. In Table 5.1 we observe a very small drop of volumetric DSC with respect to the (original) high resolution segmentations. Interestingly, we have the smallest drop when we compare to the *bundle-based* ground-truths. This suggests an excellent preservation of segmentation quality albeit the use of low resolution CFA maps.

In the case of clinical APSS data instead we have a threefold challenge given by the low resolution data, the low quality data source i.e. 1.5T acquired in time-restricted clinical scenario, and the presence of additional noise in the MR signal due to tumors. Despite this tough scenario, the segmentation examples illustrated in Figure 5.8 are all successful. In SUB 1 the large volume of the tumor severely alters the trajectory of the IFOF in the occipital terminations. We may consider anatomically correct the segmentation of streamlines that circumnavigate the tumor mass, according to existing studies showing a similar behavior in the same circumstances of DTI tractography and low-grade glioma (such as the one of SUB 1) (Campanella et al., 2014; Yeh et al., 2021). By looking at SUB 5 we observe a similar situation, but in this case the alteration is less severe. Instead, in SUB 4 the IFOF is missing a portion of the frontal inferior terminations, which would be in correspondence of the tumor area. In this case, we may attribute most of the responsibility to the tractography pipeline since a more detailed inspection of the input tractogram revealed the lack of such streamlines — probably due to the strong signal noise caused by the tumor.

Overall, considering the APSS segmentations in Figure 5.9 we may confirm a good robustness of our StemSeg. We suggest that one of the reason might be the little reliance on linear or non-linear co-registrations. We do not need any of them for the critical stem-ROI segmentation, while we use them only to co-register the large frontal and occipital regions, which are less sensitive to small alignment errors. However, in Figure 5.9 we notice the presence of some false positives streamlines, which are clearly artifacts of the tractography pipeline. Although we use an anatomically sound segmentation rule, the use of ROIs to characterize the anatomy is not able to filter all the anatomically non-plausible fibers. Indeed, as visible in Figure 5.1b there is a mismatch between the anatomical definition of the stem and the corresponding stem-ROI given by the different data representations. We argue that a better approach to detect false positive fibers should cast the learning problem directly on the fiber trajectory — as we do with Verifyber described in the next chapter.

Chapter 6

Tractogram Filtering with Geometric Deep Learning

In this Chapter, we present our main contribution *Verifyber*, which tackles the open questions of RQ2. In the following sections we first introduce the main idea of the work, and describe the existing labeling strategies for the extensional encoding of white matter anatomical plausibility (see 6.1.1). Next, we present our Geometric Deep Learning model in 6.2), and a comprehensive discussion of the alternative model to perform learning on streamlines (6.3). Finally, we report the material and a large set of experiments in 6.4 and 6.5, respectively. We discuss the results in the subsequent 6.6. The presented results refer in part to (Astolfi et al., 2020b) published at the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2020. But, they have been extended to a large extent for submission to a journal.

Synopsis The purpose of this contribution is to leverage the brain anatomical knowledge to design a method for tractogram filtering based on the notion of anatomically plausible and 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. The proposed method may substantially reduce the amount of noise in tractograms, enabling more accurate tractography-based analysis as well as neuroanatomical investigations.

6.1 Introduction

Tractograms are, to date, the most detailed in-vivo representation of the brain white matter. They are obtained through complex computational pipelines that culminate in tracking millions of virtual fibers, which approximate the white matter pathways. However, despite such fibers might be highly informative of the WM structural connectivity, they may also be *false positives* of the tracking, having no correspondence in the WM anatomy, i.e., *anatomically non-plausible*. Unfortunately, according to multiple recent evaluation studies (Maier-Hein et al., 2017; Nath et al., 2020; Schilling et al., 2019b), false positives often dominate tractograms.

One possible solution is to post-process tractograms by filtering out fibers based on anatomical plausibility criteria. This task is known as *tractogram filtering*. Unfortunately, the current knowledge of WM anatomy does not permit a complete definition of what is plausible and what is non-plausible. The design of a loss function suitable to capture the notion of anatomical plausibility remains an open challenge. For this reason, state of the art methods (see 3.2) adopt *unsupervised* and *intensional* criteria either based on the signal explainability of fibers and/or fibers consistency in local bundles or across individuals. However, such unsupervised criteria only capture limited anatomical plausibility aspects, often leading to suboptimal removal of false positives. They lack explicit anatomical priors such as extensional fiber labeling, which would guarantee anatomically sound filtering.

In this work, we propose Verifyber, a novel tractography-based method to perform filtering of non-plausible fibers from a tractogram. Unlike previously done in the literature, the task of tractogram filtering is shaped as a supervised learning problem. A binary classifier takes in input a fiber and outputs 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),. GDL better fits the learning with 3D data without forcing Euclidean vector representations. The notion of anatomical plausibility is not hardcoded in the loss function but derived from fiber examples, labeled either as anatomically plausible or anatomically non-plausible.

Regardless of the labeling source, the challenge of supervised learning is to train a binary classifier based only on a digital representation of fibers. Choosing 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 dealt with the constraint of learning algorithms requiring a fixed-length embedding. The most common solutions are the computation of a Euclidean embedding such as dissimilarity representation (Olivetti and Avesani, 2011; Bertò et al., 2021). Unfortunately, these fiber embeddings are lossy.

To overcome these limitations of traditional learning models and to preserve the full geometrical information encoded in the fiber pathways, we propose to investigate the use of Geometric Deep Learning (GDL) 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 operators whose combination allows a model to perform convolution in a non-grid 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 handle traditional fibers encoding as ordered sequences of a variable number of 3D points. The proposal extends the existing Edge Convolution (EC) layer (Wang et al., 2019) to take into account the information encoded by the edges between two subsequent 3D points in a fiber. The architecture comprises a global pooling layer that compresses each fiber to a single descriptor and by a Multi-Layer Perceptron (MLP) to discriminate between the two categories, either anatomically plausible or non-plausible. Unlike 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 to prove the properties of the proposed model and assess 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 to different types of tractography and equally effective on *inclusive* and *exclusive* policies to elicit the notion of neuroanatomical plausibility. An additional comparison aims to show the different behavior of supervised and unsupervised filtering approaches such as 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 method's behavior when the labeling of fiber is dynamically evolving over time.

6.1.1 Ground truth anatomy: inclusion vs. exlusion policies

We envision the task of elicitation of brain knowledge as binary labeling of fibers. Despite the evolutionary nature of the knowledge of the human brain, we may conceive two main labeling policies: *inclusive* and *exclusive*. The *inclusive* policy leans to be more conservative and aims to prevent false positives. According to this bias, only fibers following the pathways of well-known bundles are labeled 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 labeled anatomically non-plausible. It is out of the scope of this work to establish which policy might be more effective and appropriate. Instead, our goal is to investigate whether the proposed method is equally robust for the two policies.

Our empirical analysis is considering datasets labeled with both *inclusive* and *exclusive* policies. As a reference example of *inclusive* policy, we point out 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., 2019). In this work, the notion of non-anatomical 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 labeled as anatomically non-plausible, usually half portion of fibers. This declarative knowledge can be applied to any tractogram enabling the annotation of training and test sets for learning purposes.

6.2 Verifyber 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, and the base of our contribution *sequence EC*.

6.2.1 Edge convolution layer

Considering a point cloud $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, $\mathbf{x}_i \in \mathbb{R}^3$, 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, $\text{knn}(\mathbf{x}_i) = \{\mathbf{x}_{j_{i_1}}, \ldots, \mathbf{x}_{j_{i_k}}\}$, using the Euclidean distance as metrics (see Figure 6.1a). The result is a k-nn graph composed of $\mathcal{V}$ nodes and $\mathcal{E}$ edges:

$$\mathcal{G}(\mathcal{V}, \mathcal{E}),\ \mathcal{V} = \mathcal{X},\ e_{ij} \in \mathcal{E} \colon \mathbf{x}_i \to \mathbf{x}_j \iff \mathbf{x}_j \in \text{knn}(\mathbf{x}_i). \tag{6.1}$$

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

$$\mathbf{e}_{ij} = h_{\mathbf{\Theta}}(\mathbf{x}_i \oplus (\mathbf{x}_j - \mathbf{x}_i)), \tag{6.2}$$

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. $\mathbf{x}'_i = \text{pool}(\mathbf{e}_{ij})$, $j \colon (i, j) \in \mathcal{E}$, where pool is either max or mean.

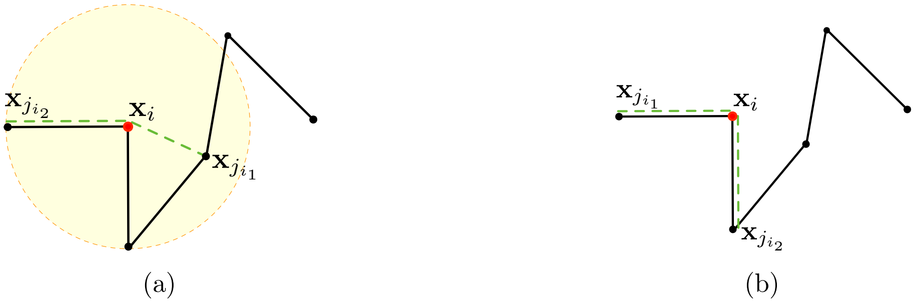


Figure 6.1: Comparison between Euclidean k-nn (a), and graph k-nn on the streamline (b).

6.2.2 Sequence edge convolution layer

A remarkable property an 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 order 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 induces a new input graph structure, with a graph-based k-nn (see Figure 6.1b) that instead preserves an 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:

$$\mathcal{G}(\mathcal{V}, \mathcal{E}'),\; e'_{ij} \in \mathcal{E}' \colon \mathbf{x}_i \to \mathbf{x}_j, \iff j = i + 1 \vee j = i - 1. \tag{6.3}$$

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.

6.2.3 Verifyber model

The Verifyber (VF) model is characterized by the stacking in depth of one sEC layer with one or more EC layers. Specifically, referring to the architecture adopted for our experiments in Figure 6.2, 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 dynamical adjustment of the local neighborhood of points guided by the semantics of the task at hand. Then, $\mathcal{X}'$ and $\mathcal{X}''$ are concatenated, encoded to 1024 features with a learning layer $g_{\mathbf{\Phi}}$, and pooled to obtain a single descriptor of the whole point cloud,

$$\mathbf{z} = \text{pool}(g_{\mathbf{\Phi}}(\mathcal{X}' \oplus \mathcal{X}'')). \tag{6.4}$$

Finally, the 1024d 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).

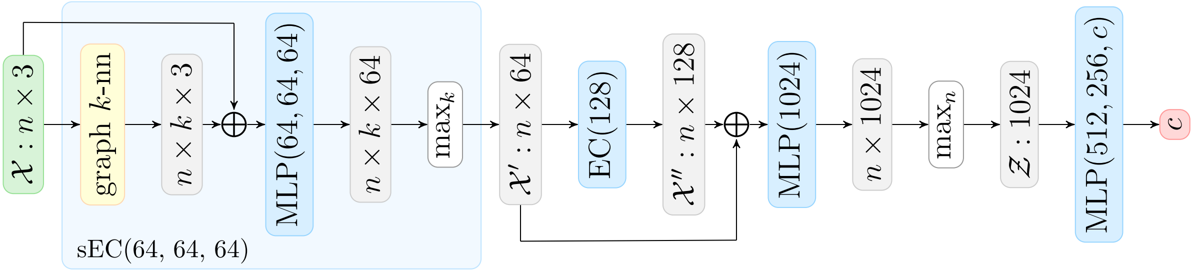


Figure 6.2: Verifyber architecture. Green, gray, and red blocks represent input, intermediate, and output tensors, respectively. Parametric layers are colored in blue, while fixed layers in white. In yellow we highlight the graph k-nn that allows the model to be sequence sensitive.

6.3 Streamline representation learning

The proposed Verifyber model is the result of a step-by-step investigation that arises 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 most common neuroimaging data like images and volumes: (i) the length of a streamline is variable, (ii) a streamline is a sequence of points without orientation.

Streamline embedding and traditional models These properties prevent many learning methods from being directly applied to streamlines. One common requirement of learning methods is to have fixed length vectors as input, and thus existing works resorted to different preprocessing solutions to match that 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 many methods, especially traditional machine learning algorithms like support vector machines or linear classifiers, the resulting 2-dimensional vectors of fixed size (n_points, 3) 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) or more recently (Bertò et al., 2021), which combines it with a set of handcrafted features based on the white matter anatomy. 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.

CNNs for streamlines Such a limitation might be bypassed using deep learning techniques as they are able to learn embeddings directly from the input data. 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 Analisys (deepWMA) (Zhang et al., 2020), instead, train standard CNN models like (Krizhevsky et al., 2012) and (He et al., 2016) directly on streamlines using bundle supervision. In this way they learn an embedding optimized with respect to the bundle segmentation task.

However, we notice a controversial use of convolutional filters in FiberNet and DCNN as they treat a streamline of fixed size (n_points, 3) like it is an image of size (height, width). As operated in FINTA and DeepWMA, the proper use of convolution considers the three component (x, y, z) of a streamline to be channels like (r, g, b) in images so that different filters are learned for each channel. We may also notice a general drawback concerning the adoption of standard CNNs to perform learning on the streamline structure. Indeed, the success of CNNs is strictly related with their translation invariance (or equivariance) property that is crucial for the image domain. In other words the grid-based structure of images allows the parameter sharing of spatial convolution filters. However, streamlines do not have a grid structure, and thus the translation invariance/equivariance of CNNs is not a crucial requirement, as it is for images. Streamlines, instead, require flip invariance as they are unoriented sequences of points, but unfortunately this is not a property of standard CNNs despite it is neglected by the approaches mentioned above

RNNs and GDL Based on these premises we investigate supervised deep learning approaches different from standard CNNs. We seek for NN 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 to deal 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, like point clouds and graphs. To deal with such data, GDL models perform convolutions and pooling that are permutation invariant instead of translation invariant and that can be applied to batches of size-varying samples.

bLSTM (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 e.g., concatenation, and then forwards it to a FC network, which performs classification.

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. However, 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. We may approximate a streamline to be a point cloud by neglecting its sequential structure as point clouds do not assume any order in the points. Despite such a representation loses sequentiality it maintains both the streamline spatial information and its invariance to the flipping of orientation.

PointNet (Qi et al., 2017) In our experiment we investigate the precursor and most adopted NN for point cloud, namely PointNet (PN) (Qi et al., 2017) (see Chapter 2.2.2). 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 perming the output classification.

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).

DGCNN (Wang et al., 2019) The DGCNN model is, according to Wang et al. a generalization of PN. Instead of considering a single 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 6.2.1) 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 DGCNN and PN, resulting able to work with size-varying input and to consider point relations.

6.4 Material

In this section we present the datasets used for the experiments. A summary is reported in Table 6.1. The name adopt for the datasets reflect the data source, the type of labeling (E for exclusive and I for inclusive policies), and the initial letter of the first author of the labeling policy.

Table 6.1: 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
HCP-EP[*]	HCP	Exclusive Petit et al. (2019)	-	20	1M	CSD PF-ACT	3T DWI, 1.25mm, 270g multi-b=(1,2,3)K
HCP-IZ[†]	HCP	Inclusive Zhang et al. (2018)	74	1[‡]	1M	HARDI UKF	3T DWI, 1.25mm, 108g single-b=3K
HCP-IW[§]	HCP	Inclusive Wasserthal et al. (2018)	72	23	10M	CSD iFOD2	3T DWI, 1.25mm, 270g multi-b=(1,2,3)K
APSS-IS	APSS	Inclusive expert: S.S.	4	5	100K	DTI EuDX	1.5T DWI, 2.5mm, 60g single-b=1K

HCP diffusion The Human Connectome Project (HCP) (Van Essen et al., 2013) diffusion dataset is the base of most of the datasets we present in the next paragraphs. It is a public and widely adopted dataset where the acquisition and pre-processing pipeline for diffusion MRI is carefully validated (Milchenko and Marcus, 2013; Sotiropoulos et al., 2013; Glasser et al., 2013). The dataset comprises healthy subjects (all genders) aged between 24 and 35. Each subject has both the structural T1 image and the 3T Diffusion Weighted Image (DWI). The DWI has resolution 1.25mm with 270 gradients multi-shell (Andersson and Sotiropoulos, 2016), and is corrected with eddy currents (Andersson and Sotiropoulos, 2015).

[*] `https://doi.org/10.25663/brainlife.pub.13`

[†] `https://github.com/SlicerDMRI/ORG-Atlases`

[‡] averaged from 100 subjects.

[§] `https://zenodo.org/record/1477956#.Ya67UyzMKL8`

HCP-EP This dataset is composed of 20 randomly selected HCP subjects, for which a tractogram has been reconstructed. The tracking pipeline comprised the estimation of the diffusivity model using the Constrained Spherical Deconvolution (CSD) (Tournier et al., 2007), and the (ii) Particle Filtering Anatomically Constrained Tractography (PF-ACT) algorithm (Girard et al., 2014). More in detail, the tracking generated around ∼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, we have available an *exclusive* labeling generated by the rule-based method Extractor (Petit et al., 2019). The rules are manually defined based on the white matter anatomy and provide a label for each streamline of a tractogram. The label given is either anatomically plausible (p) or anatomically non-plausible (np). The anatomical criteria encoded by the rules regards the non-plausibility i.e., the labeling is conservative with respect to unknown pathways. Non-plausible streamlines are identified with a 2-step procedure: (i) the first step individuates 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) the second step makes use of clustering (Côté et al., 2015) and based on on the three main categories of pathways, i.e., associative, projective, commissural, marks as non-plausible local outlier streamlines. Eventually, the labeling results in a balanced partitioning: the average on the 20 tractograms is 49.7 ± 1.5% of p streamlines. Among the non-plausible, the most prominent are short streamlines, representing the 31.8±1.2% of the whole set of non-plausible.

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 ∼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 through time. Each split considers the streamlines belonging to certain classes of bundles as plausible and all the others as non-plausible.

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.

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[1] 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^2 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 streamlines.

An expert neurosurgeon 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

[1] These patients are not the same of the APSS dataset of Chapter 5.3. The different selection is due to the lack of structural T1w without contrast agent for certain patients. VF requires the structural volume as it needs the tractogram to be mapped to MNI space

(SLF), and the Inferior Fronto-Occipital Fascicle (IFOF), and one bundle in the lesioned hemisphere: Pyramidal Tract (CST).

6.5 Empirical analysis

The design of the empirical analysis is organized into two parts: model-related and task-related experiments. The former is devoted to assessing the properties of the proposed model; the latter aims to investigate the effectiveness as a solution for the task of tractogram filtering. The performances 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:

- looking at the distribution of the misclassification error,
- considering the behavior on different types of tractograms,
- simulating the evolving definition of anatomical plausibility to test the adaptation in the case of concept drift.

6.5.1 Model related experiments

5-fold CV filtering on HCP-EP The first experiment is designed to measure the learning performances of Verifyber according to the usual setting of cross-validation. For this purpose, we consider a sample of 20 individuals from the HCP dataset where the fibers have been labeled following the heuristic rules of Extractor. 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 alfa 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, i.e., a fiber is either anatomically plausible or non-plausible, is carried out by measuring the accuracy, the precision, the recall, and the Dice-Sørensen coefficient (DSC). Results are reported in Table 6.3.

Comparison with alternative DL architectures The second experiment aims to compare Verifyber with the competing deep learning models, namely bLSTM, PN, and DGCNN. In this experiment, we operate the 5-fold cross-validation setting adopted for Verifyber to all other methods and measure the same evaluation metrics. For a fair comparison,

Table 6.2: Models architecture used for experiments.

Method	Architecture	params
bLSTM	MLP(128)→LSTM(256)⊕LSTM^{-1}(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

we enforce a uniform configuration among the different models: the choice of the input representation, the number of parameters of the NN models, and the hyper-parameters for training. We report the training curves in Figure 6.3.

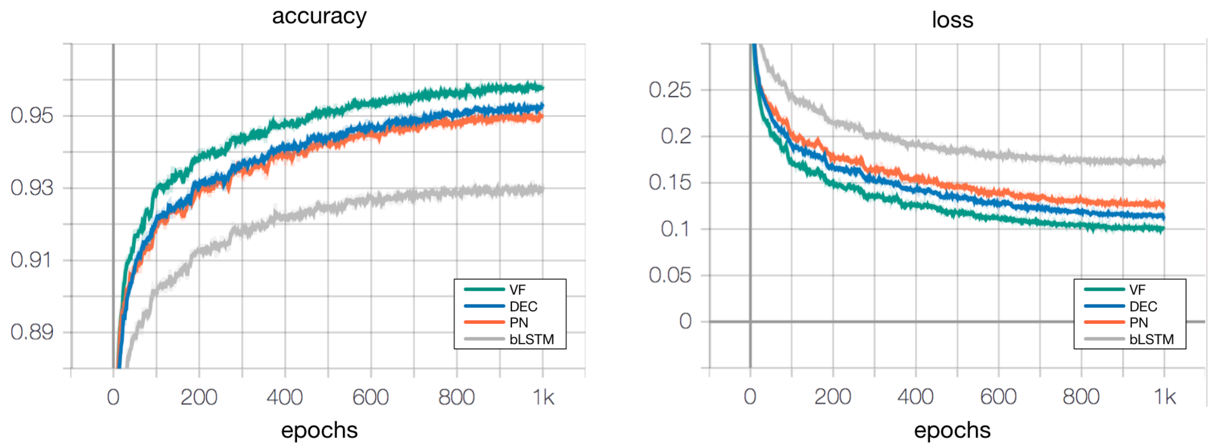


Figure 6.3: Training curves of the four methods compared. Despite the similar number of parameters of all the models we can observe that bLSTM (grey) has clear negative gap with respect to all the other models. Conversely, VF (green) has always a positive gap compared to all the others.

Regarding the input representation, GDL models can deal with size-varying input, e.g., streamlines with different 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 a resampling to 12, 16, or 20 points per streamline. Since a side empirical assessment did not provide any significant difference in performance, we operate a resampling to 16 points for all the subsequent experiments. To make the size of NN models homogeneous, we set the architecture of the different methods with a uniform number of parameters, as reported in Table 6.2.

In Table 6.3 we show the performance achieved by all methods on the HCP-EP dataset. The scores obtained by Verifyber are the highest and most stable across different metrics, even though the gap compared to PN and DGCNN is small. The lowest performance is obtained by bLSTM, for which the computation of t-test provides a statistical significance, with a p-value $< 10^{-3}$. The inspection of behavior at training time, see the Figure 6.3, confirms a poor minimization of the loss for bLSTM compared to other GDL models.

Table 6.3: 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 (±0.1)	93.8 (±0.1)	96.2 (±0.2)	95.0 (±0.1)
PN	94.7 (±0.1)	95.5 (±0.2)	96.9 (±0.2)	96.2 (±0.1)
DGCNN	94.4 (±0.1)	95.4 (±0.2)	96.5 (±0.1)	96.0 (±0.1)
VF	**95.2** (±0.1)	**96.1** (±0.2)	**96.9** (±0.1)	**96.6** (±0.1)

Permutation invariance test The third empirical investigation, concerned with the property of the proposed model, aims to assess the sensitivity to order of points in a fiber. 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 anatomically non-plausible likely. 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 6.4. 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.

Table 6.4: 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 6.3 of the paper.

Method	Accuracy	Precision	Recall	DSC
bLSTM$_{\text{perm}}$	64.1 (±1.1)	89.8 (±1.0)	55.1 (±1.1)	68.3 (±0.9)
PN$_{\text{perm}}$	94.5 (±0.1)	95.4 (±0.2)	96.8 (±0.2)	96.1 (±0.2)
DGCNN$_{\text{perm}}$	94.3 (±0.1)	95.4 (±0.3)	96.5 (±0.2)	95.9 (±0.2)
VF$_{\text{perm}}$	**30.0** (±2.8)	**87.7** (±0.7)	**00.3** (±0.0)	**00.6** (±0.1)

6.5.2 Task related experiments

Performance analysis for types of streamline A second purpose of the empirical investigation is to understand how the different models behave with respect to the geometrical and anatomical properties of fibers. The geometrical properties of fibers might be captured considering two features: the length and the curvature. The combination of these features may represent a good proxy of the anatomical properties. 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 (computed per-point as in the Frenet-Serret Space Curve Invariants): 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, we obtain 9 groups of fibers, see Figure 6.4.

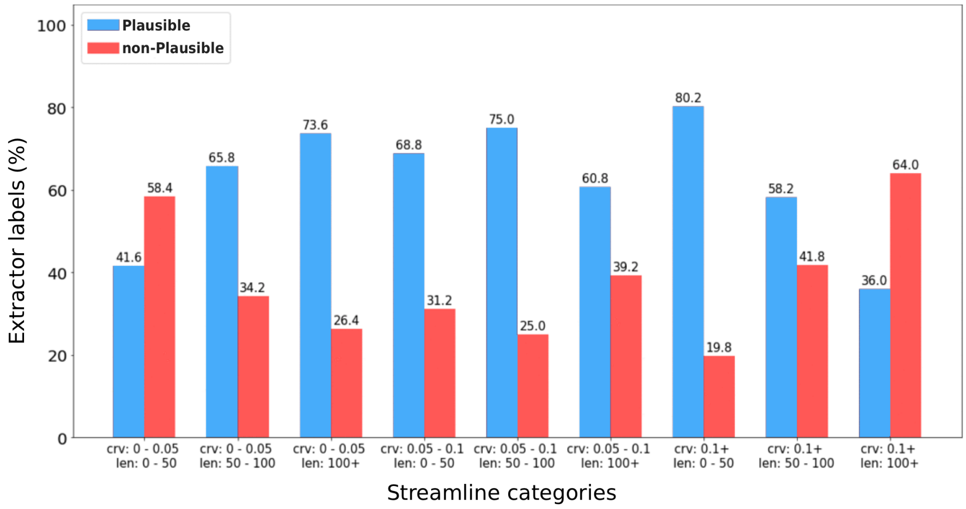
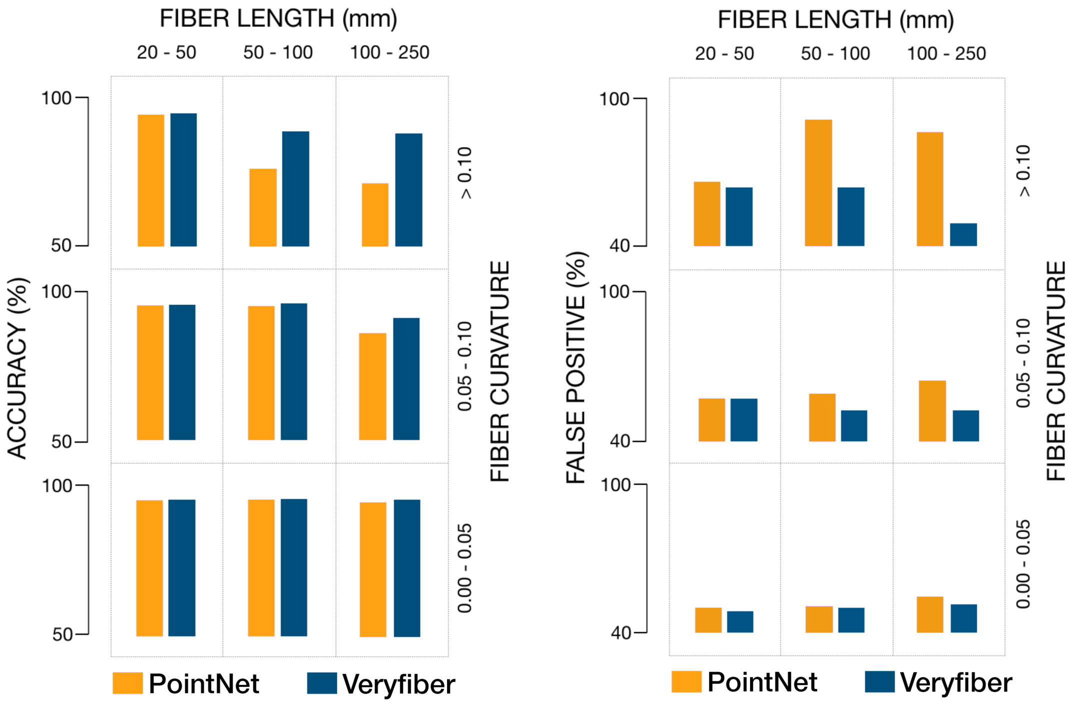


Figure 6.4: Distribution of Extractor labels per streamline category.

Comparison with PointNet (Qi et al., 2017) We may proceed by looking at these groups and inspecting where the predictions fail to discriminate between anatomically plausible and non-plausible fibers properly. A deeper analysis of the misclassification error might provide a better comprehension of how Verifyber, based on edge representation, differs from PointNet, based on bare point clouds. Despite a similar score of classification accuracy, the two methods share only a 60% of the error while the remaining 40% concerns different fibers. The distribution of error with respect to these 9 groups is reported in Figure 6.5a.

Figure 6.5a highlights that on longer and more curved fibers, Verifyber outperforms PointNet. An interpretation of the source of such a difference might be achieved by looking at the internal representation of the two models. The relevance weights associated with each fiber point are uniformly distributed in PointNet, while in Verifyber, the learning



(a)

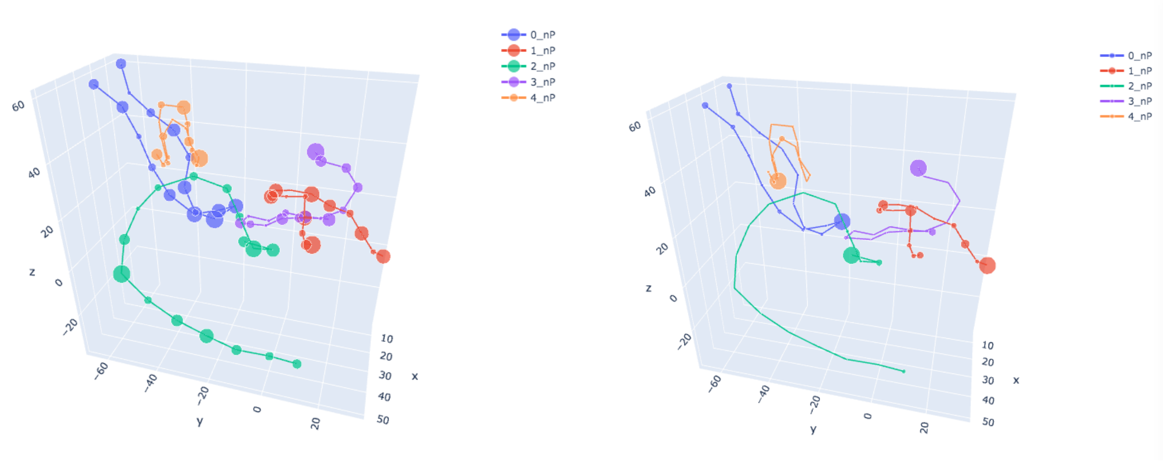
(b)

Figure 6.5: 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).

process clearly identifies a few more discriminating points. We show visual evidence of this difference for long and more curved fibers in Figure 6.6.

In-depth error characterization We may 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., 2019) 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 6.5b 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 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.

Figure 6.6: Given a set of streamlines, we compared 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 analysed 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 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.



(a) PN

(b) Verifyber

Considering the conservative approach of exclusive labeling, 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 might 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 6.7. 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 with HCP-IZ 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,

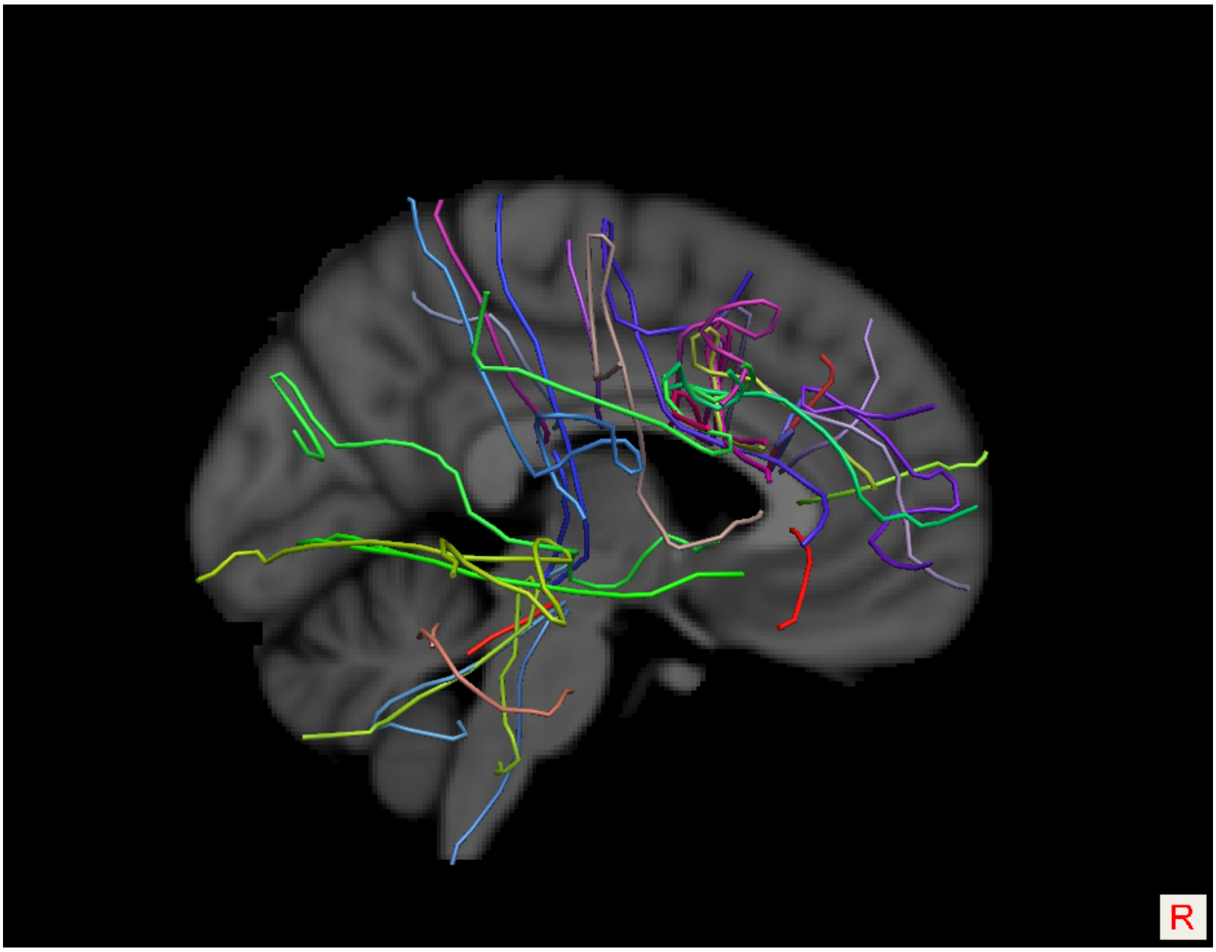


Figure 6.7: Example of FN i.e., labeled as plausible but classified as non-plausible, produced by VF. The streamlines shown are clearly non-plausible showing the presence of noise in the rule-based labeling, and the capacity of VF of extending the concept learnt by the rules.

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 6.5 shows the results of the incremental learning for the five stages. The scores confirm that Verifyber is quite stable and does not suffer the drift of the 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 might be explained by the balance shift between the number of plausible and non-plausible fibers.

Table 6.5: 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 considers all the streamlines of the Zhang et al. atlas (Zhang et al., 2018) as plausible and the rest as non-plausible.

Method	Plausible	Accuracy	Precision	Recall	DSC
Verifyber	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
	All bundles	97.1	97.6	98.0	97.8

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 6.8. 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 FINTA (Legarreta et al., 2021) An interesting open question is whether our supervised approach is better when compared to a state of the art unsupervised tractogram filtering approach such as FINTA (Legarreta et al., 2021). FINTA is a recent approach based on a convolutional autoencoder, which, as in our case, is trained directly on the raw streamline structure. We think that such a comparison might be relevant to clarify the difference between unsupervised and supervised approaches. Unfortunately, neither the code nor the data utilized 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 6.5.3.

Analogously to what is done in (Legarreta et al., 2021) we train the autoencoder of FINTA with a single average tractogram, the one of HCP-IZ, whose streamlines are

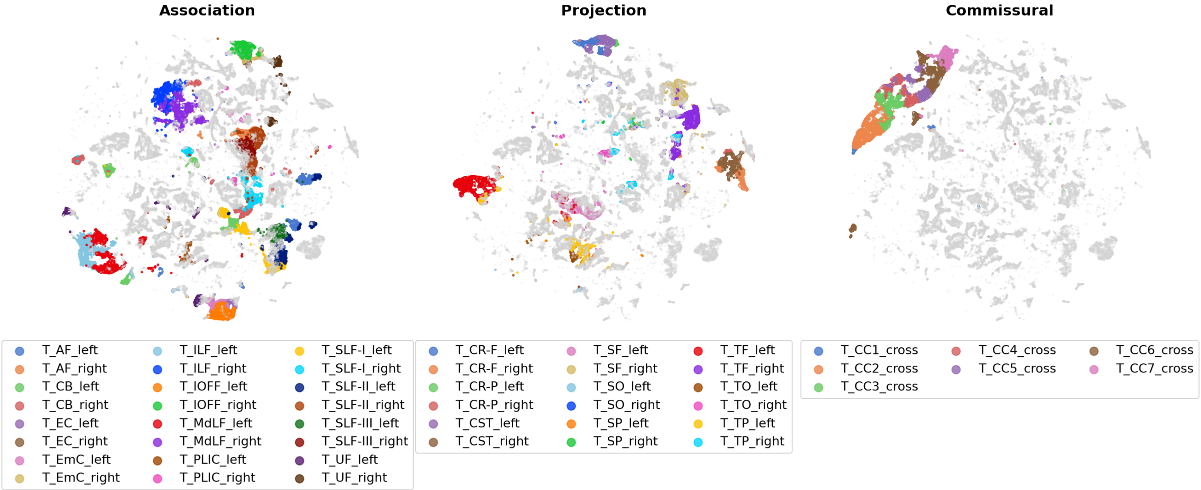


Figure 6.8: Latent space learned by VF$^{\mathrm{IZ}}$. Each point corresponds to a streamline in the 1024D space reduced to 2D by means of t-SNE (Van der Maaten and Hinton, 2008). The three plots show the non-plausible streamlines in light-gray, together with one macro category of bundle i.e., association, projection, commissural at a time.

randomly split into 80/20 for train and test. The training is performed using the hyper-parameters reported by the authors where possible; default values are used otherwise. We trained the autoencoder until convergence (see Appendix B.3.1). In addition, we performed a qualitative assessment of streamlines that were reconstructed with the trained autoencoder to check that the reproduced approach worked as expected (see Appendix B.3.1). After the training, we have run nearest neighbor classification of p/np streamlines considering both the dataset HCP-IZ and HCP-EP. For each dataset, we computed the latent features of a portion of the training streamlines labeled as p using the trained autoencoder. Subsequently, we tuned a plausibility threshold, i.e., Euclidean distance in the latent space, using a portion of streamlines randomly sampled from the validation set. Finally, the threshold has been used to classify streamlines of the test set.

In Table 6.6 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 obtain scores similar to the result published in (Legarreta et al., 2021) and reported in the first (light grey) row of the table. Though, the performance of Verifyber is six points higher in terms of DSC and nine points 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.

Model deployment on HCP-IW In the previous experiments, we trained Verifyber on HCP-EP and HCP-IZ datasets, where 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 where to do inference on a new dataset, namely HCP-IW,

Table 6.6: FINTA performance evaluation. We report the score of accuracy, precision, recall, and DSC for each run. The training of the FINTA autoencoder is performed using data from the dataset indicated in the apex, e.g., FINTA$^{\text{IZ}}$ is trained on the dataset HCP-IZ. Specifically we use the same training/test set split used for our Verifyber experiments e.g., HCP-IZ is split into 80/20 training/test. Then, we select a portion of p fibers is used as reference (Ref. p) for the embedded nearest neighbor step of FINTA. The selection is performed from the training set of the dataset reported in the second column. The radius threshold (reported in column two) is tuned on the embedded space using as query fibers a portion of the validation set of the same dataset.

Method	Thr	Reference p	Acc	Prec	Rec	DSC
FINTA	13.6	Legarreta et al. (2021)	91.0	91.0	91.0	91.0
		HCP-IZ all				
FINTA$^{\text{IZ}}$	36.9	Zhang et al. (2018)	88.0	87.3	95.8	91.3
Verifyber	-		**97.1**	**97.6**	**98.0**	**97.8**
		HCP-EP				
FINTA$^{\text{IZ}}$	46.0	Petit et al. (2019)	74.3	75.1	94.6	83.8
Verifyber	-		**95.2**	**96.2**	**96.9**	**96.6**

using the models trained on HCP-EP and HCP-IZ. Since a ground truth is not available on HCP-IW, we need to revise the evaluation procedure. The segmented bundles in HCP-IW might be considered fiducial regions, where to focus on a 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 6.10. We operate 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 of HCP-IW datasets. In addition, we investigate the agreement of the two models trained on two different ground truths, i.e., an inclusive and an exclusive policy.

In Figure 6.9 and Figure 6.10 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 Figure 6.11 and Figure 6.12. 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.

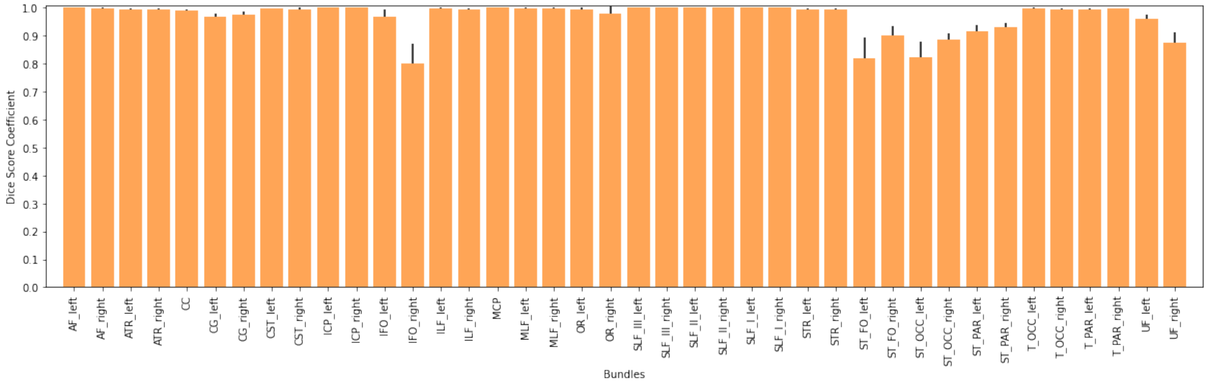


Figure 6.9: Volumetric DSC of Tractseg bundles after filtering using $\mathrm{VF}^{\mathrm{EP}}$. Reported mean and standard deviation on 23 subjects. The filtering does not impact significantly the shape of the bundle guaranteeing at least 80% of DSC. The mean DSC is 96.8 ± 5.4.

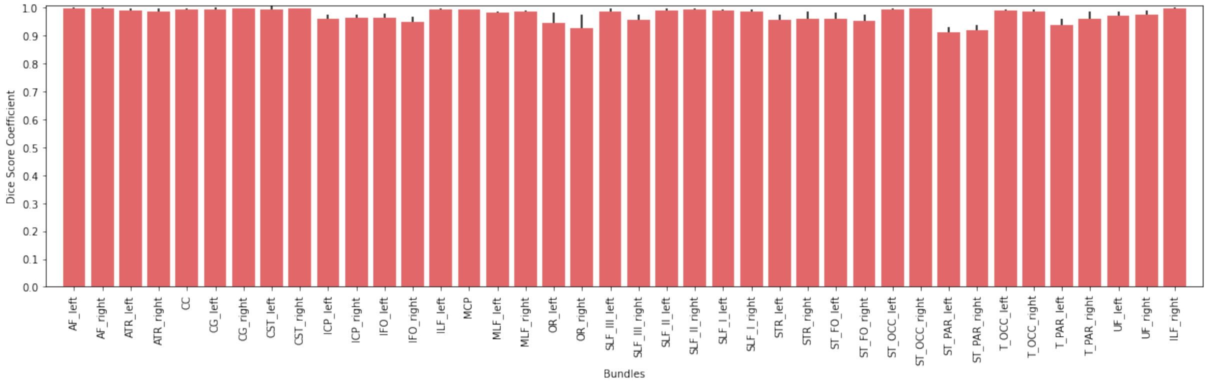


Figure 6.10: Volumetric DSC of Tractseg bundles after filtering using $\mathrm{VF}^{\mathrm{IZ}}$. Reported mean and std on 23 subjects. The filtering does not impact significantly the shape of the bundle guaranteeing at least 85% of DSC. In most of the cases the DSC is above 95%. The mean DSC is 97.5 ± 2.4.

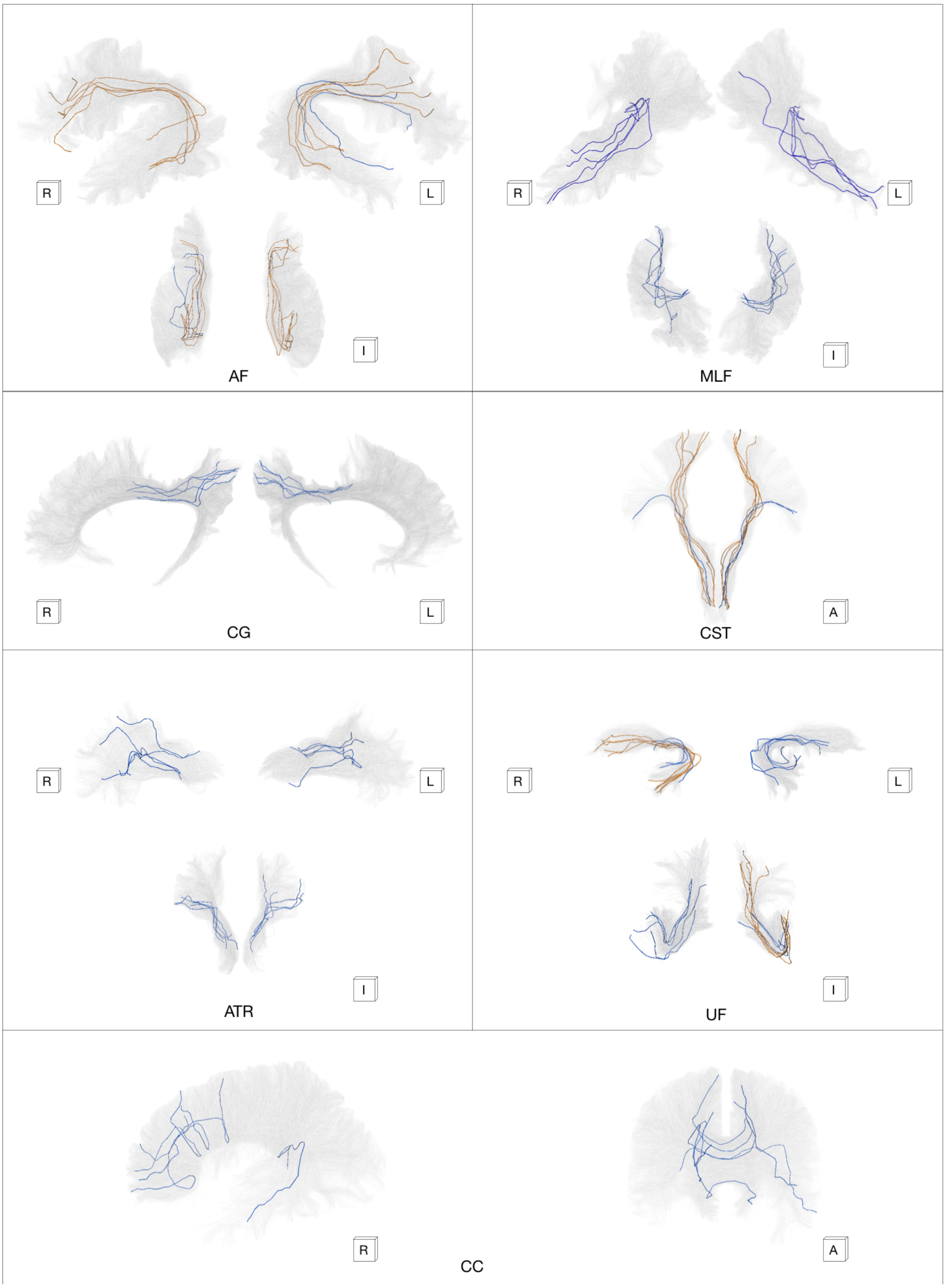


Figure 6.11: Qualitative example of non-plausible streamlines belonging to Tractseg bundles individuated using VF^{EP}. The figure shows in blue streamlines individuated as non-plausible both by VF^{EP} and VF^{IZ} (shared non-plausible). The orange streamlines are instead exclusive of VF^{EP}. In some 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^{EP} and when we use VF^{IZ}.

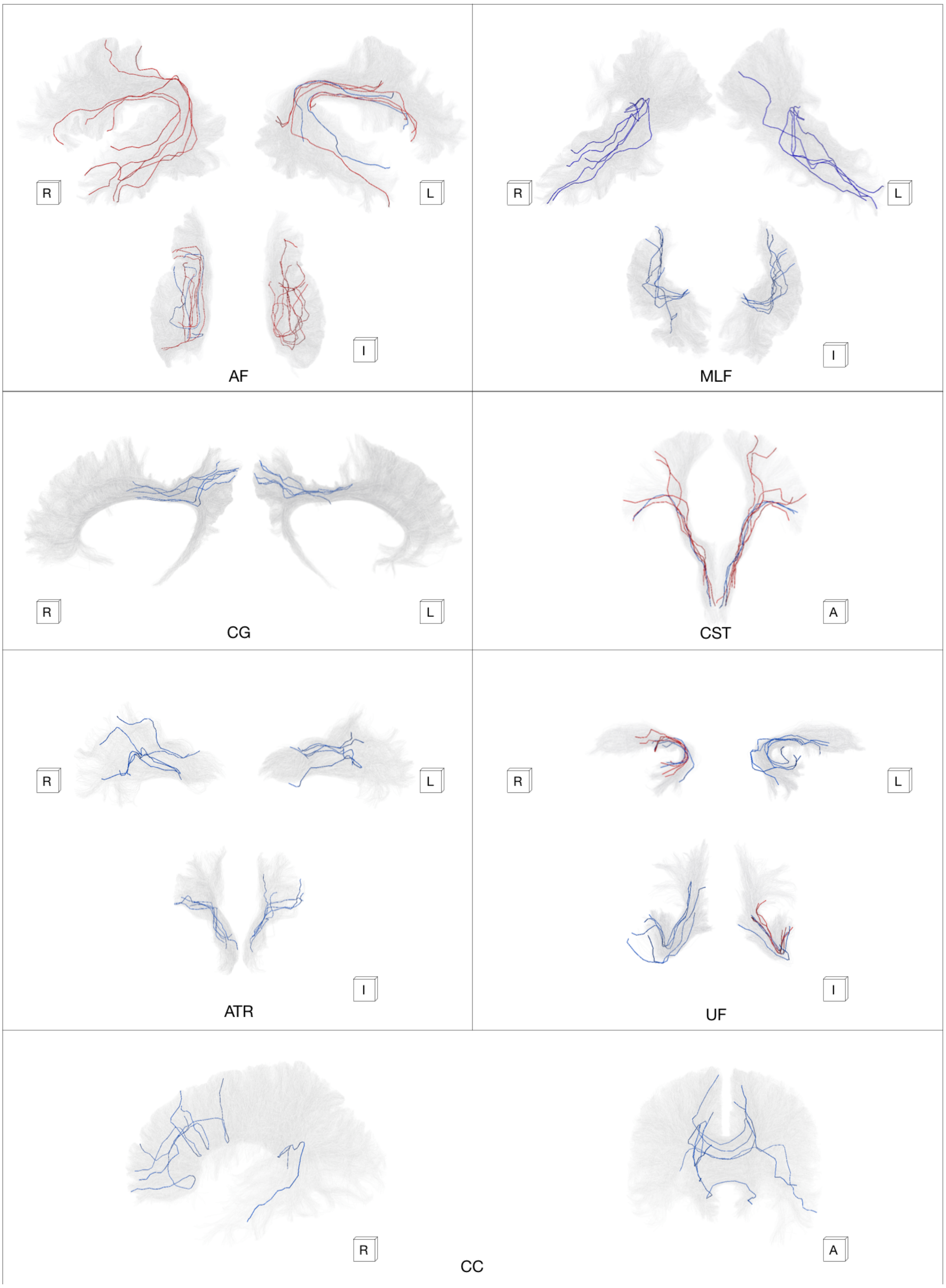


Figure 6.12: Qualitative example of non-plausible streamlines belonging to Tractseg bundles individuated using VF^{IZ}. Blue streamlines are shared non-plausible examples, while red streamlines are exclusive non-plausible of VF^{IZ}.

Model deployment on APSS-IS As an additional real-world experiment, we present a possible clinical application of our approach. This experiment considers the patients with tumors of the APSS-IS dataset, and we filter their tractograms with VF^{EP}. Note that the choice of filtering with VF^{EP} rather than VF^{IZ} is driven by the more conservative approach of the underlying labeling. Similar to the experiments involving HCP TractSeg, we also use, in this case, the segmented bundles as a proxy evaluator. However, since the clinical bundles are much poorer, i.e., less dense, than the TractSeg ones (due to data quality and time restrictions), the filtered bundles could change their shape significantly even with the removal of very few non-plausible streamlines. Therefore, as we cannot do any supposition on the expected shape, we opt to evaluate the filtering impact only qualitatively, inspecting three bundles from the healthy hemisphere and one bundle from the tumored hemisphere of each subject.

To evaluate the qualitative analysis of VF^{EP} deployed on the APSS-IS dataset, we report two figures. Figure 6.13 depicts the filtering of bundles segmented from the healthy hemisphere of the subjects, while Figure 6.14 shows one bundle segmented from the tumored hemisphere. In both cases, the individuated non-plausible streamlines (colored in black) strongly disagree with the 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.

6.5.3 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/tractogram_filtering`. The Github repository also contains our FINTA (Legarreta et al., 2021) implementation. In addition, to simplify the deployment of VF^{EP} and VF^{IZ}, we published an App on the BrainLife platform (Avesani et al., 2019), accessible at `https://doi.org/10.25663/brainlife.app.390`.

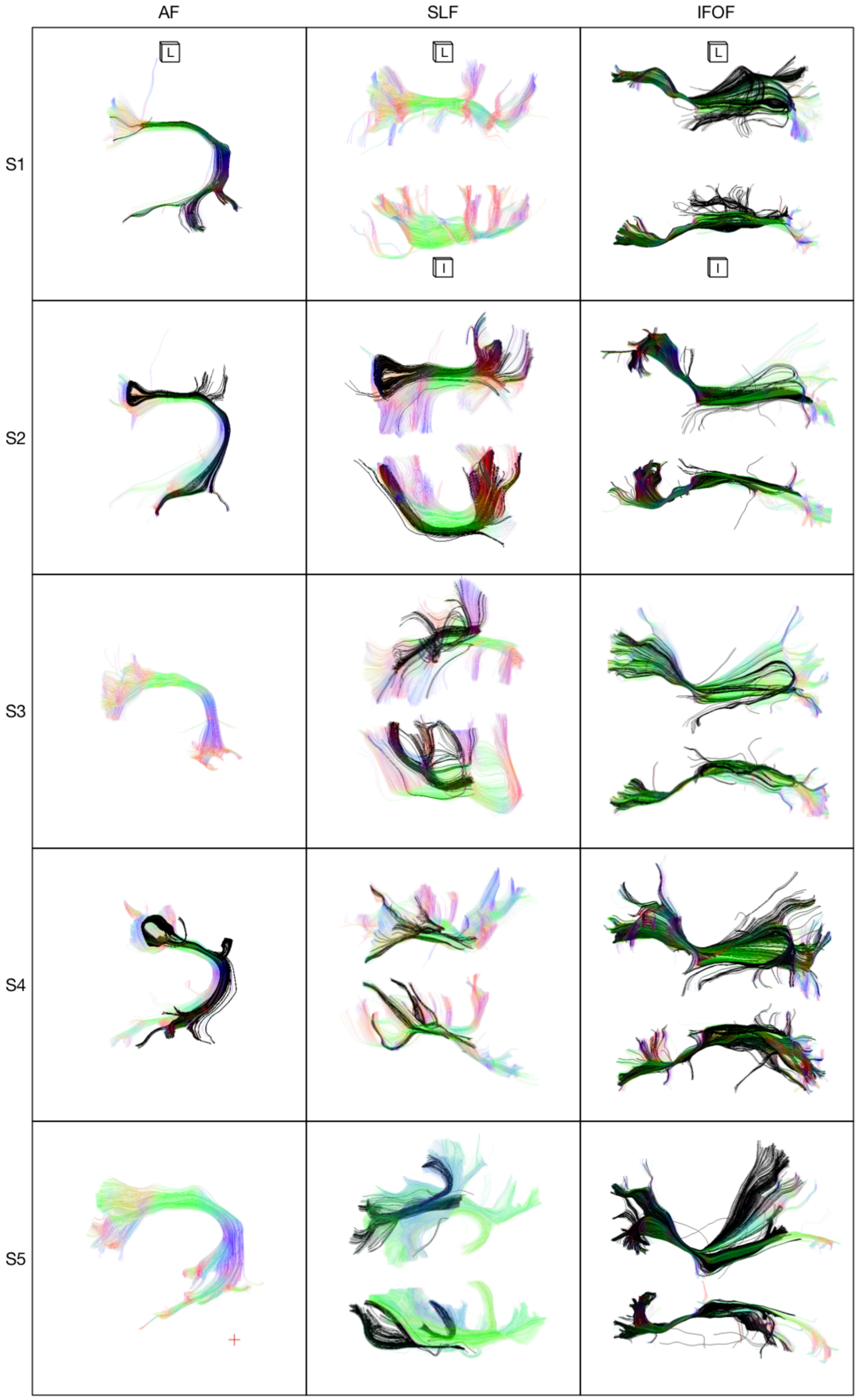


Figure 6.13: Qualitative results after filtering APSS clinical dataset using VF$^{\mathrm{EP}}$. Black streamlines are classified as non-plausible. Bundles reported in this figure belong to the healthy hemisphere of patients.

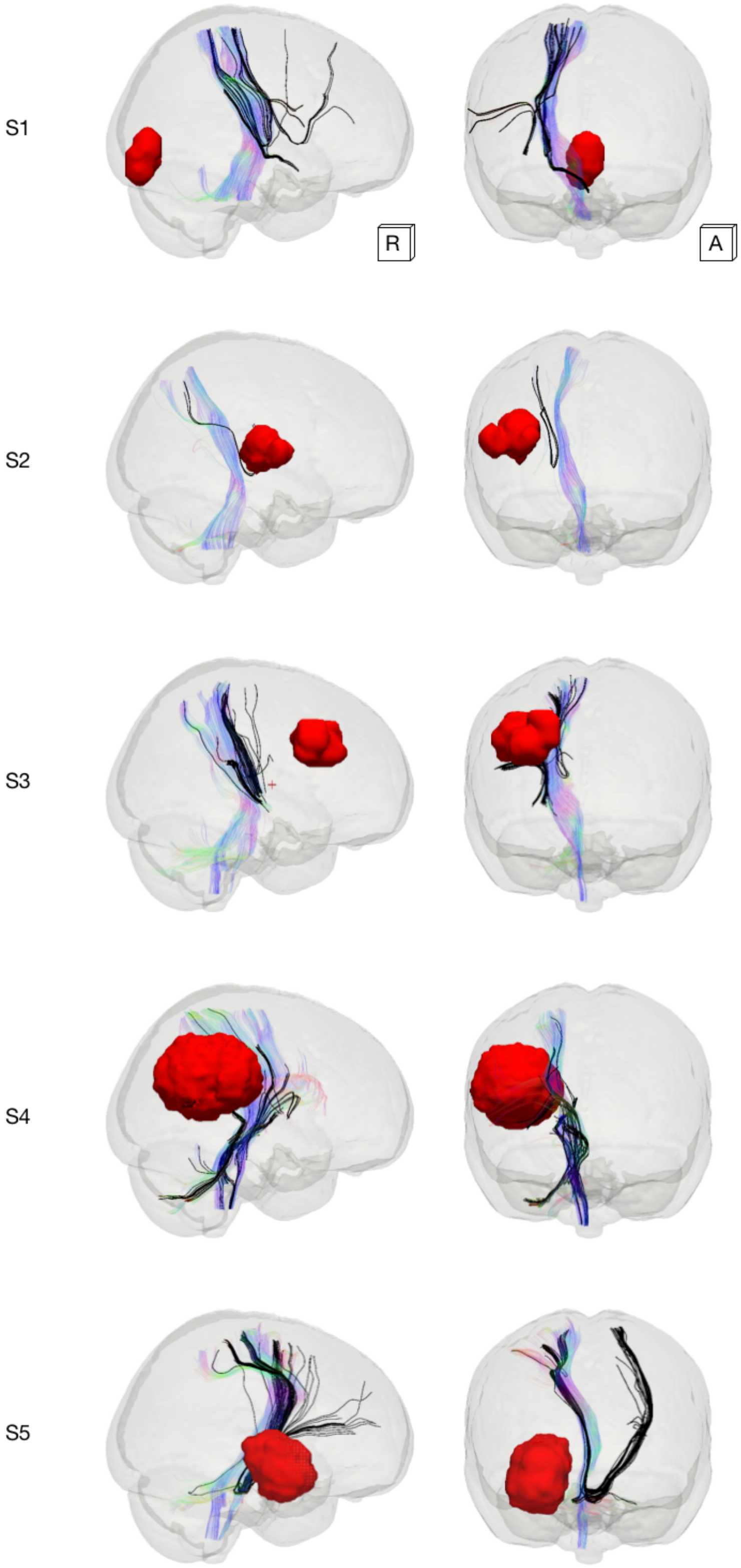


Figure 6.14: Qualitative results of the tumored hemisphere after filtering APSS clinical dataset using $\mathrm{VF}^{\mathrm{EP}}$. The bundle shown is the pyramidal tract (PT). Black streamlines are classified as non-plausible. The red ROI is the segmented tumor.

6.6 Discussion

6.6.1 Verifyber performance

Verifyber effective for tractogram filtering The proposed method, Verifyber, resulted successful for the tractogram filtering task. It can learn from an external supervision, i.e., Extractor labeling (see Figure 6.3), and then maintain the high performance reached during training also at test time, see Table 6.3. In all the metrics reported, very high scores are coupled with a very low standard deviation in the cross-validation setting, e.g., accuracy 95.2±0.01% and DSC 96.6%±0.01, proving robustness to the choice of the train/test subjects. The higher recall compared to precision suggests more conservative filtering, in agreement with the labeling's *exclusive* policy. Additionally, VF revealed fast inference at test time, classifying 1M streamlines in less than a minute, i.e., 46.2 sec using a GPU NVIDIA Titan XP 12GB. Although we did not present a speed comparison with common filtering methods, e.g., signal-based or rule-based, we remind that such methods usually require tens of minutes.

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 orientation, at the cost to be invariant to the permutations of streamline points. In VF, we overcome such a drawback, as confirmed by the permutation test results in Table 6.4 while remaining invariant to the fiber orientation. Furthermore, we show that considering the edges of the input graph (streamline) matters, obtaining a percentage of non-shared error between PN and VF around 40%. When streamlines are longer and curved, edges become more informative and provide a competitive advantage. In these categories, VF has only a small drop in accuracy while significantly outperforming PN, 89% vs. 76% and 88% vs. 71% in the very curved mid and long fibers, respectively (see Figure 6.5). This gap is explainable by the different representation learning abilities of the two models. (Sequence) Edge convolutional layers of VF retain more local information into the learned embedding of fibers, capturing the most salients in complex pathways. On the contrary, PN struggles to identify salient points as the learned global embedding is not informative for local patterns. See Figure 6.6.

Error analysis In *exclusive* labeling policies, the false positive error is more relevant than the false negative. Observing the FP analysis in Figure 6.5, VF disclosed robustness despite the uneven distribution of the plausible and non-plausible labels, as observable in Figure 6.4. Compared to PN, which seems to be influenced by the higher number of plausible labels, 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 6.7) proves to some extent the ability of the model to generalize the rules beyond the labeling. Despite some noisy labels of these streamlines, the large number of streamlines seen during training (around 10M) provide generalization ability to VF. We may remark the importance of this property as manual labeling is intrinsically error-prone and always leads to having some noisy labels. Also, the lack of complete knowledge of the white matter anatomy does not allow the definition of a gold standard labeling.

Incremental learning The lack of a complete WM knowledge also leads to a continuous evolution of the notion of anatomical plausibility, which reflects on tractogram labelings, i.e., *concept drift*. New bundles are added to the definition of plausibility as long the neuroanatomist consolidates their definition. We emulated such a real-world scenario in the incremental learning experiments, where VF obtained convincing results, Table 6.5. 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 6.8). This result suggests that VF effectively captures the trajectory of fibers. Therefore, we may expect robustness to (inclusive) incremental learning also in the real world. 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 DL approach in contrast with unsupervised to guarantee higher flexibility to different definitions of anatomical plausibility. The comparison with FINTA, chosen as representative of DL unsupervised methods, in Table 6.6 highlights a significant difference between the two types of approach. The performance of Verifyber is high and stable on both the *inclusive* and *exclusive* labeling policies, i.e., only 2% of accuracy and 1pt of DSC difference. On the contrary, FINTA is not robust to two labeling policies, obtaining 14% of accuracy and 7% of DSC difference. We remark that such a difference occurs although we tuned the filtering threshold of FINTA specifically for each dataset/labeling. In our experience, adopting the same threshold across different datasets highly worsen the results (see Appendix B.3.1).

An additional observation concerning FINTA is that it performs considerably better on the inclusive labeling of the HCP-IZ dataset. This labeling policy discriminates between plausible and non-plausible using the concept of bundle or cluster. Hence, non-plausible fibers are isolated anatomically unknown bundles or outliers (false positives) manually removed from clusters. Both the situations are favorable for autoencoder-based solutions such as FINTA, as AEs are known to perform well in the task of anomaly/outlier detection (Pang et al., 2021). The low dimensional bottleneck of the AE 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. Nevertheless, the filtering done with FINTA is not perfect and tends to produce false positives, i.e., it misclassifies non-plausible fibers as plausible(see Appendix B.3.1).

When we move from inclusive to exclusive labelings, the hypothesis of non-plausible as outliers/isolated 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 B.3.2), 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% 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 of 20% accuracy and 13% DSC on the exclusive labeling.

6.6.2 Model deployment

Impact on TractSeg bundles In the deployment of VF to unseen data, we proposed the use of bundles from the HCP-IW as proxy evaluators of the filtering. We tested the use of both $\mathrm{VF}^{\mathrm{EP}}$ and $\mathrm{VF}^{\mathrm{IZ}}$. From the results reported in Figure 6.9 and 6.10, we notice in both cases there is no meaningful change of volumetric bundle mask 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, not having a ground truth labeling to refer to, such high percentages of non-plausible may be an error of our model. For this reason, we needed to operate an additional visual check to validate the fibers' non-plausibility. In the specific case of the IFOF, the visual check reported Figure 6.15 revealed a premature termination in the frontal lobe. This situation is probably due to the lack of anatomical constraints in the tracking procedure that TractSeg adopted for the IFOF ((Wasserthal et al., 2018)). For all the other bundles, an expert visually investigated a portion of the predicted non-plausible fibers confirming that such fibers conduct non-plausible pathways, see Figure 6.11 and 6.12. 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/safe VF filtering. Using a common test dataset, i.e., HCP-IW, for $\mathrm{VF}^{\mathrm{EP}}$ and $\mathrm{VF}^{\mathrm{IZ}}$ allows us to compute the agreement of the predictions as a rough estimate of the labeling policies overlaps. 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. and Extractor, i.e., inclusive vs. exclusive.

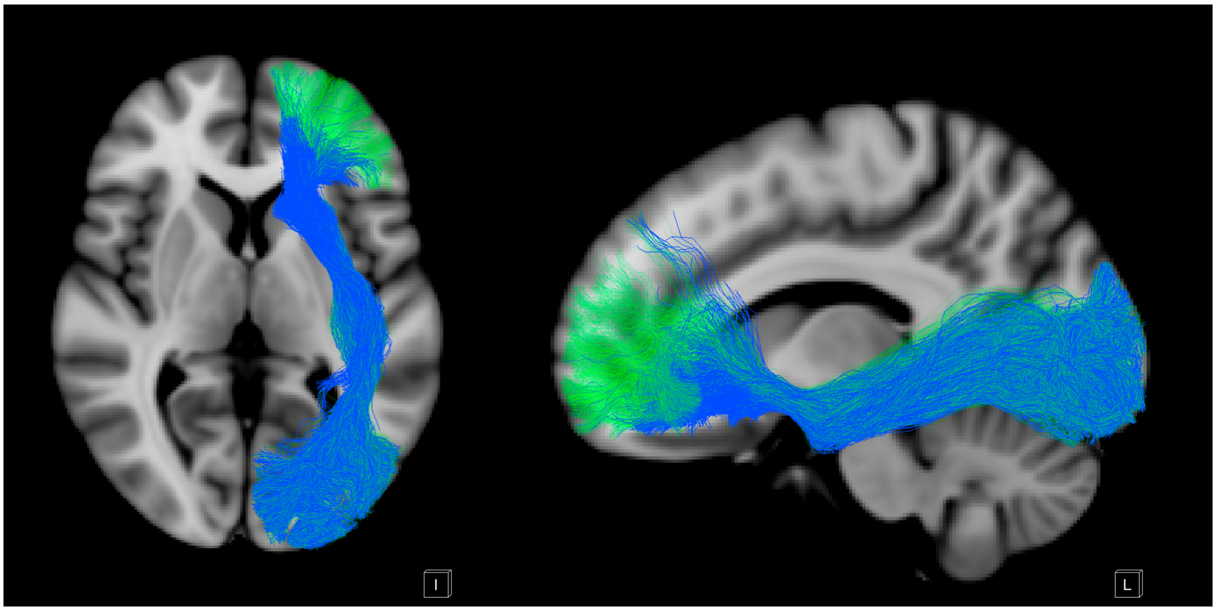


Figure 6.15: Filtering of Tractseg IFOF using VF^{IZ}. Blue streamlines are classified as non-plausible, while transparent green are classified as plausible.

Clinical application The last validation of VF on clinical data, i.e., APSS dataset, shows that the method is suitable for clinical use. 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 6.13 and 6.14. Indeed, the characteristics of the detected non-plausible agree with the principles of some of the Extractor rules, e.g., truncation, loop/strict U-turn. In most cases, it is easy to notice the disagreement of filtered fibers compared to the retained fibers 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 otherwise necessary.

6.6.3 Possible weaknesses and limitation

Choosing between VF^{EP} and VF^{IZ} Some precautions must be taken into account when deploying Verifyber to new data, particularly considering the choice of VF^{EP} versus VF^{IZ}. In the case of VF^{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 VF^{IZ}, as the HCP-IZ training dataset contains fibers that are only linearly registered to a shared anatomical reference. However, $\mathrm{VF}^{\mathrm{IZ}}$ 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 (see Appendix B.2.3). For the sake of completeness, in additional experiments, we explored using VF to learn the SIFT2 filtering. Such experiments are reported in Appendix B.2.3.

Quantitative generalization experiments are missing Despite the extensive set of experiments employed to test Verifyber, we miss a second labeled dataset where quantitatively confirms 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 B.2.2.

Chapter 7

Conclusion

In this Chapter, we draw the conclusions of the thesis. First, we provide an overall discussion of the thesis impact, highlighting the novel contributions and limitations (Section 7.1 and 7.2). Then, we discuss future research directions (Section 7.3).

White matter characterization with deep learning The ultimate goal of this thesis was to reduce artifactual fibers in tractograms and enable a more accurate automatic process of white matter characterization. The presented approaches, i.e., StemSeg and Verifyber, enrich the neuroimaging literature with novel ways of characterizing the white matter using deep learning. Our work constitutes a contribution from both an applicative and a methodological perspective. First, we propose effective strategies for more accurate characterization of the white matter, reducing the number of artifacts (see Section 7.1). Then, we demonstrate the power of deep learning for accurately processing tractography data (see Section 7.2).

7.1 White matter characterization with supervised learning

The accurate characterization of the brain white matter is a task/topic of pivotal importance for several neuroanatomical, clinical, and neuroscientific applications. In this thesis, we proposed to improve the WM characterization by injecting anatomical priors into tractograms. The application of anatomical priors as constraints for tractography and tractograms was already proved to be one of the best ways to obtain a correct characterization of the white matter (Schilling et al., 2020; Girard et al., 2020). However, one of the necessary conditions is that anatomical priors must be accurate and specific for the anatomy at hand. For this reason, such works resorted to experts' manual intervention. On the contrary, in this thesis, we explored the encoding of anatomical priors using an automatic strategy. In particular, to guarantee anatomical grounding, we proposed casting the white matter characterization as a supervised learning problem. Such an approach constitutes a novelty for the task of learning anatomical priors not related only to bundles characterization.

Anatomical priors can encode different types of anatomical knowledge. Typically it encodes either anatomical regions of the white matter or directly the white matter fibers. In this thesis, we challenged the learning of both these types of anatomical priors via supervised learning:

- In Chapter 5 we presented *StemSeg*, a supervised learning approach to learn anatomical waypoints regions, namely *stems*, crucial for the definition/segmentation of bundles. We proposed a two-step procedure that (i) automatically segments the stem-ROI of a bundle using a Convolutional Neural Network (CNN), and (ii) exploits the predicted stem to improve the accuracy of the segmentation rule for that bundle. To the best of our knowledge, StemSeg is the first automatic *stem-based* method for bundle segmentation. The experimental results show the effectiveness of the proposed method for IFOF segmentation. StemSeg, substantially outperformed the competitor approach, TractSeg (Wasserthal et al., 2018), which is the state-of-the-art for bundle segmentation. Finally, as a side contribution of StemSeg, we published the training dataset, composed of manually curated stem-ROIs for a large set of individuals (more than 150 subjects).

- In Chapter 6, we proposed Verifyber, the first method for tractogram filtering based on the supervised learning of fibers' anatomical plausibility. Verifyber achieves high accuracy on in-distribution and out-of-distribution tractograms from different tractography pipelines, including clinical cases, showing a good generalization capability. Moreover, it can be effectively trained with differently labeled data — as long as they are based on fiber anatomy. In our analyses, we validated this property considering two different anatomical labelings and a simulated evolutionary scenario with five stages of labeling refinement. Further, the supervised strategy of Verifyber is more beneficial compared to the unsupervised competing methods, such as FINTA (Legarreta et al., 2021), which show a lower flexibility to different labelings. We remark that Verifyber's flexibility to labelings is extremely valuable as the white matter structural connectivity is still under investigation, and the community continuously proposes novel accurate tractogram labelings. As soon as new labelings are available, Verifyber can be re-trained to learn meaningful features and enable more accurate tractogram filtering at a large scale.

7.2 GDL for streamline representation learning

From the methodological perspective, Verifyber provides an original and principled method for learning on the streamline structure. Unlike images, which have a grid-based regular structure, streamlines are represented by an irregular structure with no orientation and variable length. Due to this irregularity existing machine learning approach on streamlines always resorted to embeddings or heuristic approximations, which lose part of the trajectory information. With Verifyber, such pre-processings are not necessary. The proposed deep neural network has a built-in invariance to the fibers' orientation while being sensitive to the order of points. This property differentiates our model from other pre-existing geometric deep learning models. Moreover, Verifyber, like other GDL models, can deal with streamlines of different lengths. All in all, Verifyber is the first GDL model designed for streamlines.

7.3 Future directions

This thesis constitutes a small step toward characterizing white matter using deep learning. The proposed methods are original but come with some limitations that could be overcome in future studies.

Extending StemSeg to other bundles In StemSeg, we proposed an empirical analysis only focused on the IFOF bundle. However, the stem also exists in other association bundles, such as Uncinate and Arcuate Fascicle. Thus it would be interesting to extend StemSeg also to such bundles. To enable the training of StemSeg on new bundles, it will require a dataset of manually curated stem-ROIs for the bundle at hand. Unfortunately, we cannot estimate how many examples StemSeg requires to be trained, as no ablation on low-data regimes was done. Another aspect that could play a key role in the extension of StemSeg to other bundles is the choice of the input image. In the case of the IFOF stem, the fronto-occipital direction of fibers corresponds to a green region in the simplistic CFA. However, it could be that the stem of other bundles does not align with one of the three main directions, and thus it would be encoded by a uniform color in the CFA volume. Hence, richer representations such as the fODF peaks will need to be tried.

Exploiting pre-trained Verifyber One of the main features of Verifyber is the flexibility to new/refined labelings via retraining. However, the training from scratch could be risky in some situations where the new labeling is available for a limited number of fibers/subjects. An unexplored solution would be to fine-tune fully or partially a pre-trained VF with the new labels available. Otherwise, the pre-training of VF can be exploited to extend the model to new tasks like bundle segmentation. In such cases the last layer of the model must substitute with a new layer encoding the correct output, i.e., number of classes. Then, one can decide if fully fine-tuning the model or just learning the last layer (transfer learning).

Learning relation inter-fibers Verifyber is an effective model for learning the streamline structure, i.e., anatomy/trajectory. We can call this ability intra-fiber learning. However, we know that fibers are related to each other forming bundles. Such inter-fiber relations have not been taken into account in the current approach. Thus, an intriguing open direction is learning latent relations between fibers. While we still do not have a solution to this problem, we believe that a self-supervised task to regularize the learned latent space might help, e.g., latent clustering. An additional challenge of learning inter-fiber relations is the computational complexity. There will be the need to simultaneously process the highest number of fibers to have a comprehensive view of the white matter anatomy.

Bibliography

Alexander, D. C. and Seunarine, K. K. (2010). Mathematics of crossing fibers. In *Diffusion MRI: Theory, Methods, and Applications*, pages 451–464. Oxford University Press New York, NY.

Andersson, J. L. R. and Sotiropoulos, S. N. (2015). Non-parametric representation and prediction of single- and multi-shell diffusion-weighted MRI data using Gaussian processes. *NeuroImage*, 122:166–176.

Andersson, J. L. R. and Sotiropoulos, S. N. (2016). An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. *NeuroImage*, 125:1063–1078.

Astolfi, P., De Benedictis, A., Sarubbo, S., Bertó, G., Olivetti, E., Sona, D., and Avesani, P. (2020a). A Stem-Based Dissection of Inferior Fronto-Occipital Fasciculus with A Deep Learning Model. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 267–270. IEEE.

Astolfi, P., Verhagen, R., Petit, L., Olivetti, E., Masci, J., Boscaini, D., and Avesani, P. (2020b). Tractogram Filtering of Anatomically Non-plausible Fibers with Geometric Deep Learning. In Martel, A. L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M. A., Zhou, S. K., Racoceanu, D., and Joskowicz, L., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, Lecture Notes in Computer Science, pages 291–301, Cham. Springer International Publishing.

Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1):26–41.

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

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., and 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):1–13.

Aydogan, D. B. and Shi, Y. (2015). Track Filtering via Iterative Correction of TDI Topology. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science, pages 20–27, Cham. Springer International Publishing.

Aydogan, D. B. and Shi, Y. (2018). Tracking and validation techniques for topographically organized tractography. *NeuroImage*, 181:64–84.

Bach, F. (2017). Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Bengio, Y. and LeCun, Y., editors, *International Conference on Learning Representations*.

Basser, P. J., Mattiello, J., and LeBihan, D. (1994). MR diffusion tensor spectroscopy and imaging. *Biophysical journal*, 66(1):259–267.

Basser, P. J., Pajevic, S., Pierpaoli, C., Duda, J., and Aldroubi, A. (2000). In vivo fiber tractography using DT-MRI data. *Magnetic Resonance in Medicine*, 44(4):625–632.

Bastiani, M., Shah, N. J., Goebel, R., and Roebroeck, A. (2012). Human cortical connectome reconstruction from diffusion weighted MRI: The effect of tractography algorithm. *NeuroImage*, 62(3):1732–1749.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Berman, M., Triki, A. R., and Blaschko, M. B. (2018). The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4413–4421.

Bertò, G., Bullock, D., Astolfi, P., Hayashi, S., Zigiotto, L., Annicchiarico, L., Corsini, F., De Benedictis, A., Sarubbo, S., Pestilli, F., Avesani, P., and Olivetti, E. (2021). Classifyber, a robust streamline-based linear classifier for white matter bundle segmentation. *NeuroImage*, 224:117402.

Bitar, R., Leung, G., Perng, R., Tadros, S., Moody, A. R., Sarrazin, J., McGregor, C., Christakis, M., Symons, S., and Nelson, A. (2006). MR pulse sequences: What every radiologist wants to know but is afraid to ask. *Radiographics*, 26(2):513–537.

Bonilha, L., Gleichgerrcht, E., Nesland, T., Rorden, C., and Fridriksson, J. (2015). Gray Matter Axonal Connectivity Maps. *Frontiers in Psychiatry*, 6:35.

Bronstein, M. M., Bruna, J., Cohen, T., and Veličković, P. (2021). Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*.

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. (2017). Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Buchanan, C. R., Pernet, C. R., Gorgolewski, K. J., Storkey, A. J., and Bastin, M. E. (2014). Test–retest reliability of structural brain networks from diffusion MRI. *NeuroImage*, 86:231–243.

Calamante, F. (2019). The Seven Deadly Sins of Measuring Brain Structural Connectivity Using Diffusion MRI Streamlines Fibre-Tracking. *Diagnostics*, 9(3):115.

Campanella, M., Ius, T., Skrap, M., and Fadiga, L. (2014). Alterations in fiber pathways reveal brain tumor typology: A diffusion tractography study. *PeerJ*, 2:e497.

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

Catani, M. and Thiebaut de Schotten, M. (2008). A diffusion tensor imaging tractography atlas for virtual in vivo dissections. *Cortex*, 44(8):1105–1132.

Catani, M. and Thiebaut de Schotten, M. (2012). *Atlas of Human Brain Connections*. Oxford University Press, 1 reprint edition.

Chamberland, M., Scherrer, B., Prabhu, S. P., Madsen, J., Fortin, D., Whittingstall, K., Descoteaux, M., and Warfield, S. K. (2017). Active delineation of Meyer's loop using oriented priors through MAGNEtic tractography (MAGNET). *Human Brain Mapping*, 38(1):509–527.

Chandio, B. Q., Risacher, S. L., Pestilli, F., Bullock, D., Yeh, F.-C., Koudoro, S., Rokem, A., Harezlak, J., and Garyfallidis, E. (2020). Bundle analytics, a computational framework for investigating the shapes and profiles of brain pathways across populations. *Scientific Reports*, 10(1):17149.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 424–432. Springer.

Clark, C. A. and Byrnes, T. (2010). DTI and Tractography in Neurosurgical Planning. In *Diffusion MRI: Theory, Methods, and Applications*, chapter Diffusion MRI. Oxford University Press.

Conturo, T. E., Lori, N. F., Cull, T. S., Akbudak, E., Snyder, A. Z., Shimony, J. S., McKinstry, R. C., Burton, H., and Raichle, M. E. (1999). Tracking neuronal fiber pathways in the living human brain. *Proceedings of the National Academy of Sciences of the United States of America*, 96(18):10422–10427.

Côté, M.-A., Garyfallidis, E., Larochelle, H., and 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., and Descoteaux, M. (2013). Tractometer: Towards validation of tractography pipelines. *Medical Image Analysis*, 17(7):844–857.

Daducci, A., Dal Palu, A., Lemkaddem, A., and Thiran, J.-P. (2015). COMMIT: Convex Optimization Modeling for Microstructure Informed Tractography. *IEEE Transactions on Medical Imaging*, 34(1):246–257.

De Benedictis, A., Avesani, P., and Astolfi, P. (2019). STEM-IFOF: Segmentation of the stem region of inferior fronto-occipital fasciculus.

De Benedictis, A., Duffau, H., Paradiso, B., Grandi, E., Balbi, S., Granieri, E., Colarusso, E., Chioffi, F., Marras, C. E., and Sarubbo, S. (2014). Anatomo-functional study of the temporo-parieto-occipital region: Dissection, tractographic and brain mapping evidence from a neurosurgical perspective. *Journal of Anatomy*, 225(2):132–151.

De Benedictis, A., Nocerino, E., Menna, F., Remondino, F., Barbareschi, M., Rozzanigo, U., Corsini, F., Olivetti, E., Marras, C. E., Chioffi, F., Avesani, P., and Sarubbo, S. (2018). Photogrammetry of the Human Brain: A Novel Method for Three-Dimensional Quantitative Exploration of the Structural Connectivity in Neurosurgery and Neurosciences. *World Neurosurgery*, 115:e279–e291.

De Benedictis, A., Petit, L., Descoteaux, M., Marras, C. E., Barbareschi, M., Corsini, F., Dallabona, M., Chioffi, F., and 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(12):4718–4735.

De Benedictis, A., Sarubbo, S., and Duffau, H. (2012). Subcortical surgical anatomy of the lateral frontal region: Human white matter dissection and correlations with functional insights provided by intraoperative direct brain stimulation: Laboratory investigation. *Journal of Neurosurgery*, 117(6):1053–1069.

de Groot, M., Vernooij, M. W., Klein, S., Ikram, M. A., Vos, F. M., Smith, S. M., Niessen, W. J., and Andersson, J. L. (2013). Improving alignment in Tract-based spatial statistics: Evaluation and optimization of image registration. *NeuroImage*, 76:400–411.

Dell'Acqua, F., Rizzo, G., Scifo, P., Clarke, R. A., Scotti, G., and Fazio, F. (2007). A model-based deconvolution approach to solve fiber crossing in diffusion-weighted MR imaging. *IEEE Transactions on Biomedical Engineering*, 54(3):462–472.

Descoteaux, M. (2015). High Angular Resolution Diffusion Imaging (HARDI). In *Wiley Encyclopedia of Electrical and Electronics Engineering*, pages 1–25. American Cancer Society.

Descoteaux, M., Angelino, E., Fitzgibbons, S., and Deriche, R. (2007). Regularized, fast, and robust analytical Q-ball imaging. *Magnetic Resonance in Medicine*, 58(3):497–510.

Descoteaux, M., Deriche, R., Knosche, T. R., and Anwander, A. (2009). Deterministic and Probabilistic Tractography Based on Complex Fibre Orientation Distributions. *IEEE Transactions on Medical Imaging*, 28(2):269–286.

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

Destrieux, C., Fischl, B., Dale, A., and Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, 53(1):1–15.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, pages 4171–4186. Association for Computational Linguistics.

Dong, H., Yang, G., Liu, F., Mo, Y., and Guo, Y. (2017a). Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks. In *Annual Conference on Medical Image Understanding and Analysis*, pages 506–517. Springer.

Dong, X., Zhang, Z., and Srivastava, A. (2017b). Bayesian Tractography Using Geometric Shape Priors. *Frontiers in Neuroscience*, 0.

Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., and Pal, C. (2016). The Importance of Skip Connections in Biomedical Image Segmentation. In Carneiro, G., Mateus, D., Peter, L., Bradley, A., Tavares, J. M. R. S., Belagiannis, V., Papa, J. P., Nascimento, J. C., Loog, M., Lu, Z., Cardoso, J. S., and Cornebise, J., editors, *Deep Learning and Data Labeling for Medical Applications*, Lecture Notes in Computer Science, pages 179–187, Cham. Springer International Publishing.

Duffau, H. (2015). Stimulation mapping of white matter tracts to study brain functional connectivity. *Nature Reviews Neurology*, 11(5):255–265.

Dziedzic, T. A., Balasa, A., Jeżewski, M. P., Michałowski, Ł., and Marchel, A. (2021). White matter dissection with the Klingler technique: A literature review. *Brain Structure and Function*, 226(1):13–47.

Eickhoff, S. B., Yeo, B. T. T., and Genon, S. (2018). Imaging-based parcellations of the human brain. *Nature Reviews Neuroscience*, 19(11):672–686.

Essayed, W. I., Zhang, F., Unadkat, P., Cosgrove, G. R., Golby, A. J., and 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.

Fey, M. and Lenssen, J. E. (2019). Fast Graph Representation Learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.

Figley, T. D., Mortazavi Moghadam, B., Bhullar, N., Kornelsen, J., Courtney, S. M., and Figley, C. R. (2017). Probabilistic White Matter Atlases of Human Auditory, Basal Ganglia, Language, Precuneus, Sensorimotor, Visual and Visuospatial Networks. *Frontiers in Human Neuroscience*, 11:306.

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., and Poupon, C. (2011). Quantitative evaluation of 10 tractography algorithms on a realistic diffusion MR phantom. *NeuroImage*, 56(1):220–234.

Fischl, B. (2012). FreeSurfer. *NeuroImage*, 62(2):774–781.

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., and Dale, A. M. (2002). Whole Brain Segmentation: Automated Labeling of Neuroanatomical Structures in the Human Brain. *Neuron*, 33(3):341–355.

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

Forkel, S. J., Thiebaut de Schotten, M., Kawadler, J. M., Dell'Acqua, F., Danek, A., and Catani, M. (2014). The anatomy of fronto-occipital connections from early blunt dissections to contemporary tractography. *Cortex*, 56:73–84.

Frigo, M., Deslauriers-Gauthier, S., Parker, D., Aziz Ould Ismail, A., John Kim, J., Verma, R., and Deriche, R. (2020). Diffusion MRI tractography filtering techniques change the topology of structural connectomes. *Journal of Neural Engineering*, 17(6):065002.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202.

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

Garyfallidis, E., Brett, M., Correia, M. M., Williams, G. B., and Nimmo-Smith, I. (2012). QuickBundles, a Method for Tractography Simplification. *Frontiers in Neuroscience*, 6.

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

Garyfallidis, E., Ocegueda, O., Wassermann, D., and Descoteaux, M. (2015). Robust and efficient linear registration of white-matter fascicles in the space of streamlines. *NeuroImage*, 117:124–140.

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., and Innocenti, G. M. (2020). On the cortical connectivity in the macaque brain: A comparison of diffusion tractography and histological tracing data. *NeuroImage*, 221:117201.

Girard, G., Caruyer, E., Rafael-Patino, J., Pizzolato, M., Truffet, R., and Thiran, J.-P. (2021). Diffusion-simulated connectivity challenge.

Girard, G., Whittingstall, K., Deriche, R., and Descoteaux, M. (2014). Towards quantitative connectivity analysis: Reducing tractography biases. *NeuroImage*, 98:266–278.

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., and Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178.

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., and Jenkinson,

M. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, 80:105–124.

Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610.

Gray, H. (1918). *Anatomy of the Human Body.* Lea & Febiger,, Philadelphia,, 20th ed., by warren h. lewis edition.

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

Gupta, V., Thomopoulos, S. I., Corbin, C. K., Rashid, F., and 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)*, pages 708–711.

Gupta, V., Thomopoulos, S. I., Rashid, F. M., and Thompson, P. M. (2017). FiberNET: An Ensemble Deep Learning Framework for Clustering White Matter Fibers. In Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D. L., and Duchesne, S., editors, *Medical Image Computing and Computer Assisted Intervention − MICCAI 2017*, Lecture Notes in Computer Science, pages 548–555, Cham. Springer International Publishing.

Hansen, C. B., Yang, Q., Lyu, I., Rheault, F., Kerley, C., Chandio, B. Q., Fadnavis, S., Williams, O., Shafer, A. T., Resnick, S. M., Zald, D. H., Cutting, L. E., Taylor, W. D., Boyd, B., Garyfallidis, E., Anderson, A. W., Descoteaux, M., Landman, B. A., and Schilling, K. G. (2021). Pandora: 4-D White Matter Bundle Population-Based Atlases Derived from Diffusion MRI Fiber Tractography. *Neuroinformatics*, 19(3):447–460.

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., and Petit, L. (2017). Revisiting the human uncinate fasciculus, its subcomponents and asymmetries with stem-based tractography and microdissection validation. *Brain Structure and Function*, pages 1–18.

Hau, J., Sarubbo, S., Perchey, G., Crivello, F., Zago, L., Mellet, E., Jobard, G., Joliot, M., Mazoyer, B. M., Tzourio-Mazoyer, N., and Petit, L. (2016). Cortical Terminations of the Inferior Fronto-Occipital and Uncinate Fasciculi: Anatomical Stem-Based Virtual Dissection. *Frontiers in Neuroanatomy*, 10.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Hebb, D. O. (1949). *The Organization of Behavior; a Neuropsychological Theory*. The Organization of Behavior; a Neuropsychological Theory. Wiley, Oxford, England.

Henderson, F., Abdullah, K. G., Verma, R., and Brem, S. (2020). Tractography and the connectome in neurosurgical treatment of gliomas: The premise, the progress, and the potential. *Neurosurgical Focus*, 48(2):E6.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

House, E. L. and Pansky, B. (1960). A functional approach to neuroanatomy. *Academic Medicine*, 35(11):1067–1068.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Hubel, D. H. and Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, 148(3):574–591.

Jbabdi, S., Woolrich, M., Andersson, J., and Behrens, T. (2007). A Bayesian framework for global tractography. *NeuroImage*, 37:116–29.

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

Jeurissen, B., Descoteaux, M., Mori, S., and Leemans, A. (2017). Diffusion MRI fiber tractography of the brain. *NMR in Biomedicine*, 32(4):e3785.

Jones, D. K. (2008). Studying connections in the living human brain with diffusion MRI. *Cortex*, 44(8):936–952.

Jones, D. K. (2010). *Diffusion Mri*. Oxford University Press.

Jörgens, D., Descoteaux, M., and Moreno, R. (2021). Challenges for Tractogram Filtering. In Özarslan, E., Schultz, T., Zhang, E., and Fuster, A., editors, *Anisotropy Across Fields and Scales*, pages 149–168, Cham. Springer International Publishing.

Khatami, M., Schmidt-Wilcke, T., Sundgren, P. C., Abbasloo, A., Schölkopf, B., and Schultz, T. (2017). BundleMAP: Anatomically localized classification, regression, and hypothesis testing in diffusion MRI. *Pattern Recognition*, 63:593–600.

Kier, E. L., Staib, L. H., Davis, L. M., and Bronen, R. A. (2004). MR imaging of the temporal stem: Anatomic dissection tractography of the uncinate fasciculus, inferior occipitofrontal fasciculus, and Meyer's loop of the optic radiation. *American Journal of Neuroradiology*, 25(5):677–691.

Kipf, T. N. and Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations*. OpenReview.net.

Klingler, J., Klingler, J., and Klingler, J. P. (1935). *Erleichterung Der Makroskopischen Präparation Des Gehirn Durch Den Gefrierprozess*. Orell Füssli.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25:1097–1105.

Kruper, J., Yeatman, J. D., Richie-Halford, A., Bloom, D., Grotheer, M., Caffarra, S., Kiar, G., Karipidis, I. I., Roy, E., Chandio, B. Q., Garyfalldis, E., and Rokem, A. (2021). Evaluating the reliability of human brain white matter tractometry. *bioRxiv*, page 2021.02.24.432740.

Le Bihan, D., Breton, E., Lallemand, D., Grenier, P., Cabanis, E., and Laval-Jeantet, M. (1986). MR imaging of intravoxel incoherent motions: Application to diffusion and perfusion in neurologic disorders. *Radiology*, 161(2):401–407.

LeCun, Y. and Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The Handbook of Brain Theory and Neural Networks*, 3361(10):1995.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551.

Lee, M.-H., O'Hara, N., Sonoda, M., Kuroda, N., Juhasz, C., Asano, E., Dong, M., and 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(11):3151–3162.

Legarreta, J. H., Petit, L., Rheault, F., Theaud, G., Lemaire, C., Descoteaux, M., and Jodoin, P.-M. (2021). Filtering in Tractography using Autoencoders (FINTA). *Medical Image Analysis*, page 102126.

Leitner, Y., Travis, K. E., Ben-Shachar, M., Yeom, K. W., and Feldman, H. M. (2015). Tract Profiles of the Cerebellar White Matter Pathways in Children and Adolescents. *The Cerebellum*, 14(6):613–623.

Lemkaddem, A., Skiöldebrand, D., Dal Palú, A., Thiran, J.-P., and Daducci, A. (2014). Global tractography with embedded anatomical priors for quantitative connectivity analysis. *Frontiers in Neurology*, 5:232.

Li, B., de Groot, M., Steketee, R. M., Meijboom, R., Smits, M., Vernooij, M. W., Ikram, M. A., Liu, J., Niessen, W. J., and Bron, E. E. (2020). Neuro4Neuro: A neural network approach for neural tract segmentation using large-scale population-based diffusion imaging. *NeuroImage*, 218:116993.

Lin, G., Milan, A., Shen, C., and Reid, I. (2017). Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1925–1934.

Lu, Q., Li, Y., and Ye, C. (2021). Volumetric white matter tract segmentation with nested self-supervised learning using sequential pretext tasks. *Medical Image Analysis*, 72:102094.

Maffei, C., Jovicich, J., De Benedictis, A., Corsini, F., Barbareschi, M., Chioffi, F., and 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(1):449–459.

Maffei, C., Lee, C., Planich, M., Ramprasad, M., Ravi, N., Trainor, D., Urban, Z., Kim, M., Jones, R. J., Henin, A., Hofmann, S. G., Pizzagalli, D. A., Auerbach, R. P., Gabrieli, J. D. E., Whitfield-Gabrieli, S., Greve, D. N., Haber, S. N., and Yendiki, A. (2021). Using diffusion MRI data acquired with ultra-high gradients to improve tractography in routine-quality data. *bioRxiv*, page 2021.06.28.450265.

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., Auría, A., Esteban, O., Lemkaddem, A., Thiran, J.-P., Cetingul, H. E., Odry, B. L., Mailhe, B., Nadar, M. S., Pizzagalli, F., Prasad, G., Villalon-Reina, J. E., Galvis, J., Thompson, P. M., Requejo, F. D. S., Laguna, P. L., Lacerda, L. M., Barrett, R., Dell'Acqua, F., Catani, M., Petit, L., Caruyer, E., Daducci, A., Dyrby, T. B., Holland-Letz, T., Hilgetag, C. C., Stieltjes, B., and Descoteaux, M. (2017). The challenge of mapping the human connectome based on diffusion tractography. *Nature Communications*, 8(1):1349.

Mandonnet, E., Sarubbo, S., and Petit, L. (2018). The Nomenclature of Human White Matter Association Pathways: Proposal for a Systematic Taxonomic Anatomical Classification. *Frontiers in Neuroanatomy*, 0.

Mangin, J. F., Fillard, P., Cointepas, Y., Le Bihan, D., Frouin, V., and Poupon, C. (2013). Toward global tractography. *NeuroImage*, 80:290–296.

Martino, J., Brogna, C., Robles, S. G., Vergani, F., and Duffau, H. (2010). Anatomic dissection of the inferior fronto-occipital fasciculus revisited in the lights of brain stimulation data. *Cortex*, 46(5):691–699.

Martino, J., De Witt Hamer, P. C., Vergani, F., Brogna, C., de Lucas, E. M., Vázquez-Barquero, A., García-Porrero, J. A., and Duffau, H. (2011). Cortex-sparing fiber dissection: An improved method for the study of white matter anatomy in the human brain. *Journal of Anatomy*, 219(4):531–541.

Masci, J., Rodolà, E., Boscaini, D., Bronstein, M. M., and Li, H. (2016). Geometric deep learning. In *SIGGRAPH ASIA 2016 Courses*, SA '16, pages 1–50, Macau. Association for Computing Machinery.

Mazoyer, B., Mellet, E., Perchey, G., Zago, L., Crivello, F., Jobard, G., Delcroix, N., Vigneau, M., Leroux, G., Petit, L., Joliot, M., and Tzourio-Mazoyer, N. (2016). BIL&GIN: A neuroimaging, cognitive, behavioral, and genetic database for the study of human brain lateralization. *NeuroImage*, 124:1225–1231.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133.

Meynert, T. (1885). *Psychiatry: A Clinical Treatise on Diseases of the Fore-Brain Based Upon a Study of Its Structure, Functions, and Nutrition.* GP Putnam's Sons.

Milchenko, M. and Marcus, D. (2013). Obscuring surface anatomy in volumetric imaging data. *Neuroinformatics*, 11(1):65–75.

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., Bartsch, A. J., Jbabdi, S., Sotiropoulos, S. N., Andersson, J. L. R., Griffanti, L., Douaud, G., Okell, T. W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., Matthews, P. M., and Smith, S. M. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*, 19(11):1523–1536.

Minati, L. and Węglarz, W. P. (2007). Physical foundations, models, and methods of diffusion magnetic resonance imaging of the brain: A review. *Concepts in Magnetic Resonance Part A*, 30(5):278–307.

Mitchell, T. M. (1980). *The Need for Biases in Learning Generalizations.* Department of Computer Science, Laboratory for Computer Science Research ….

Mori, S., Crain, B. J., Chacko, V. P., and Van Zijl, P. (1999). Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging. *Annals of Neurology*, 45(2):265–269.

Mori, S., Wakana, S., Van Zijl, P. C. M., and Nagae-Poetscher, L. M. (2005). *MRI Atlas of Human White Matter*, volume 16. Am Soc Neuroradiology, Amsterdam, The Netherlands.

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., and Landman, B. A. (2020). Tractography reproducibility challenge with empirical data (TraCED): The 2017 ISMRM diffusion study group challenge. *Journal of Magnetic Resonance Imaging*, 51(1):234–249.

Neher, P. F., Descoteaux, M., Houde, J.-C., Stieltjes, B., and 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(1):287–305.

Neher, P. F., Stieltjes, B., and Maier-Hein, K. H. (2018). Anchor-Constrained Plausibility (ACP): A Novel Concept for Assessing Tractography and Reducing False-Positives. In Frangi, A. F., Schnabel, J. A., Davatzikos, C., Alberola-López, C., and Fichtinger, G., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, Lecture Notes in Computer Science, pages 20–27, Cham. Springer International Publishing.

Nie, X. and Shi, Y. (2019). Topographic Filtering of Tractograms as Vector Field Flows. In Shen, D., Liu, T., Peters, T. M., Staib, L. H., Essert, C., Zhou, S., Yap, P.-T., and Khan, A., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, Lecture Notes in Computer Science, pages 564–572, Cham. Springer International Publishing.

Norton, I., Essayed, W. I., Zhang, F., Pujol, S., Yarmarkovich, A., Golby, A. J., Kindlmann, G., Wassermann, D., Estepar, R. S. J., Rathi, Y., Pieper, S., Kikinis, R., Johnson, H. J., Westin, C.-F., and O'Donnell, L. J. (2017). SlicerDMRI: Open Source Diffusion MRI Software for Brain Cancer Research. *Cancer Research*, 77(21):e101–e103.

Ocampo-Pineda, M., Schiavi, S., Rheault, F., Girard, G., Petit, L., Descoteaux, M., and Daducci, A. (2021). Hierarchical Microstructure Informed Tractography. *Brain Connectivity*.

O'Donnell, L. J., Kubicki, M., Shenton, M. E., Dreusicke, M. H., Grimson, and Westin, C.-F. (2006). A method for clustering white matter fiber tracts. *American Journal of Neuroradiology*, 27(5):1032–1036.

O'Donnell, L. J., Wells, W. M., Golby, A. J., and Westin, C.-F. (2012). Unbiased Groupwise Registration of White Matter Tractography. In Ayache, N., Delingette, H., Golland, P., and Mori, K., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012*, Lecture Notes in Computer Science, pages 123–130, Berlin, Heidelberg. Springer.

O'Donnell, L. J. and Westin, C.-F. (2007). Automatic Tractography Segmentation Using a High-Dimensional White Matter Atlas. *IEEE Transactions on Medical Imaging*, 26(11):1562–1575.

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

Olivetti, E. and Avesani, P. (2011). Supervised segmentation of fiber tracts. In *Proceedings of the First International Conference on Similarity-based Pattern Recognition*, SIMBAD'11, pages 261–274, Berlin, Heidelberg. Springer-Verlag.

Olivetti, E., Gori, P., Astolfi, P., Bertó, G., and Avesani, P. (2020). Nonlinear Alignment of Whole Tractograms with the Linear Assignment Problem. In *International Workshop on Biomedical Image Registration*, pages 3–11. Springer.

Olivetti, E., Sharmin, N., and Avesani, P. (2016). Alignment of Tractograms As Graph Matching. *Frontiers in Neuroscience*, 10.

Orringer, D. A., Golby, A., and Jolesz, F. (2012). Neuronavigation in the surgical management of brain tumors: Current and future trends. *Expert Review of Medical Devices*, 9(5):491–500.

Panesar, S. S., Yeh, F.-C., Deibert, C. P., Fernandes-Cabral, D., Rowthu, V., Celtikci, P., Celtikci, E., Hula, W. D., Pathak, S., and Fernández-Miranda, J. C. (2017). A diffusion spectrum imaging-based tractographic study into the anatomical subdivision and cortical connectivity of the ventral external capsule: Uncinate and inferior fronto-occipital fascicles. *Neuroradiology*, 59(10):971–987.

Pang, G., Shen, C., Cao, L., and Hengel, A. V. D. (2021). Deep learning for anomaly detection: A review. *ACM Computing Surveys (CSUR)*, 54(2):1–38.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., and Antiga, L. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.

Pestilli, F., Yeatman, J. D., Rokem, A., Kay, K. N., and Wandell, B. A. (2014). Evaluation and statistical inference for human connectomes. *Nature Methods*, 11(10):1058–1063.

Petit, L., Rheault, F., Descoteaux, M., and Tzourio-Mazoyer, N. (2019). Half of the streamlines built in a whole human brain tractogram is anatomically uninterpretable. In *Organization for Human Brain Mapping (OHBM)*. F1000 Research Limited.

Pierpaoli, C. (2010). Artifacts in diffusion MRI. In *Diffusion MRI: Theory, Methods, and Applications*, pages 303–317. Oxford University Press, New York.

Pierpaoli, C., Barnett, A., Pajevic, S., Chen, R., Penix, L., Virta, A., and Basser, P. (2001). Water Diffusion Changes in Wallerian Degeneration and Their Dependence on White Matter Architecture. *NeuroImage*, 13(6):1174–1185.

Pierpaoli, C., Jezzard, P., Basser, P. J., Barnett, A., and Di Chiro, G. (1996). Diffusion tensor MR imaging of the human brain. *Radiology*, 201(3):637–648.

Porro-Muñoz, D., Olivetti, E., Sharmin, N., Nguyen, T., Garyfallidis, E., and Avesani, P. (2015). Tractome: A visual data mining tool for brain connectivity analysis. *Data Mining and Knowledge Discovery*, 29(5):1258–1279.

Presseau, C., Jodoin, P.-M., Houde, J.-C., and Descoteaux, M. (2015). A new compression format for fiber tracking datasets. *NeuroImage*, 109:73–83.

Qi, C. R., Su, H., Mo, K., and 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*, pages 652–660.

Raffelt, D., Tournier, J. D., Rose, S., Ridgway, G. R., Henderson, R., Crozier, S., Salvado, O., and Connelly, A. (2012). Apparent Fibre Density: A novel measure for the analysis of diffusion-weighted magnetic resonance images. *NeuroImage*, 59(4):3976–3994.

Reddy, C. P. and Rathi, Y. (2016). Joint Multi-Fiber NODDI Parameter Estimation and Tractography Using the Unscented Information Filter. *Frontiers in Neuroscience*, 10.

Reisert, M., Coenen, V. A., Kaller, C., Egger, K., and Skibbe, H. (2018). HAMLET: Hierarchical Harmonic Filters for Learning Tracts from Diffusion MRI. *arXiv:1807.01068 [cs]*.

Rheault, F., Bayrak, R. G., Wang, X., Schilling, K. G., Greer, J. M., Hansen, C. B., Kerley, C., Ramadass, K., Remedios, L. W., Blaber, J. A., Williams, O., Beason-Held,

L. L., Resnick, S. M., Rogers, B. P., and Landman, B. A. (2022). TractEM: Evaluation of protocols for deterministic tractography white matter atlas. *Magnetic Resonance Imaging*, 85:44–56.

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., and Descoteaux, M. (2020a). Tractostorm: The what, why, and how of tractography dissection reproducibility. *Human Brain Mapping*, 41(7):1859–1874.

Rheault, F., Houde, J.-C., Goyette, N., Morency, F., and Descoteaux, M. (2016). MI-Brain, a software to handle tractograms and perform interactive virtual dissection. In *Proceedings of International Society of Magnetic Resonance in Medicine (ISMRM)*.

Rheault, F., Poulin, P., Valcourt Caron, A., St-Onge, E., and Descoteaux, M. (2020b). Common misconceptions, hidden biases and modern challenges of dMRI tractography. *Journal of Neural Engineering*, 17(1):011001.

Rheault, F., St-Onge, E., Sidhu, J., Maier-Hein, K., Tzourio-Mazoyer, N., Petit, L., and Descoteaux, M. (2019). Bundle-specific tractography with incorporated anatomical and orientational priors. *NeuroImage*, 186:382–398.

Robbins, H. and Monro, S. (1951). A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400–407.

Rolls, E. T., Joliot, M., and Tzourio-Mazoyer, N. (2015). Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas. *NeuroImage*, 122:1–5.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science, pages 234–241. Springer International Publishing.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Rosenblatt, F. (1961). *Principles of Neurodynamics. Perceptrons and the Theory of Brain Mechanisms*. Cornell Aeronautical Lab Inc Buffalo NY.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Salat, D. H., Greve, D. N., Pacheco, J. L., Quinn, B. T., Helmer, K. G., Buckner, R. L., and Fischl, B. (2009). Regional white matter volume differences in nondemented aging and Alzheimer's disease. *NeuroImage*, 44(4):1247–1258.

Sarubbo, S., De Benedictis, A., Maldonado, I., Basso, G., and Duffau, H. (2013). Frontal terminations for the inferior fronto-occipital fascicle: Anatomical dissection, DTI study and functional considerations on a multi-component bundle. *Brain Structure and Function*, 218(1):21–37.

Sarubbo, S., Petit, L., De Benedictis, A., Chioffi, F., Ptito, M., and Dyrby, T. B. (2019). Uncovering the inferior fronto-occipital fascicle and its topological organization in non-human primates: The missing connection for language evolution. *Brain Structure and Function*, 224(4):1553–1567.

Sarwar, T., Ramamohanarao, K., and Zalesky, A. (2019). Mapping connectomes with diffusion MRI: Deterministic or probabilistic tractography? *Magnetic Resonance in Medicine*, 81(2):1368–1384.

Schiavi, S., Ocampo-Pineda, M., Barakovic, M., Petit, L., Descoteaux, M., Thiran, J.-P., and Daducci, A. (2020). A new method for accurate in vivo mapping of human brain connections using microstructural and anatomical information. *Science Advances*, 6(31):eaba8245.

Schiffler, P., Tenberge, J.-G., Wiendl, H., and Meuth, S. G. (2017). Cortex Parcellation Associated Whole White Matter Parcellation in Individual Subjects. *Frontiers in Human Neuroscience*, 0.

Schilling, K. G., Daducci, A., Maier-Hein, K., Poupon, C., Houde, J.-C., Nath, V., Anderson, A. W., Landman, B. A., and Descoteaux, M. (2019a). Challenges in diffusion MRI tractography – Lessons learned from international benchmark competitions. *Magnetic Resonance Imaging*, 57:194–209.

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., Fischi, 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., and Landman, B. A. (2019b). Limits to anatomical accuracy of diffusion tractography using modern approaches. *NeuroImage*, 185:1–11.

Schilling, K. G., Petit, L., Rheault, F., Remedios, S., Pierpaoli, C., Anderson, A. W., Landman, B. A., and 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(8):2387–2402.

Schilling, K. G., Rheault, F., Petit, L., Hansen, C. B., Nath, V., Yeh, F.-C., Girard, G., Barakovic, M., Rafael-Patino, J., Yu, T., Fischi-Gomez, E., Pizzolato, M., Ocampo-Pineda, M., Schiavi, S., Canales-Rodríguez, E. J., Daducci, A., Granziera, C., Innocenti, G., Thiran, J.-P., Mancini, L., Wastling, S., Cocozza, S., Petracca, M., Pontillo, G., Mancini, M., Vos, S. B., Vakharia, V. N., Duncan, J. S., Melero, H., Manzanedo, L., Sanz-Morales, E., Peña-Melián, Á., Calamante, F., Attyé, A., Cabeen, R. P., Korobova, L., Toga, A. W., Vijayakumari, A. A., Parker, D., Verma, R., Radwan, A., Sunaert, S., Emsell, L., De Luca, A., Leemans, A., Bajada, C. J., Haroon, H., Azadbakht, H., Chamberland, M., Genc, S., Tax, C. M. W., Yeh, P.-H., Srikanchana, R., Mcknight, C. D., Yang, J. Y.-M., Chen, J., Kelly, C. E., Yeh, C.-H., Cochereau, J., Maller, J. J., Welton, T., Almairac, F., Seunarine, K. K., Clark, C. A., Zhang, F., Makris, N., Golby, A., Rathi, Y., O'Donnell, L. J., Xia, Y., Aydogan, D. B., Shi, Y., Fernandes, F. G., Raemaekers, M., Warrington, S., Michielse, S., Ramírez-Manzanares, A., Concha, L., Aranda, R., Meraz, M. R., Lerma-Usabiaga, G., Roitman, L., Fekonja, L. S., Calarco, N., Joseph, M., Nakua, H., Voineskos, A. N., Karan, P., Grenier, G., Legarreta, J. H., Adluru, N., Nair, V. A., Prabhakaran, V., Alexander, A. L., Kamagata, K., Saito, Y., Uchida, W., Andica, C., Abe, M., Bayrak, R. G., Wheeler-Kingshott, C. A. M. G., D'Angelo, E., Palesi, F., Savini, G., Rolandi, N., Guevara, P., Houenou, J., López-López, N., Mangin, J.-F., Poupon, C., Román, C., Vázquez, A., Maffei, C., Arantes, M., Andrade, J. P., Silva, S. M., Calhoun, V. D., Caverzasi, E., Sacco, S., Lauricella, M., Pestilli, F., Bullock, D., Zhan, Y., Brignoni-Perez, E., Lebel, C., Reynolds, J. E., Nestrasil, I., Labounek, R., Lenglet, C., Paulson, A., Aulicka, S., Heilbronner, S. R., Heuer, K., Chandio, B. Q., Guaje, J., Tang, W., Garyfallidis, E., Raja, R., Anderson, A. W., Landman, B. A., and Descoteaux, M. (2021). Tractography dissection variability: What happens when 42 groups dissect 14 white matter bundles on the same dataset? *NeuroImage*, 243:118502.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117.

Sharmin, N., Olivetti, E., and Avesani, P. (2016). Alignment of Tractograms as Linear Assignment Problem. In Fuster, A., Ghosh, A., Kaden, E., Rathi, Y., and Reisert, M., editors, *Computational Diffusion MRI*, Mathematics and Visualization, pages 109–120. Springer International Publishing.

Sharmin, N., Olivetti, E., and Avesani, P. (2018). White Matter Tract Segmentation as Multiple Linear Assignment Problems. *Frontiers in Neuroscience*, 11.

Siless, V., Chang, K., Fischl, B., and Yendiki, A. (2018). AnatomiCuts: Hierarchical clustering of tractography streamlines based on anatomical similarity. *NeuroImage*, 166:32–45.

Smith, R., Raffelt, D., Tournier, J.-D., and Connelly, A. (2020a). Quantitative streamlines tractography: Methods and inter-subject normalisation.

Smith, R. E., Calamante, F., and Connelly, A. (2020b). Notes on "A cautionary note on the use of SIFT in pathological connectomes". *Magnetic Resonance in Medicine*, 84(5):2303–2307.

Smith, R. E., Tournier, J.-D., Calamante, F., and Connelly, A. (2012). Anatomically-constrained tractography: Improved diffusion MRI streamlines tractography through effective use of anatomical information. *NeuroImage*, 62(3):1924–1938.

Smith, R. E., Tournier, J.-D., Calamante, F., and Connelly, A. (2015a). The effects of SIFT on the reproducibility and biological accuracy of the structural connectome. *NeuroImage*, 104:253–265.

Smith, R. E., Tournier, J.-D., Calamante, F., and Connelly, A. (2015b). SIFT2: Enabling dense quantitative assessment of brain white matter connectivity using streamlines tractography. *NeuroImage*, 119:338–351.

Smith, R. E., Tournier, J.-D. D., Calamante, F., and Connelly, A. (2013). SIFT: Spherical-deconvolution informed filtering of tractograms. *NeuroImage*, 67:298–312.

Sotiropoulos, S. N., Moeller, S., Jbabdi, S., Xu, J., Andersson, J. L., Auerbach, E. J., Yacoub, E., Feinberg, D., Setsompop, K., Wald, L. L., and Others (2013). Effects of image reconstruction on fiber orientation mapping from multichannel diffusion MRI: Reducing the noise floor using SENSE. *Magnetic Resonance in Medicine*, 70(6):1682–1689.

Sporns, O. (2013). The human connectome: Origins and challenges. *Neuroimage*, 80:53–61.

Sporns, O., Tononi, G., and Kötter, R. (2005). The Human Connectome: A Structural Description of the Human Brain. *PLoS Computational Biology*, 1(4):e42+.

Srivastava, R. K., Greff, K., and Schmidhuber, J. (2015). Highway networks. *arXiv preprint arXiv:1505.00387*.

Sydnor, V. J., Rivas-Grajales, A. M., Lyall, A. E., Zhang, F., Bouix, S., Karmacharya, S., Shenton, M. E., Westin, C.-F., Makris, N., Wassermann, D., O'Donnell, L. J., and Kubicki, M. (2018). A comparison of three fiber tract delineation methods and their impact on white matter analysis. *NeuroImage*, 178:318–331.

Takemura, H., Caiafa, C. F., Wandell, B. A., and Pestilli, F. (2016). Ensemble Tractography. *PLoS computational biology*, 12(2).

Thiebaut de Schotten, M., Ffytche, D. H., Bizzi, A., Dell'Acqua, F., Allin, M., Walshe, M., Murray, R., Williams, S. C., Murphy, D. G. M., and Catani, M. (2011). Atlasing location, asymmetry and inter-subject variability of white matter tracts in the human brain with MR diffusion tractography. *NeuroImage*, 54(1):49–59.

Thomas, C., Ye, F. Q., Irfanoglu, M. O., Modi, P., Saleem, K. S., Leopold, D. A., and 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(46):16574–16579.

Toga, A. W., Clark, K. A., Thompson, P. M., Shattuck, D. W., and Van Horn, J. D. (2012). Mapping the human connectome. *Neurosurgery*, 71(1):1.

Tournier, J., Calamante, F., Connelly, A., and Others (2012). MRtrix: Diffusion tractography in crossing fiber regions. *International Journal of Imaging Systems and Technology*, 22(1):53–66.

Tournier, J.-D. (2019). Diffusion MRI in the brain – Theory and concepts. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 112–113:1–16.

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

Tournier, J. D., Calamante, F., and Connelly, A. (2010). Improved probabilistic streamlines tractography by 2nd order integration over fibre orientation distributions. In *Proceedings of International Society of Magnetic Resonance in Medicine (ISMRM)*, page 1670.

Tournier, J.-D., Calamante, F., Gadian, D. G., and Connelly, A. (2004). Direct estimation of the fiber orientation density function from diffusion-weighted MRI data using spherical deconvolution. *NeuroImage*, 23(3):1176–1185.

Tournier, J.-D., Smith, R., Raffelt, D., Tabbara, R., Dhollander, T., Pietsch, M., Christiaens, D., Jeurissen, B., Yeh, C.-H., and Connelly, A. (2019). MRtrix3: A fast, flexible and open software framework for medical image processing and visualisation. *NeuroImage*, 202:116137.

Tuch, D. S. (2004). Q-ball imaging. *Magnetic Resonance in Medicine*, 52(6):1358–1372.

Turken, Umit, and Dronkers, N. F. (2011). The Neural Architecture of the Language Comprehension Network: Converging Evidence from Lesion and Connectivity Analyses. *Frontiers in Systems Neuroscience*, 0.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in

SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, 15(1):273–289.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11).

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., and Ugurbil, K. (2013). The WU-Minn Human Connectome Project: An overview. *NeuroImage*, 80:62–79.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2018). Graph Attention Networks. In *International Conference on Learning Representations*.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A., and Bottou, L. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(12).

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

Wang, J., Aydogan, D. B., Varma, R., Toga, A. W., and Shi, Y. (2018). Modeling topographic regularity in structural brain connectivity with application to tractogram filtering. *NeuroImage*, 183:87–98.

Wang, R., Benner, T., Sorensen, A. G., and 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)*, volume 15, Berlin. Springer.

Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., and Solomon, J. M. (2019). Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics*, 38(5):146.

Warrington, S., Bryant, K. L., Khrapitchev, A. A., Sallet, J., Charquero-Ballester, M., Douaud, G., Jbabdi, S., Mars, R. B., and Sotiropoulos, S. N. (2020). XTRACT - Standardised protocols for automated tractography in the human and macaque brain. *NeuroImage*, 217:116923.

Wassermann, D., Makris, N., Rathi, Y., Shenton, M., Kikinis, R., Kubicki, M., and Westin, C.-F. F. (2013). On describing human white matter anatomy: The white matter query language. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, 16(Pt 1):647–654.

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

Wassermann, D., Rathi, Y., Bouix, S., Kubicki, M., Kikinis, R., Shenton, M., and Westin, C.-F. F. (2011). White matter bundle registration and population analysis based on Gaussian processes. In *Information Processing in Medical Imaging*, volume 22, pages 320–332.

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

Wasserthal, J., Neher, P. F., Hirjak, D., and Maier-Hein, K. H. (2019). Combined tract segmentation and orientation mapping for bundle-specific tractography. *Medical Image Analysis*, 58:101559.

Wedeen, V. J., Hagmann, P., Tseng, W.-Y. I., Reese, T. G., and Weisskoff, R. M. (2005). Mapping complex tissue architecture with diffusion spectrum magnetic resonance imaging. *Magnetic Resonance in Medicine*, 54(6):1377–1386.

Werbos, P. J. (1988). Generalization of backpropagation with application to a recurrent gas market model. *Neural Networks*, 1(4):339–356.

Wu, Y., Sun, D., Wang, Y., and Wang, Y. (2016). Subcomponents and Connectivity of the Inferior Fronto-Occipital Fasciculus Revealed by Diffusion Spectrum Imaging Fiber Tracking. *Frontiers in Neuroanatomy*, 10.

Xia, Y. and Shi, Y. (2020). Groupwise track filtering via iterative message passing and pruning. *NeuroImage*, 221:117147.

Xu, H., Dong, M., Lee, M.-H., O'Hara, N., Asano, E., and 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(8):1910–1922.

Yang, J. Y.-M., Yeh, C.-H., Poupon, C., and Calamante, F. (2021). Diffusion MRI tractography for neurosurgery: The basics, current state, technical reliability and challenges. *Physics in Medicine & Biology*.

Yeatman, J. D., Dougherty, R. F., Myall, N. J., Wandell, B. A., and Feldman, H. M. (2012). Tract Profiles of White Matter Properties: Automating Fiber-Tract Quantification. *PLoS ONE*, 7(11):e49790.

Yeh, C.-H., Jones, D. K., Liang, X., Descoteaux, M., and Connelly, A. (2020). Mapping Structural Connectivity Using Diffusion MRI: Challenges and Opportunities. *Journal of Magnetic Resonance Imaging*, page jmri.27188.

Yeh, F.-c. (2021). DSI Studio. Zenodo.

Yeh, F.-C., Irimia, A., Bastos, D. C. d. A., and Golby, A. J. (2021). Tractography methods and findings in brain tumors and traumatic brain injury. *NeuroImage*, 245:118651.

Yeh, F.-C., Panesar, S., Barrios, J., Fernandes, D., Abhinav, K., Meola, A., and Fernandez-Miranda, J. C. (2019). Automatic Removal of False Connections in Diffusion MRI Tractography Using Topology-Informed Pruning (TIP). *Neurotherapeutics*, 16(1):52–58.

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

Yeh, F.-C. and Tseng, W.-Y. I. (2011). NTU-90: A high angular resolution brain atlas constructed by q-space diffeomorphic reconstruction. *NeuroImage*, 58(1):91–99.

Yendiki, A., Panneck, P., Srinivasan, P., Stevens, A., Zöllei, L., Augustinack, J., Wang, R., Salat, D., Ehrlich, S., Behrens, T., Jbabdi, S., Gollub, R., and 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.

Yushkevich, P. A., Piven, J., Cody Hazlett, H., Gimpel Smith, R., Ho, S., Gee, J. C., and Gerig, G. (2006). User-Guided 3D Active Contour Segmentation of Anatomical Structures: Significantly Improved Efficiency and Reliability. *Neuroimage*, 31(3):1116–1128.

Zalesky, A., Fornito, A., Cocchi, L., Gollo, L. L., van den Heuvel, M. P., and Breakspear, M. (2016). Connectome sensitivity or specificity: Which is more important? *NeuroImage*, 142:407–420.

Zalesky, A., Sarwar, T., and Kotagiri, R. (2020a). SIFT in pathological connectomes: Follow-up response to Smith and colleagues. *Magnetic Resonance in Medicine*, 84(5):2308–2311.

Zalesky, A., Sarwar, T., and Ramamohanarao, K. (2020b). A cautionary note on the use of SIFT in pathological connectomes. *Magnetic Resonance in Medicine*, 83(3):791–794.

Zemmoura, I., Blanchard, E., Raynal, P.-I., Rousselot-Denis, C., Destrieux, C., and Velut, S. (2016). How Klingler's dissection permits exploration of brain structural connectivity? An electron microscopy study of human white matter. *Brain Structure and Function*, 221(5):2477–2486.

Zhang, F., Cetin Karayumak, S., Hoffmann, N., Rathi, Y., Golby, A. J., and O'Donnell, L. J. (2020). Deep white matter analysis (DeepWMA): Fast and consistent tractography segmentation. *Medical Image Analysis*, 65:101761.

Zhang, F., Daducci, A., He, Y., Schiavi, S., Seguin, C., Smith, R., Yeh, C.-H., Zhao, T., and O'Donnell, L. J. (2021). Quantitative mapping of the brain's structural connectivity using diffusion MRI tractography: A review. *arXiv:2104.11644 [q-bio]*.

Zhang, F., Wu, Y., Norton, I., Rigolo, L., Rathi, Y., Makris, N., and 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.

Zhuang, J., Hrabe, J., Kangarlu, A., Xu, D., Bansal, R., Branch, C. A., and Peterson, B. S. (2006). Correction of eddy-current distortions in diffusion tensor images using the known directions and strengths of diffusion gradients. *Journal of Magnetic Resonance Imaging*, 24(5):1188–1193.

Appendix A

Dataset specifications

A.1 Stem-ROI

Additional specifications for the manually curated stem-ROIs. See Figures A.1.1, A.1.2, A.1.3, A.1.4

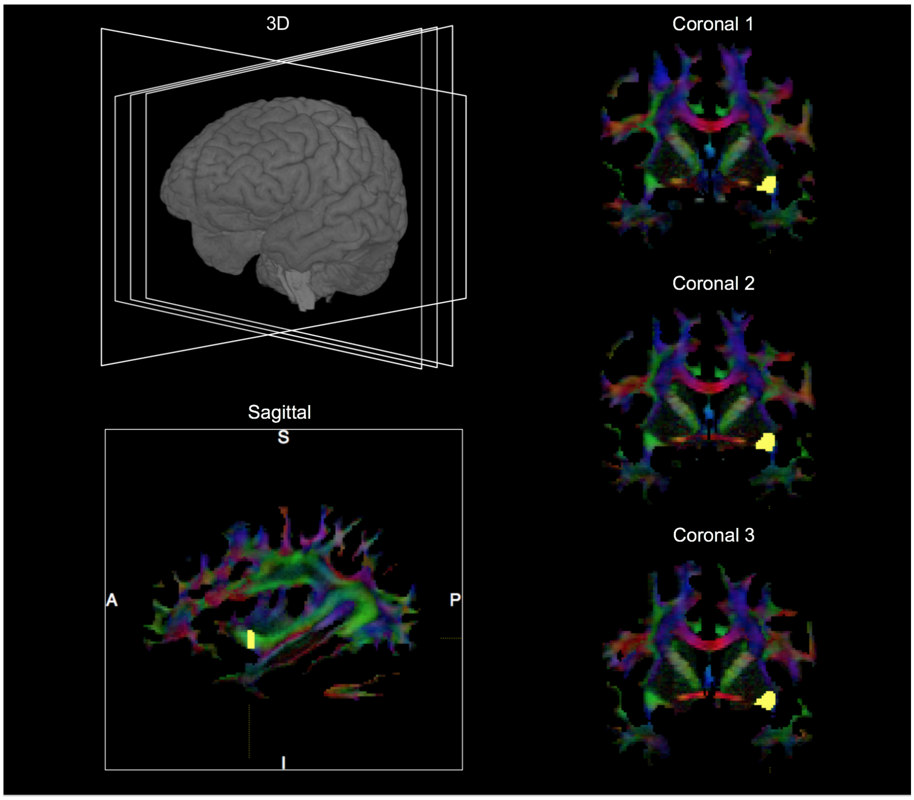


Figure A.1.1: Manual stem segmentation. The stem-ROI is segmented in three consecutive coronal slices. The sagittal view is used as reference for the antero-posterior location of the stem.

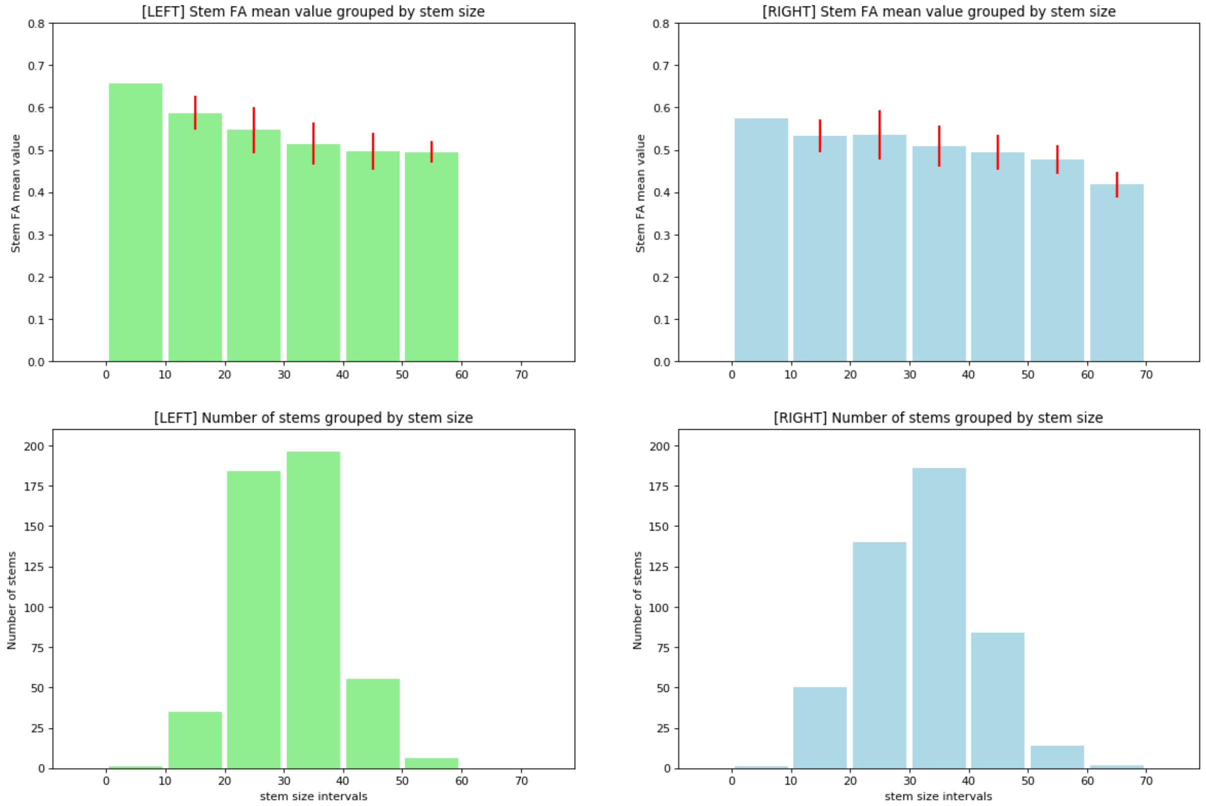


Figure A.1.2: Stem-ROI population statistics. The upper plots show the mean Fractional Anisotropy (FA) value of the populations stems grouped by stem size (number of voxels). The slightly decrease of mean FA value with the increase stem-ROI size is expected considering the anatomy of stem: super-dense of antero-posterior streamlines in its center, less on its borders. The bottom plots instead shows the stem-ROI size distribution across the population. We observe a normal distribution centered in the range [30-40]voxels.

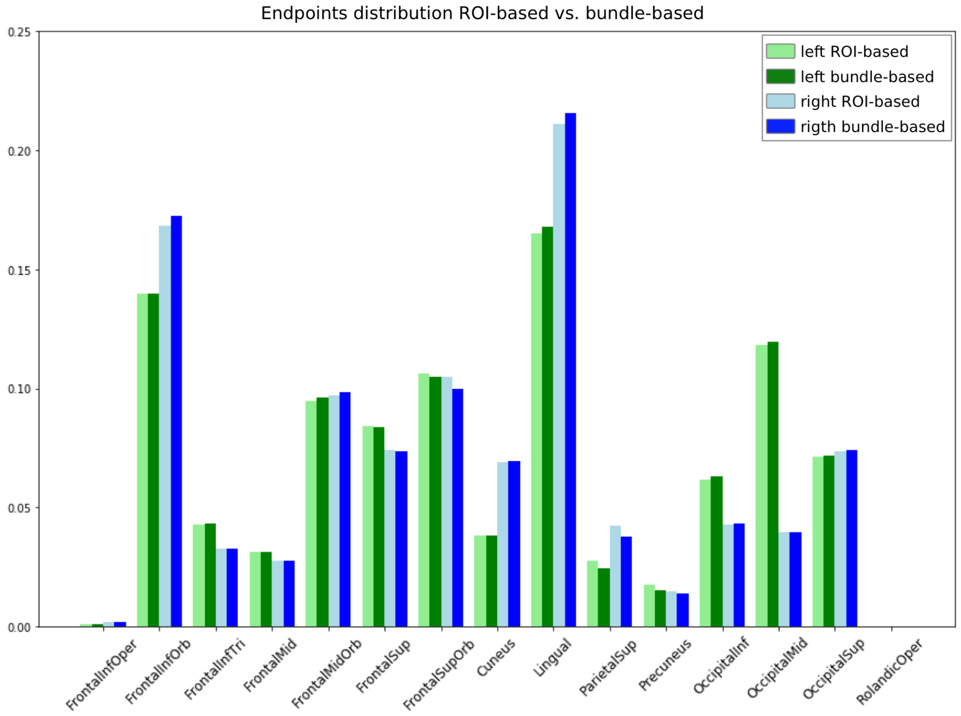


Figure A.1.3: Termination statistics of reference IFOFs. Considering the different strategies of labeling: ROI-based and bundle-based, we indicate with light colors IFOFs manually segmented with the former strategy, and dark colors for bundle-based IFOFs. We report all the termination regions defined according to (Hau et al., 2016), and derived from the Automated Anatomical Labeling 2 (AAL2) atlas (Rolls et al., 2015). In addition we report also the ”null“ statistics for the Rolandic Operculus, which is an exclusion ROI for the IFOF. Overall, we notice a good alignment of the endpoints distribution for the two strategies.

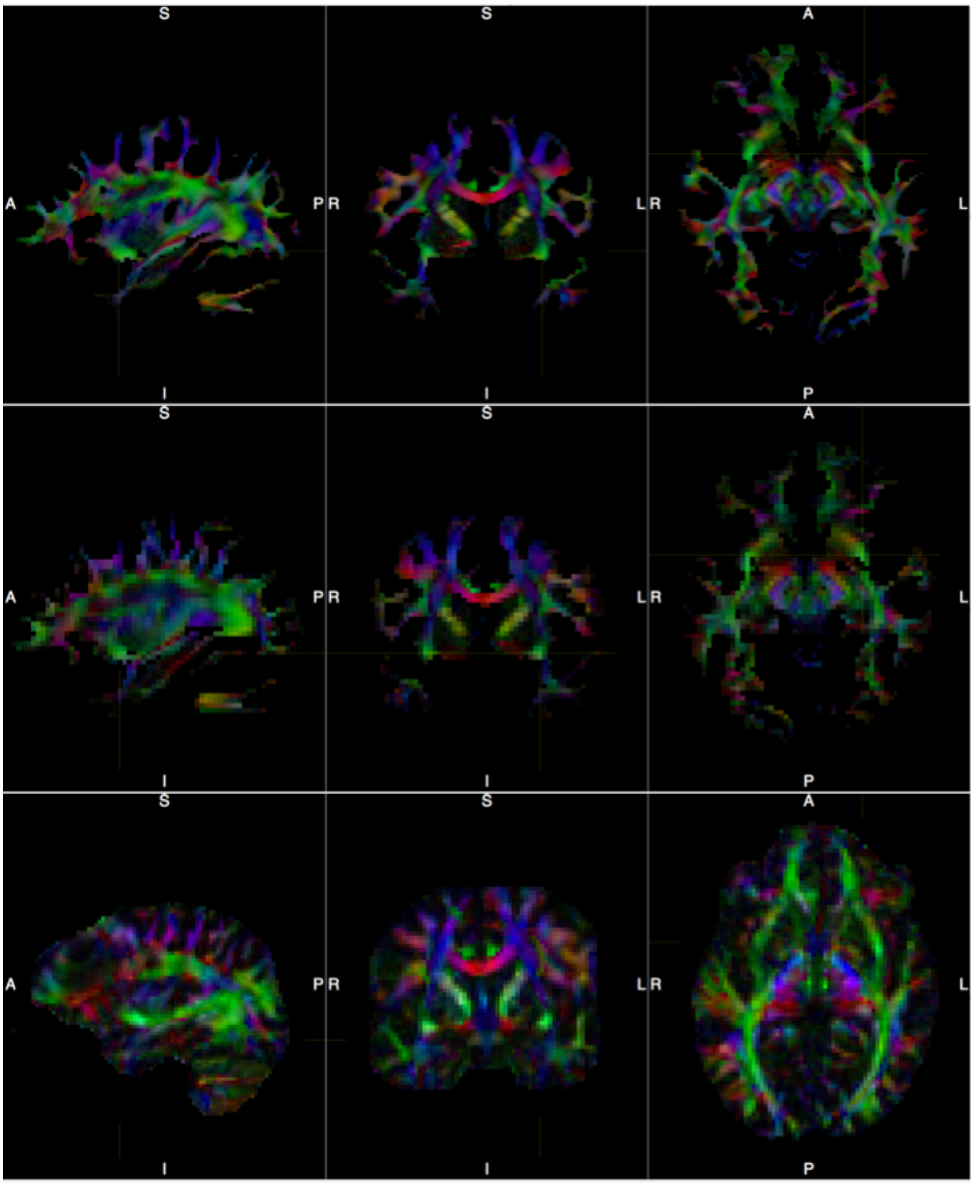


Figure A.1.4: Visual comparison of high quality colored FA (CFA) image versus clinical quality CFA. First row shows the HCP high quality CFA (3T scanner, resolution 1.25mm), the mid row shows the simulated low quality CFA i.e. reconstructed from HCP DWI at 2.5mm, clinical quality CFA (1.5T, resolution 2mm).

A.2 HCP-IW

Additional specifications for the dataset HCP-IW. See Table A.2.1

Table A.2.1: Full name list of (Wasserthal et al., 2018) bundles shared with (Zhang et al., 2018).

AF_left	(Arcuate fascicle)	SLF_I_left	(Superior longitudinal fascicle I)
AF_right		SLF_I_right	
ATR_left	(Anterior Thalamic Radiation)	SLF_II_left	(Superior longitudinal fascicle II)
ATR_right		SLF_II_right	
CC	(Corpus Callosum - all)	SLF_III_left	(Superior longitudinal fascicle III)
CG_left	(Cingulum left)	SLF_III_right	
CG_right		STR_left	(Superior Thalamic Radiation)
CST_left	(Corticospinal tract)	STR_right	
CST_right		UF_left	(Uncinate fascicle)
MLF_left	(Middle longitudinal fascicle)	UF_right	
MLF_right		T_PAR_left	(Thalamo-parietal)
ICP_left	(Inferior cerebellar peduncle)	T_PAR_right	
ICP_right		T_OCC_left	(Thalamo-occipital)
IFO_left	(Inferior occipito-frontal fascicle)	T_OCC_right	
IFO_right		ST_FO_left	(Striato-fronto-orbital)
ILF_left	(Inferior longitudinal fascicle)	ST_FO_right	
ILF_right		ST_PAR_left	(Striato-parietal)
OR_left	(Optic radiation)	ST_PAR_right	
OR_right		ST_OCC_left	(Striato-occipital)
MCP	(Middle cerebellar peduncle)	ST_OCC_right	

Appendix B

Additional investigations

B.1 2D CNN for stem-ROI segmentation

In this Section we present a early study on the segmentation of the stem-ROI of the IFOF, aimed at the subsequent extraction of the IFOF itself. The content we report here-below has been presented with a poster at the Organization for Human Brain Mapping (OHBM) 2019 conference.

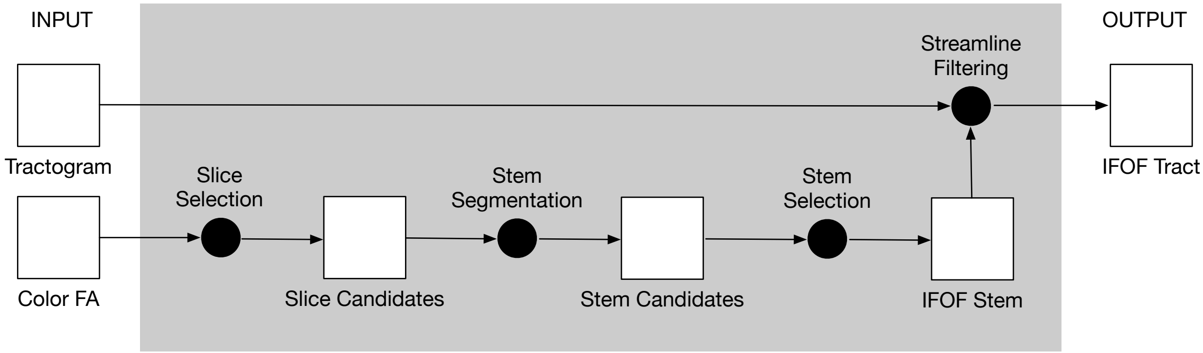


Figure B.1.1: IFOF segmentation with 2D stem-ROI segmentation.

Method The proposed method is composed by three main steps (see Figure B.1.1): (i) the computation of a set of candidates for stem-ROI in 2D coronal slices of the CFA, using a deep CNN; (ii) the selection of the slice containing the correct stem-ROI; (iii) the filtering of IFOF streamlines passing through the stem-ROI and the regions of termination as defined in (Hau et al., 2016). In details:

i From the CFA volume we select a range of 15 candidates coronal slices, and we split each of them to separate the two hemispheres (see input image in Figure B.1.2). The selection is based on the statistics of the input dataset: assuming an ACPC alignment of the input subjects the range of 15 slices is large enough to guarantee the presence of the stem-ROI also in unseen brains. The selected slices are fed into the trained CNN, which predicts a segmentation of the stem-ROI in each of the input half-slice (see Figure B.1.2).

ii The slice containing the (possibly) best candidate stem-ROI is selected using a heuristic based on the green intensity and the size. We reward higher mean green intensity in the corresponding CFA voxels, and an average size between the stem-ROI candidates.

iii The extracted stem-ROI is used as waypoint to filter the streamlines of IFOF, together with the termination points defined by the frontal (Frontal Inf. Oper., Frontal Inf. Orb., Frontal Inf. Tri., Frontal Mid., Frontal Mid. Orb., Frontal Sup., Frontal Sup. Orb.) and occipital (Cuneus, Lingual, Parietal Sup., Precuneus, Occipital Inf., Occipital Mid., Occipital Sup.) regions, determined after a non-linear registration (Avants et al., 2008) into subject space of the AAL atlas (Tzourio-Mazoyer et al., 2002).

CNN architecture As CNN we have re-implemented RefineNet (Lin et al., 2017) that is an end-to-end trainable model based on ResNet (He et al., 2016) and U-net (Ronneberger et al., 2015), depicted in Figure B.1.2. The encoder is a pre-trained ResNet (on ImageNet (Krizhevsky et al., 2012)), whose blocks are directly connected the *multi-path refinement* decoder, which is a composed of a series of RefineNet blocks. The direct connections allow the network to combine low-resolution and high-resolution features during the decoding phase, where usually many features are lost due to the deconvolution issues. In each RefineNet block the input features coming from the linked ResNet block and the previous RefineNet block are first fine-tuned with Residual Convolutional Unit (RCU), then fused with a Multi-resolution Fusion (MRF) that maps and upsample the inputs to a common feature space, and finally pooled with multiple pooling filters of different size combined together in a Chained Residual Pooling (CRP).

CNN loss The stem segmentation task presents an heavy unbalancing between the background and the ROI we want to segment. Typical crossentropy and log losses suffer such imbalance, thus we exploited a loss called Lovász-Softmax loss (Berman et al., 2018). This loss is a valid surrogate of the Intersection over Union (IoU), which is a metric that does not suffer the class unbalancing.

Experiments and results We used 150 subjects of the Human Connectome Project (Wu et al., 2016) dataset (90 gradients; b = 2000; voxel size=1.25mm). For each of them, we computed the CFA, and the tractogram using a deterministic local tracking algorithm (Garyfallidis et al., 2014). We manually segmented the stem-ROI, left and right, in three slices per subject. The subjects were split in a train set (120) and test set (30). The CNN was trained with batches of 12 single hemisphere coronal slices randomly picking left and right split for 18 epochs, using Adam optimizer with learning rate 2e-4 and weight

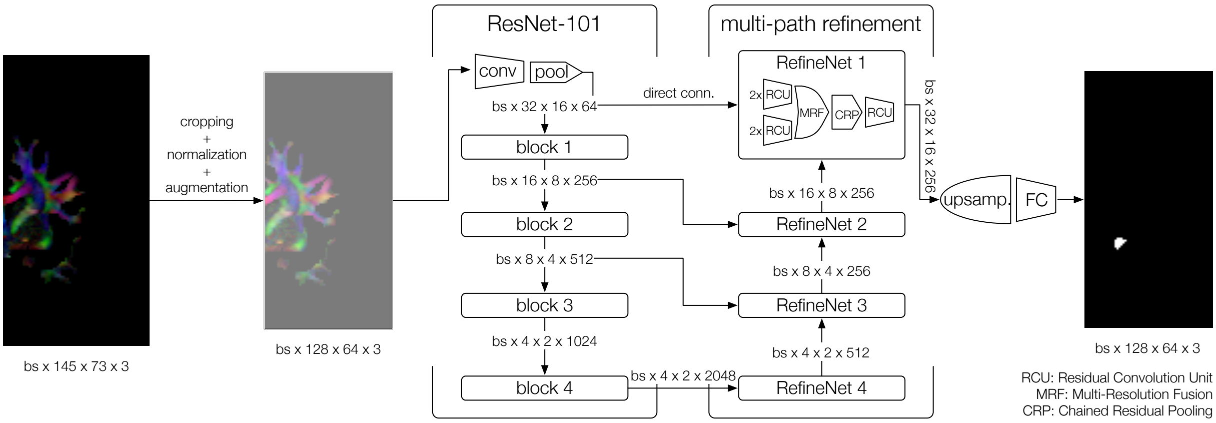


Figure B.1.2: RefineNet (Lin et al., 2017) architecture adopted.

decay 3e-4. Data augmentation was performed with horizontal flipping, rotation, zooming, and contrast distortion.

The trained CNN was evaluated by comparing the stem-ROI candidate, selected with the proposed heuristics, with the corresponding ground-truth manual segmentation in terms of IoU on voxels. To validate the optimality of the heuristic choice, we adopted as reference optimal selection the central of the three slices manually segmented, and we measured the slice offset between the selected and the reference.

The stem-ROI segmentation resulted is an average IoU value over the test subjects of 70.97%, while the optimality of the candidate selection is shown in Figure B.1.3.

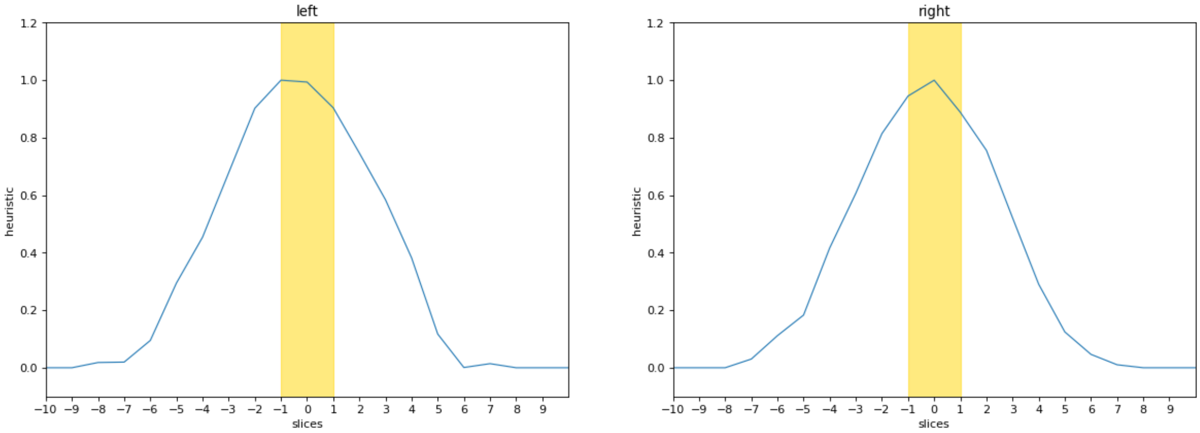


Figure B.1.3: Candidate stem-ROI selection results. The yellow column indicate the interval of three slices constituting the ground-truth stem-ROI. The blue curve is the heuristic value distribution for the ten candidate slices considered. It can be observed that the max value of the heuristic belong to the yellow column.

Finally, we evaluated the segmented IFOF with respect to two reference segmentations: a segmentation using a ROI-based approach, and a segmentation with a bundle-based approach, where all streamlines with terminations in frontal and occipital regions but not belonging to the IFOF were manually removed Porro-Muñoz et al. (2015). The evaluation was measured with the Dice Similarity Coefficient (DSC) between the corresponding voxel masks of tracts. We obtained a DSC value of 92.3% and 90.7% respectively. In Figure B.1.4 we show the variance inter-subjects, and we plot for each bundle of the test set the relation of DSC measure with respect to the corresponding reference bundle segmented with ROI-based and bundle-based approach.

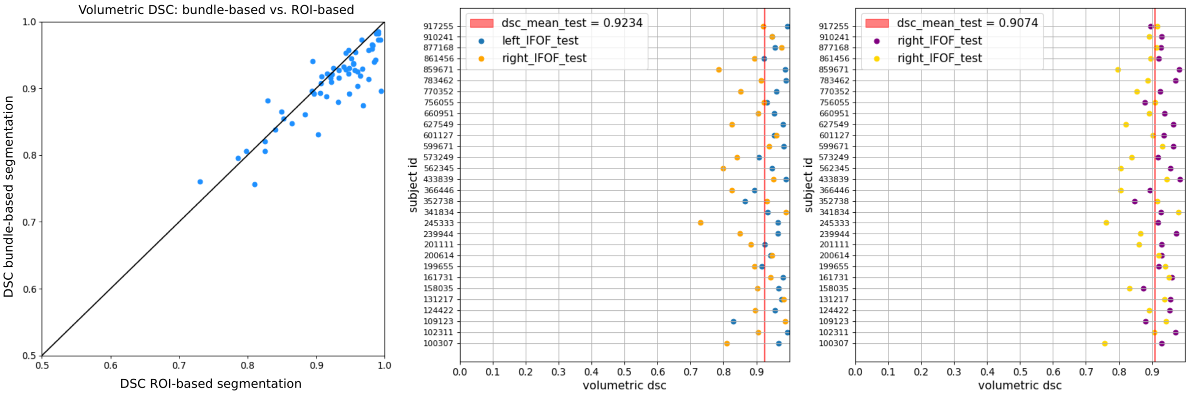


Figure B.1.4: Volumetric DSC of predicted IFOF. In the leftmost plot we highlight the (almost) linear relation between the score obtained with respect to roi-based (y-axis) and stem-based (x-axis) IFOF segmentations. In the central and rightmost plots we show the detailed results on ROI-based and bundle-based test set, respectively, differentiating subjects and the left and right IFOF.

B.2 Supplementary experiments for Verifyber

B.2.1 Alternative normalizations to MNI co-registration

Table B.2.1: Impact of different data normalizations on VF^{EP} performance. In this ablation we investigated if other non-registration based normalizations such as max-min and z-score could guarantee high accuracy results while not requiring non-linear registration to MNI at test time. However, it can be observed that MNI co-registration works best.

Normalization	Train space	Test space	Acc.	Prec.	Rec.	DSC
max-min	MNI	MNI	93.3 (±0.45)	95.3 (±0.43)	95.2 (±0.47)	95.2 (±0.34)
max-min	MNI	ACPC	83.2 (±4.54)	87.7 (±2.83)	88.7 (±2.65)	88.2 (±2.63)
z-score	MNI	MNI	92.7 (±0.63)	94.5 (±0.78)	95.1 (±0.43)	94.8 (±0.48)
z-score	MNI	ACPC	88.6 (±1.96)	91.6 (±1.39)	92.3 (±0.67)	92.0 (±1.00)
max-min wrt volume	MNI	MNI	95.2 (±0.13)	96.2 (±0.26)	96.9 (±0.19)	96.6 (±0.21)
max-min wrt volume	MNI	ACPC	81.6 (±2.80)	87.8 (±2.31)	85.7 (±2.46)	86.7 (2.37)

B.2.2 Verifyber test on BIL&GIN tractogram

BILGIN-EP BIL&GIN (Mazoyer et al., 2016) data source. Exclusive labeling (Petit et al., 2019). Single subject, with tractogram of 1.5M streamlines obtained via PF-ACT tractography (Girard et al., 2020)

Table B.2.2: Generalization test on BILIGIN-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 Veryfiber 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
VF	**92.5**	**96.2**	**93.7**	**94.9**

B.2.3 Exploring 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 B.2.1, B.2.3, B.2.2

Table B.2.3: 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 (±0.011)	0.766 (±0.039)
PointNet	DiPy	0.175 (±0.014)	0.757 (±0.033)

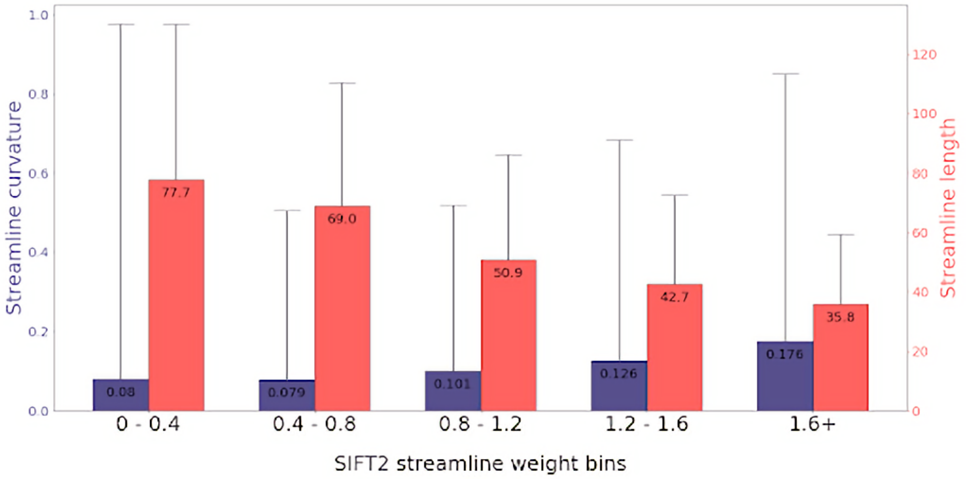


Figure B.2.1: SIFT2 weights per streamline categories in a single tractogram. Relationship between streamline anatomy and SIFT2 streamline weights. Each weight interval contains at least 10% of fibers. The trend of length and curvature shows that high-weight is assigned to very curved and short fibers. 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., 2019). Note also a very high standard deviation.

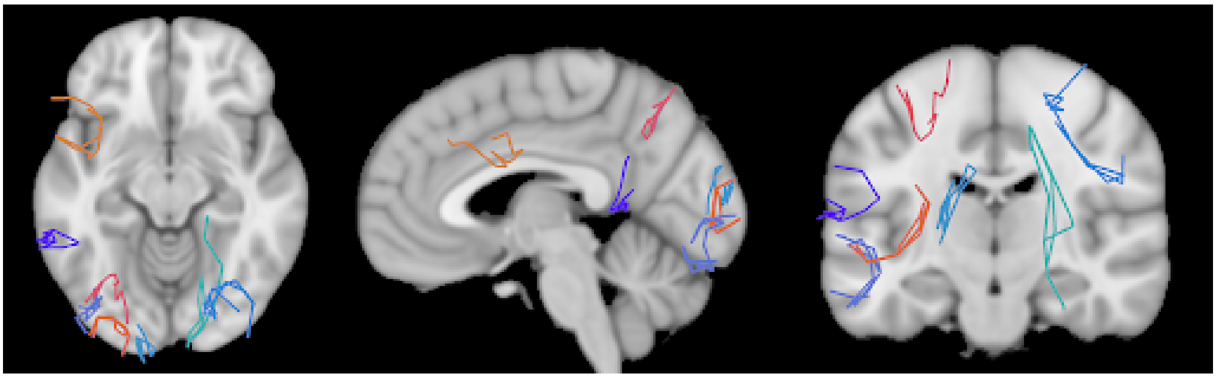


Figure B.2.2: Visual example of non-plausible fibers according to SIFT2. 8 streamlines with weights $>$ 3, lengths $>$ 92 mm, and curvatures $>$ 2 (MNI space, downsampled to 16 points). The figure shows that SIFT2 weights are not a good indicator for the anatomical plausibility of a streamline.

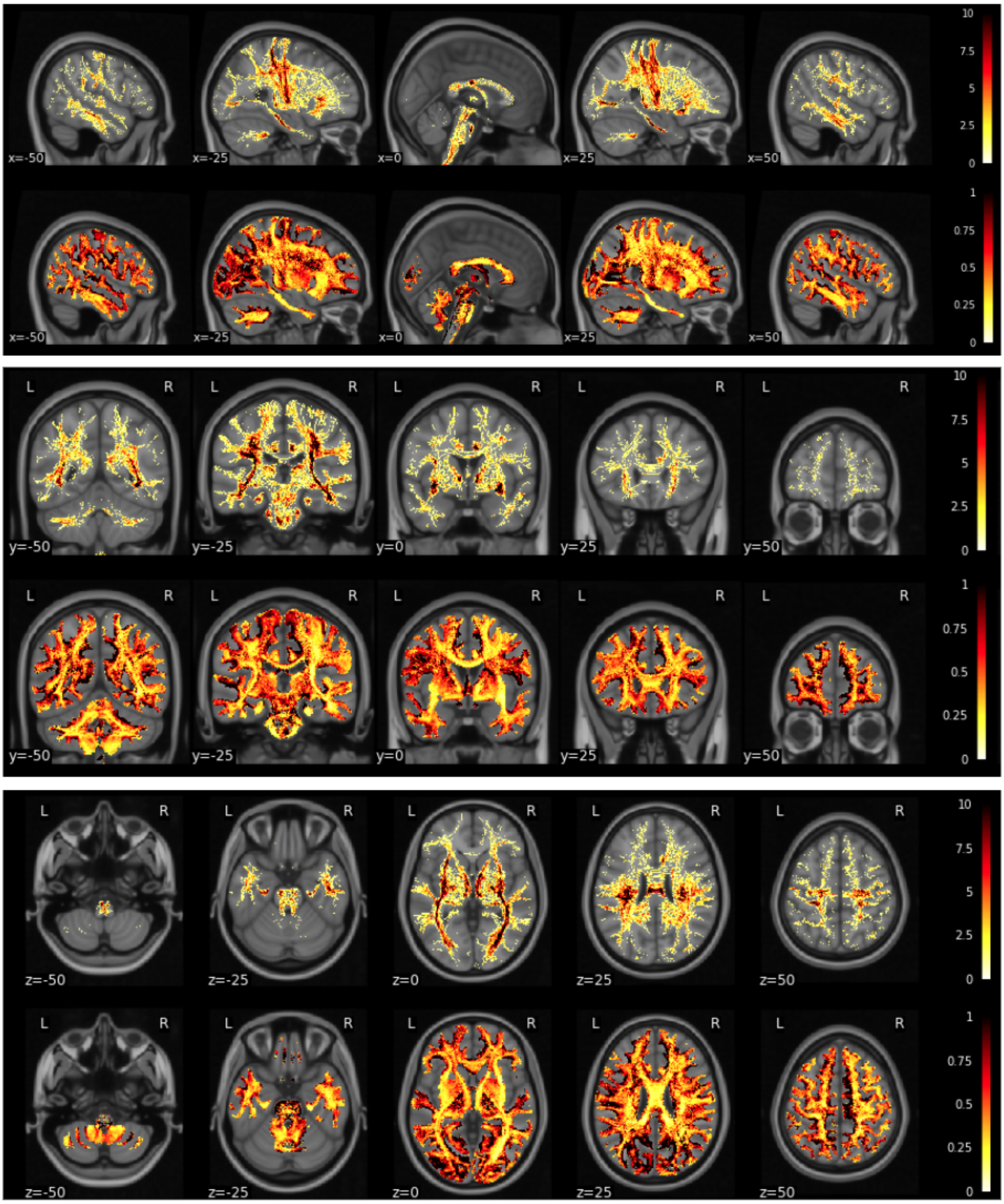


Figure B.2.3: Spatial distribution of SIFT2 weights vs. fiber density. SIFT weights distribution (top rows) compared to streamline density map (bottom rows). Observing the two types of maps, it is easy to see the high degree of correlation.

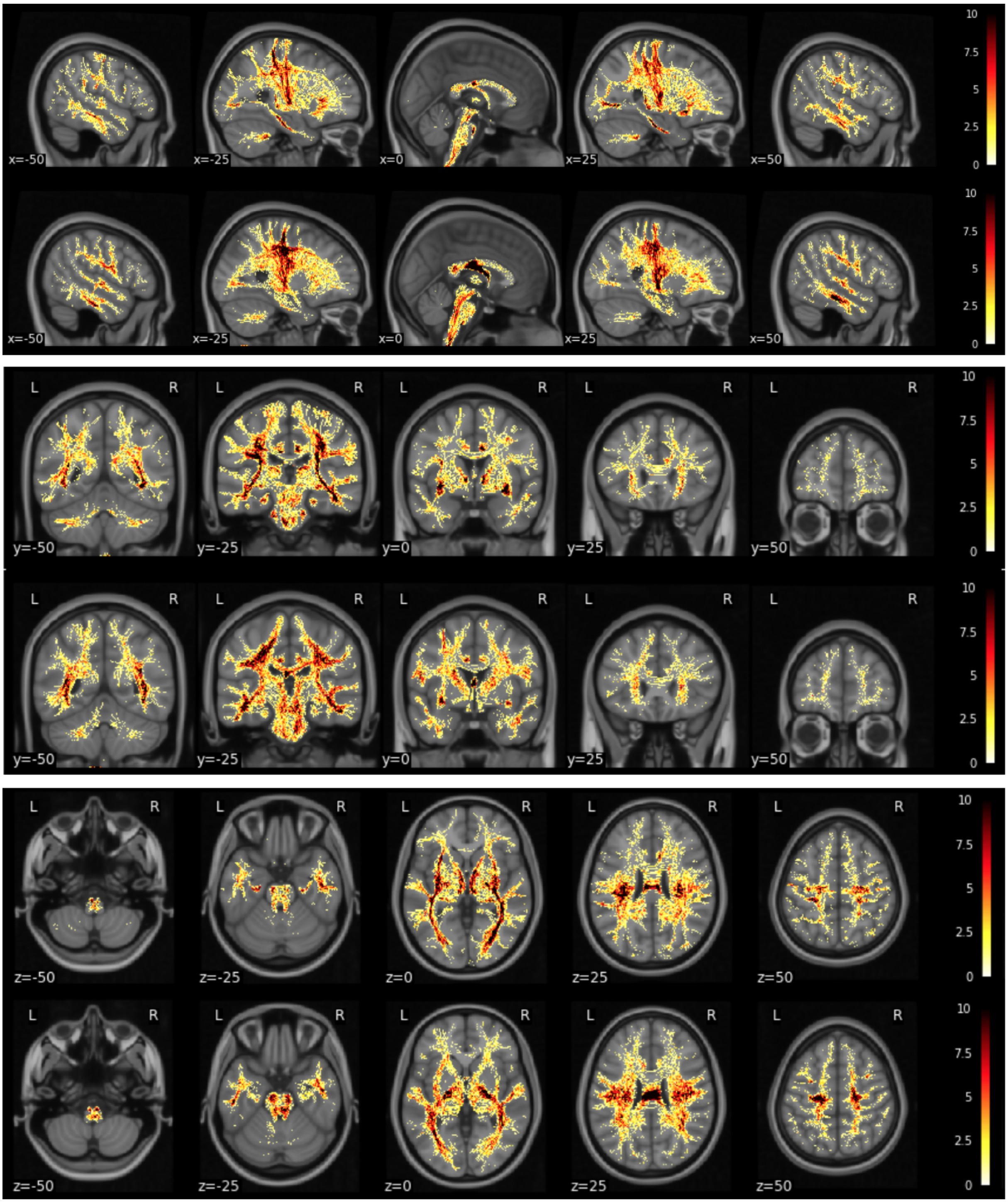


Figure B.2.4: Fiber density map obtained after Ensemble Tractography (Takemura et al., 2016) (top rows) and Particle Filtering tractography (Girard et al., 2014) (bottom rows). This figure shows that we cannot expect generalization of $\mathrm{VF}^{\mathrm{SIFT2}}$ across the two kind of tractograms.

B.3 Investigation of FINTA

Additional results obtained with the reproduced FINTA (Legarreta et al., 2021). See Table B.3.1 and Figures B.3.1, B.3.2, B.3.3.

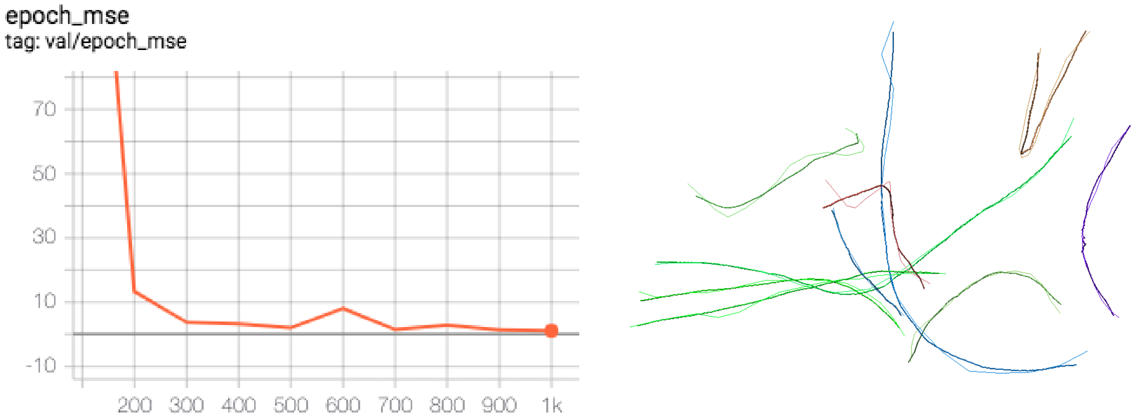


Figure B.3.1: Reproduced FINTA autoencoder. The plot on the left refer to a training performed on the dataset HCP-IZ. In this experiment, according to the setting used by the Authors of FINTA, we split the streamlines of the HCP-IZ tractogram into 80/20 train/test. The train set was then split in 4 buckets, 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 ∼1mm. 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.

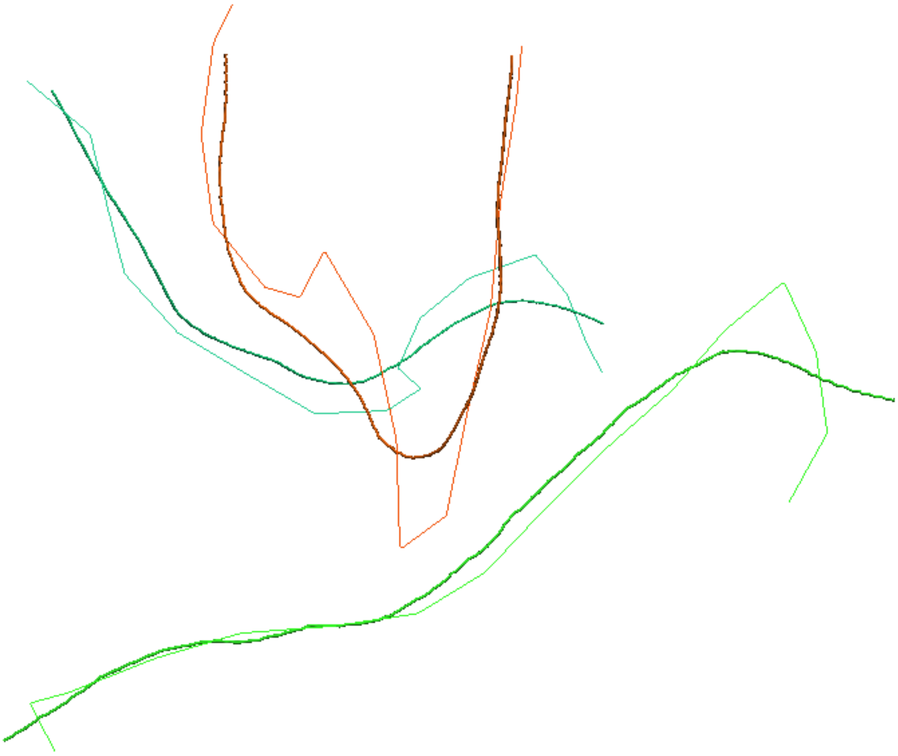


Figure B.3.2: Non-plausible fibers (HCP-EP) reconstructed using FINTA autoencoder. As can be noted original streamlines (thin) have a clear non-plausible trajectory with sharp curves. The denoising action performed by the autoencoder reconstruct smoother trajectories (fat), which we may observe to be 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.

Table B.3.1: FINTA performance evaluation. We report the score of accuracy, precision, recall, and DSC for each run. The training of the FINTA autoencoder is performed using data from the dataset indicated in the apex, e.g., FINTA$^{\text{IZ}}$ is trained on the dataset HCP-IZ. Specifically we use the same training/test set split used for our Verifyber experiments e.g., HCP-IZ is split into 80/20 training/test. From the training set a portion of p fibers is used as reference (Ref. p) for the embedded nearest neighbor step of FINTA. The radius threshold is then tuned on the same space using as query fibers a portion of the validation set of each dataset, except for BILGIN-EP v1 where we use the threshold computed on HCP-EP v1. We investigate the choice of the threshold based on two different metrics (reported in table as subscripts of the method): balanced accuracy (Bacc) and DSC. *: not tuned on this dataset.

Method	Ref. p	Thr.	Acc	Prec	Rec	DSC
			HCP-IZ all			
$\text{FINTA}^{\text{IZ}}_{\text{DSC}}$	HCP-IZ	36.9	88.0	87.3	95.8	91.3
$\text{FINTA}^{\text{IZ}}_{\text{BAcc}}$	HCP-IZ	32.9	87.6	89.9	91.5	90.7
Verifyber	-	-	97.1	97.6	98.0	97.8
			HCP-EP			
$\text{FINTA}^{\text{EP}}_{\text{DSC}}$	HCP-EP	39.9	73.6(1.2)	74.5(1.8)	94.8(0.4)	83.4(1.1)
$\text{FINTA}^{\text{IZ}}_{\text{DSC}}$	HCP-IZ	36.9*	59.9(0.8)	81.7(1.2)	55.0(1.7)	65.8(1.6)
$\text{FINTA}^{\text{IZ}}_{\text{DSC}}$	HCP-IZ	76.9	73.3(1.9)	74.0(2.4)	95.3(0.3)	83.3(1.4)
$\text{FINTA}^{\text{IZ}}_{\text{DSC}}$	HCP-EP	46.0	74.3(1.0)	75.1(1.6)	94.6(0.3)	83.8(1.0)
$\text{FINTA}^{\text{IZ}}_{\text{BAcc}}$	HCP-EP	32.9	71.2(0,7)	79.1(0.8)	80.2(1.0)	79.6(0.6)
Verifyber	-	-	95.2(0.1)	96.2(0.3)	96.9(0.1)	96.6(0.2)
			BILGIN-EP			
$\text{FINTA}^{\text{EP}}_{\text{DSC}}$	HCP-EP	39.9*	73.0	79.2	86.5	82.7
$\text{FINTA}^{\text{IZ}}_{\text{DSC}}$	HCP-IZ	36.9*	54.5	86.2	46.5	60.4
$\text{FINTA}^{\text{IZ}}_{\text{DSC}}$	HCP-IZ	76.9*	74.8	77.1	94.0	84.8
$\text{FINTA}^{\text{IZ}}_{\text{DSC}}$	HCP-EP	46.0*	74.0	80.2	86.6	83.3
$\text{FINTA}^{\text{IZ}}_{\text{BAcc}}$	HCP-EP	32.9*	61.9	84.0	60.5	70.3
Verifyber	-	*	92.5	96.2	93.7	94.9

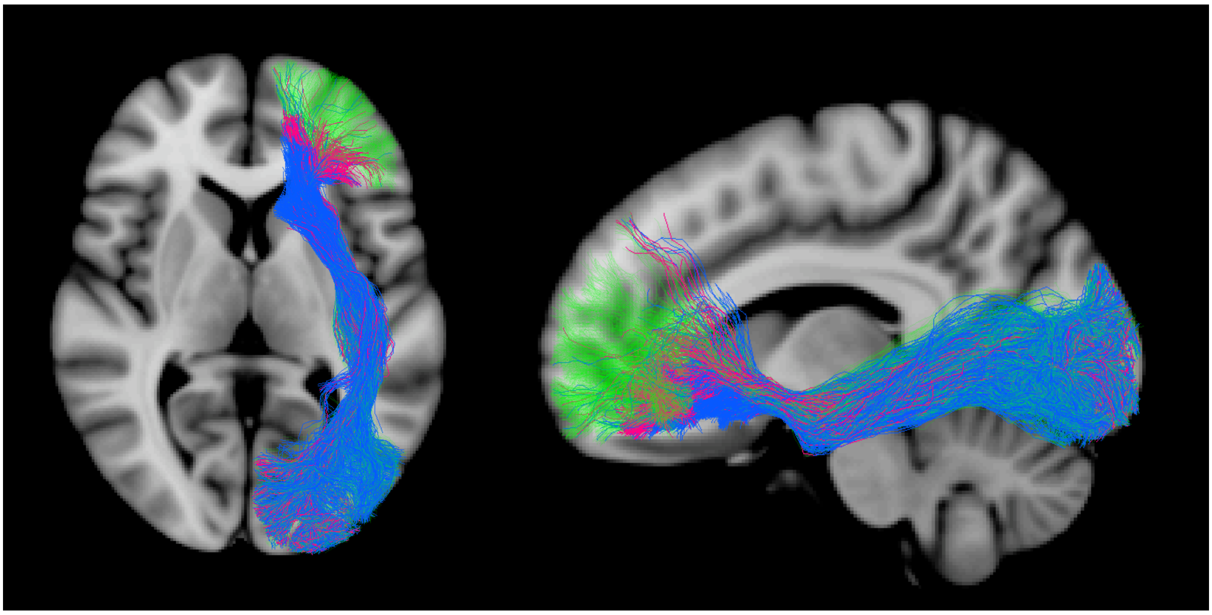


Figure B.3.3: IFOF left from HCP-IW Wasserthal et al. (2018) after classification of streamlines as plausible (p) and non-plausible (np) using FINTA trained on HCP-IZ Zhang et al. (2018). Considering the same classification performed with Verifyber (see Figure 16 of the paper) as a reference labeling, we investigate the possible false positive prediction of FINTA i.e,. fibers that are np according to Verifyber, 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.