Fiberprint: a subject fingerprint based on sparse code pooling for white matter fiber analysis

Kuldeep Kumar[a,*1], Christian Desrosiers[a], Kaleem Siddiqi[b],
Olivier Colliot[c,d,e], Matthew Toews[a]

[a] *Laboratory for Imagery, Vision and Artificial Intelligence, École de technologie supérieure, 1100 Notre-Dame W., Montreal, QC, Canada, H3C1K3*
[b] *School of Computer Science & Center for Intelligent Machines, McGill University, 3480 University Street, Montreal, QC, Canada, H3A2A7*
[c] *Sorbonne Universités, UPMC Univ Paris 06, Inserm, CNRS, Institut du cerveau et la moelle (ICM) - Hôpital Pitié-Salpêtrière, Boulevard de l'hôpital, F-75013, Paris, France*
[d] *Inria Paris, Aramis project-team, 75013, Paris, France*
[e] *AP-HP, Departments of Neurology and Neuroradiology, Hôpital Pitié-Salpêtrière, 75013, Paris, France*

Abstract

White matter characterization studies use the information provided by diffusion magnetic resonance imaging (dMRI) to draw cross-population inferences. However, the structure, function, and white matter geometry vary across individuals. Here, we propose a subject fingerprint, called *Fiberprint*, to quantify the individual uniqueness in white matter geometry using fiber trajectories. We learn a sparse coding representation for fiber trajectories by mapping them to a common space defined by a dictionary. A subject fingerprint is then generated by applying a pooling function for each bundle, thus providing a vector of bundle-wise features describing a particular subject's white matter geometry. These features encode unique properties of fiber trajectories, such as their density along prominent bundles. An analysis of data from 861 Human Connectome Project subjects reveals that a fingerprint based on approximately 3 000 fiber trajectories can uniquely identify exemplars from the same individual. We also use fingerprints for twin/sibling identification, our observations consistent with the twin data studies of white matter integrity. Our results demonstrate that the proposed Fiberprint can effectively capture the variability in white matter

[1] kkumar@livia.etsmtl.ca

fiber geometry across individuals, using a compact feature vector (dimension of 50), making this framework particularly attractive for handling large datasets.

Keywords: Subject fingerprint, dMRI, White matter geometry, fiber trajectories, Sparse Code Pooling, HCP, Twin data

1. Introduction

Diffusion magnetic resonance imaging (dMRI) is a powerful and non-invasive tool that provides key information on white matter organization and connectivity based on the diffusion of water molecules in white matter tissues [1]. Recent advances in dMRI acquisition protocols have lead to significant improvements in signal reconstruction [2, 3, 4], driving the development of novel tools for processing and interpreting dMRI data. Among the many applications using dMRI data, the quantitative characterization of white matter geometry and its genetic basis [5, 6, 7] is an important step in the study of the human brain, essential to understanding the mechanisms of neurological function and disease [8, 9, 10, 11].

Over the years, several approaches have been proposed to provide a simplified quantitative description of white matter connections, to allow for cross-population inferences [12, 13, 14, 15]. While numerous studies have focused on elucidating brain connectivity patterns that are shared across people, researchers have also acknowledged the high individual variability in brain structure [16, 17, 18], function [19, 20, 21, 22, 23, 24], and white matter geometry [25, 26]. Motivated by this, the concept of connectome fingerprinting, which characterizes individuals using unique connectivity profiles, has recently drawn significant interest from the neuroscience community [27, 28, 29, 30, 31, 32, 33].

So far, most studies on subject fingerprinting have centered around functional [27, 28, 29] and structural data [31, 34]. Recently, a novel approach was proposed for building individual connectome profiles based on dMRI data [33, 35]. This approach uses the Spin Distribution Function (SDF) at each voxel to obtain a fingerprint encoding the diffusion density along a set of prominent directions in cerebral white matter. While it captures key characteristics of white matter diffusivity, this voxel-level fingerprint lacks direct correspondence with white matter bundles, thus hindering an intuitive representation and analysis. As highlighted in [36], a direct voxelwise comparison of diffusion imaging data could also be challenging, since the high-contrast edges in diffusion MRI volumes (e.g., FA maps) make them more susceptible to small registration errors. Such comparison is also complicated by the anatomical variability of tract positions in subjects.

Building a fingerprint at the level of fiber trajectories, instead of voxels, could provide a more meaningful way of analyzing the unique connectivity properties of individuals from dMRI data. However, working with fiber trajectories also presents additional difficulties, due to the fact that the number and distribution of fiber trajectories may vary across subjects, and fiber trajectories may have very different lengths. Finding a common representation space of fiber trajectories, in different subjects, is essential to overcome these difficulties.

In recent work, we introduced a framework based on sparse coding for the compact representation and cross-population analysis of fiber trajectories [37]. This framework utilizes dictionary learning to build an atlas of fiber bundles from multi-subject dMRI data. Via sparse coding, this atlas can then be used to encode new fiber trajectory data into a compact representation, common to all subjects, and segment these fiber trajectories into prominent bundles [38]. In the current paper, we propose to use this framework to characterize the uniqueness in white matter connectivity exhibited by individual subjects, at the level of fiber trajectories. The key idea of our work is to represent each fiber trajectory as a sparse weighted combination of atlas bundles (i.e., the dictionary atoms), and use a pooling function [39] to combine the sparse codes of a subject's fiber trajectories into a single feature vector representing bundle-wise properties of fiber trajectory geometry. The resulting fingerprint, called *Fiberprint*, is used to uniquely identify subjects, as well as to discover inheritable characteristics of fiber geometry by comparing the fingerprints of twins and non-twin siblings. The use of fiber trajectories as a basis for the proposed subject fingerprint is supported by key studies, such as [25, 26], which have shown that the geometry of fiber bundles varies across subjects. However, characterizing an individual subject's white matter fiber geometry via a signature has thus far been elusive.

The main contribution of our work is the use of sparse code pooling to build a subject fingerprint, called *Fiberprint*. To our knowledge, this is the first study to propose a fingerprint based on fiber geometry. Another notable contribution of this work is the large-scale analysis and validation of our fingerprint, involving a cohort of 861 subjects from Human Connectome Project.

The rest of this paper is organized as follows. We first give an overview of related work on brain fiber analysis, sparse coding, and subject fingerprinting. Section 3 then presents the proposed Fiberprint approach, based on non-negative kernel sparse coding. In Section 4, we conduct an extensive experimental validation using the dMRI data of 861 subjects from the Human Connectome Project dataset, in which the impact of various parameters of our approach is measured. We also evaluate the usefulness of the proposed fingerprint on the task of subject, twin, and non-twin sibling identification, and use hypothesis testing to find bundles showing significant fingerprint dissimilarities across different subjects groups (i.e., males vs females). In Section 5, we discuss our main observations and experimental findings. We conclude with a summary of our contributions and a discussion of possible extensions.

2. Related work

Our presentation of relevant work is divided into three parts, focusing respectively on the representation and analysis of white matter fiber geometry, the application of sparse coding techniques in neuroimaging, and the topic of subject fingerprinting.

2.1. Representation and analysis of white matter fiber geometry

White matter fiber characterization often assumes an initial abstraction based on tractography, where local diffusion information is used to recover streamlines representing connectivity pathways in the brain [40, 41, 42]. Since tractography may output thousands of fiber trajectories, early work has focused on finding simplified quantitative descriptions of white matter connections by grouping fiber trajectories into anatomically meaningful bundles [43]. Over the years, a wide range of approaches have been proposed to cluster fiber trajectories, including methods based on hierarchical clustering [44, 45] and spectral clustering [46, 47, 48]. Most of these methods group fiber trajectories using problem-specific measures of similarity, such as the Hausdorff distance [44, 45, 49] or a mean of closest points (MCP) distance [44, 50, 45, 49].

Various studies have also focused on the segmentation of white matter tracts, toward the goal of drawing cross-population inferences [12, 13, 14, 15]. These studies either follow an atlas based approach [12, 13, 14] or align specific tracts directly across subjects [51, 52]. Multi-step or multi-level approaches have also been proposed to segment fiber trajectories, for example, by combining both voxel and fiber trajectory groupings [12], fusing labels from multiple hand-labeled atlases [13], using a white matter voxel-space atlas and a bundle representation based on maximum density paths [15], or using Gaussian processes [53]. A few studies have also investigated the representation of specific fiber trajectory bundles using different techniques such as gamma mixture models [54], hierarchical Dirichlet processes [55], and the computational model of rectifiable currents [56, 57].

2.2. Sparse coding for neuroimaging

Sparse coding, which aims at encoding a signal as a sparse combination of prototypes in a data-driven dictionary, has been applied in various domains of computer vision and pattern recognition [58, 59, 60, 39]. This technique has also shown promise for various neuroimaging applications, such as the reconstruction [61] or segmentation [62] of MRI data, and for functional connectivity analysis [63, 64]. For diffusion data, sparse coding has been used successfully for clustering white matter voxels from Orientation Density Function (ODF) data [65], and for finding a population-level dictionary of key white matter tracts [66].

To deal with the challenges of anatomic and tractographic variability, we have recently proposed a framework based on non-negative kernel dictionary learning for grouping fiber trajectories into prominent bundles [37]. This framework encodes individual fiber trajectories as a sparse non-negative combination of dictionary prototypes corresponding to bundles. Unlike other fiber trajectory clustering approaches, which assign fiber trajectories to individual bundles, the proposed framework gives fiber trajectories a membership value to each bundle, thus providing a more intuitive way of dealing with overlapping bundles and inter-subject variability. In a later study, the same framework was used to learn a multi-subject atlas of fiber bundles and for the automatic segmentation of new fiber trajectory data [38].

2.3. Subject fingerprinting

Most neuroimaging studies collapse multi-subject data to draw inferences about common patterns in a population. Although there are gross similarities, a substantial portion of a subject's connectome is unique to that individual [19, 25, 20, 17, 21, 18, 22]. A recent study has shown that functional connectivity profiles act as robust and reliable fingerprints that can identify individual subjects from a large group [28]. In this study, a functional brain atlas was employed to define target brain regions. The Pearson correlation coefficients between the time courses of region pairs were then computed, and used as a functional connectivity profile. This fingerprint was able to identify individuals across scan sessions, both for task and rest conditions.

In [31], Wachinger et al. proposed Brainprint, a subject fingerprint that characterizes brain morphology by calculating the spectrum of the Laplace-Beltrami operator on meshes from cortical and subcortical brain structures. This fingerprint was used to study morphological similarity between brains, with applications in subject identification across multiple scans of the same subject (achieving a classification accuracy of up to 99.9%), and the analysis of potential genetic influences on brain morphology.

While the majority of fingerprint studies have focused on functional and structural data, a local connectome fingerprint using Spin Distribution Function (SDF) voxel profiles obtained from dMRI data has recently been proposed in [32, 35]. This local fingerprint is built by sampling, at each voxel, the diffusion density of water along principal directions in the white matter, defined using a common fiber-direction atlas. The proposed fingerprint was used for quantifying the similarity between genetically-associated individuals, as well as measuring neuroplasticity over time, and was shown to vary substantially across individual subjects compared to traditional diffusivity measures like Fractional Anisotropy (FA). However, since this fingerprint is built using voxel-level information, it lacks a direct correspondence with white matter bundles, and a direct voxel-level comparison of diffusion imaging data could be challenging, as the high-contrast edges of diffusion MRI volumes (e.g., FA maps) make them more susceptible to small misregistration errors, as well as to anatomical variability of tract positions in health and disease [36]. Another point is that this fingerprint tries to capture both voxel-level diffusivity information and morphology. To our knowledge, the present study is the first to propose a white matter geometry fingerprint at the level of fiber trajectories and fiber bundles.

3. Materials and methods

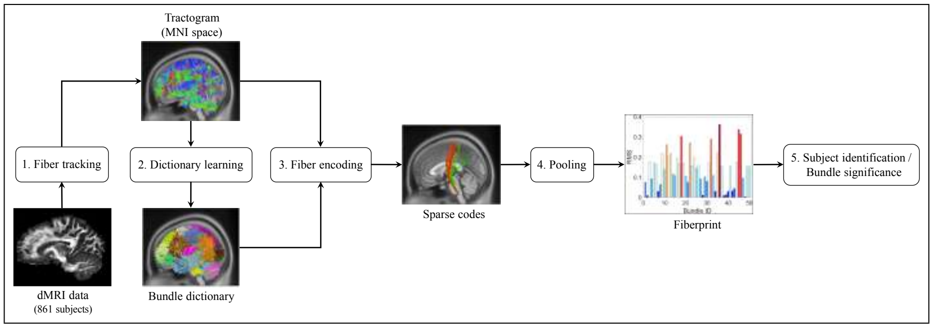


Figure 1: Pipeline of the proposed *Fiberprint* approach based on sparse code pooling.

Figure 1 summarizes the pipeline of the proposed Fiberprint method, comprised of three steps. In the first step, signal reconstruction and fiber tracking is performed on the pre-processed dMRI data of 861 subjects from the Human Connectome Project [67, 68]. Second, a dictionary of prototype fiber trajectories is then learned from a subset of subjects, based on our non-negative kernel dictionary learning framework. This dictionary can be seen as an atlas for modeling and analyzing the geometry of fiber trajectories from multiple subjects, along prominent bundles. In the third step, the learned dictionary is used to encode the fiber trajectories of the remaining subjects in a common feature space, via a sparse coding method. A fingerprint is then obtained, for each subject, by applying a pooling function to the sparse codes corresponding to each subject's fiber trajectories. This pooling function allows the comparison of subjects having a different number of fiber trajectories by aggregating the information from all fiber trajectories in a single fixed-size vector. The resulting fingerprint corresponds to an estimate of fiber trajectory density along key bundles defined by the atlas. Finally, in the last step, fingerprints are used to identify unique characteristics of genetically-related subjects, or for finding bundles showing significant differences across various subject groups (e.g., male vs female). The following subsections describe each of these steps in greater detail.

3.1. Data and pre-processing

We used the pre-processed dMRI data of 861 subjects (482 females, 378 male and 1 unknown, age 22–35) from the Q3 release of the Human Connectome Project [69, 67, 68], henceforth referred to as HCP data. All HCP data measure diffusivity along 270 directions distributed equally over 3 shells with b-values of 1000, 2000 and 3000 $^{\text{s}}/_{\text{mm}^2}$, and were acquired on a Siemens Skyra 3T scanner with the following parameters: sequence = Spin-echo EPI; repetition time (TR) = 5520 ms; echo time (TE) = 89.5 ms; resolution = $1.25 \times 1.25 \times 1.25$ mm^3 voxels. Further details can be obtained from HCP Q3 data release manual[2].

For signal reconstruction and tractography, we used the freely available DSI Studio toolbox. All subjects were reconstructed in MNI space using the Q-space

[2] `http://www.humanconnectome.org/documentation/Q3/`

diffeomorphic reconstruction (QSDR) [70] option in DSI Studio. QSDR is an extension of generalized q-sampling imaging (GQI, [71]), allowing the construction of spin distribution functions (SDF) in a given template space. DSI Studio first calculates the quantitative anisotropy (QA) mapping in the native space and then normalizes it to the MNI QA map using SPM normalization [72]. We used the SPM 21-27-21 option in DSI Studio for normalization, and set output resolution to 1 mm. For skull stripping, we used the masks provided with preprocessed diffusion HCP data. Other parameters were set to the default DSI Studio values. We also normalized T1-weighted images to MNI template space as part of this processing.

Deterministic tractography was performed with the Runge-Kutta method of DSI Studio [40, 73], using the following parameters: minimum length of 40 mm, turning angle criteria of 60 degrees, and trlinear interpolation. The termination criteria was based on the QA value, which is determined automatically by DSI Studio. As in the reconstruction step, the other parameters were set to the default DSI Studio values. Using this technique, we obtained a total of 50 000 fiber trajectories for each subject.

As a note, whether these fiber trajectories represent the actual white matter pathways remains a topic of debate [74, 75]. Fiber trajectories derived from DSI studio are hypothetical curves in space that represent, at best, the major axonal directions suggested by the orientation distribution functions of each voxel, which may contain tens of thousands of actual axonal fibers.

3.2. Learning the fiber trajectory dictionary

Out of the 861 available subjects, 10 unrelated ones [76] were used to learn the dictionary of fiber trajectory prototypes, serving as a multi-subject atlas to map new fiber trajectory data to a common space. The learning process is based on the non-negative kernel dictionary learning method presented in [37, 38], which we now summarize.

Let X be the set of n training fiber trajectories, represented as a set of 3D coordinates. For the purpose of explanation, we suppose that each trajectory i

is encoded as a feature vector $x_i \in \mathbb{R}^d$, and that X is a $d \times n$ feature matrix. Since our dictionary learning method is based on kernels, a fixed set of features is however not required, and fiber trajectories having a different number of 3D coordinates could be compared with a suitable similarity measure (i.e., the kernel function).

In the proposed model, each fiber trajectory can be described as a sparse linear combination of m prototype fiber trajectories in a dictionary D. Formally, we write this as $x_i \sim Dw_i$, where w_i is a sparse vector of non-negative weights representing the fiber trajectory's relationship to each prototype. Since fiber trajectories may have very different lengths and endpoints, encoding them using a fixed set of features can be challenging. To avoid this problem, we embed them into a q-dimensional Hilbert space via a mapping function $\phi : \mathbb{R}^d \to \mathbb{R}^q$, such that $\phi(x)^\top \phi(x') = k(x, x')$ is a kernel function. The main advantage of this approach is that fiber trajectories can now be represented based on a similarity measure tailored to this type of data, such as the Hausdorff distance [44, 45, 49], the mean of closest points (MCP) distance [44, 50, 45, 49] or the Minimum average Direct Flip (MDF) distance [77]. In this work, we considered the MDF distance, which computes the average distance between points on a fiber trajectory and corresponding points in a second fiber trajectory, or in the reverse point sequence of the second fiber trajectory if it leads to a smaller distance. A Gaussian kernel was used to convert distances to similarities, i.e. $k(x, x') = \exp\big(-\gamma \cdot \mathrm{dist}_{\mathrm{MDF}}(x, x')\big)$. The fiber trajectories were sampled to 15 equidistant points for distance computation [77] and the kernel bandwidth parameter was set empirically to $\gamma = 0.0001$.

Using $\Phi \in \mathbb{R}^{q \times n}$ to denote the matrix of mapped training fiber trajectories, the kernel matrix of pairwise similarities then corresponds to $K = \Phi^\top \Phi$. Using the idea proposed in [78], we express the dictionary as a non-negative linear combination of training examples, i.e., $D \sim \Phi A$, and formulate the dictionary learning task as the following optimization problem:

$$\underset{A,\, W \geq 0}{\arg\min}\ \frac{1}{2} \|\Phi - \Phi A W\|_F^2 \quad \text{s.t.} \quad \|w_i\|_0 \leq S_{\max}, \quad i = 1, \ldots, n, \tag{1}$$

where $\|w_i\|_0$ is the L_0 norm (i.e., number of non-zero elements) of w_i, constraining each fiber trajectory to be encoded using at most $S_{\max}$ prototypes, $A \in \mathbb{R}^{n \times m}$ is the dictionary coefficient matrix, and $W \in \mathbb{R}^{m \times n}$ is the sparse code matrix for all fiber trajectories. When $S_{\max} = 1$, this formulation corresponds to the kernel K-means problem [79]. As shown in Section 4.1.4, expressing fiber trajectories using more than one prototype (i.e., $S_{\max} > 1$) provides a better representation of complex bundles, leading to a more discriminative fingerprint.

This problem is solved using the method described in [37], which updates the sparse codes W and dictionary matrix A iteratively, until convergence. In the sparse coding step, each column of W is updated independently by optimizing the following sub-problem:

$$\underset{w_i \geq 0}{\arg\min} \ \frac{1}{2} w_i^\top A^\top K A w_i \, - \, k_i^\top A w_i \quad \text{s.t.} \ \ \|w_i\|_0 \leq S_{\max}, \tag{2}$$

where $k_i \in \mathbb{R}^n$ is the vector containing the similarities between fiber trajectory i and all training fiber trajectories. This problem is solved heuristically using a non-negative kernel Orthogonal Matching Pursuit (NKOMP) algorithm [37]. The dictionary matrix A is then obtained using a kernel version of the non-negative matrix tri-factorization approach proposed in [80], which applies the following update scheme until convergence:

$$A_{ij} \ \leftarrow \ A_{ij} \cdot \frac{\left(KW^\top\right)_{ij}}{\left(KAWW^\top\right)_{ij}}, \quad \forall\, i, j. \tag{3}$$

Due to machine precision, the above update scheme produces small positive values instead of zero entries. To resolve this problem, a small threshold is applied on A.

Since the kernel contains the similarities between each pair of fiber trajectories ($50\,000 \times 10$ fiber trajectories, squared), computing it directly is impracticable. Instead, we start with $5\,000$ fiber trajectories sampled uniformly from each subject, and approximate the resulting kernel matrix ($50\,000 \times 50\,000$) using Nystrom's method [81, 14]. This method starts with defining a subset of fiber trajectories and computing the pairwise similarities between each training fiber trajectory and this sampled subset. The missing entries in kernel matrix K are then estimated using a low-rank approximation process based on SVD. Using this technique, the entire dictionary learning process takes about 1 000 seconds on a quad-core 3.6 GHz computer with 32 GB of RAM.

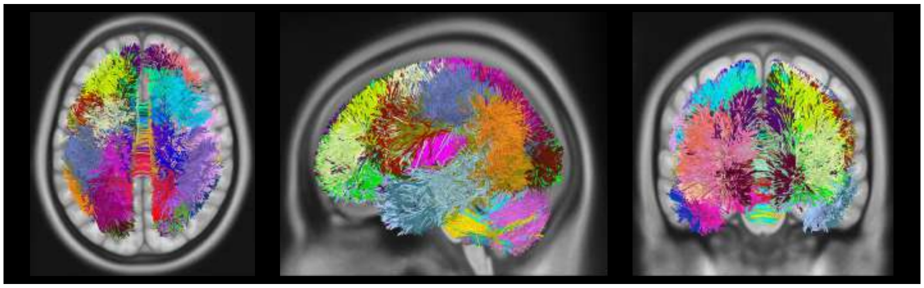


Figure 2: Dictionary visualization. Visualization of $m = 50$ fiber trajectory prototypes learned from 10 subjects, with an unique color assigned to each dictionary prototype. For this simplified visualization each fiber trajectory is assigned to a single prototype by taking the maximum for each row of the matrix A. (superior axial, left sagittal, and anterior coronal views respectively)

Figure 2 gives a qualitative visualization of $m = 50$ fiber trajectory prototypes learned in the dictionary (the impact of parameter m is analyzed in Section 4.1.2), each one corresponding to a different color. To generate this figure, we convert the soft assignment defined in A to a hard clustering, by assigning each fiber trajectory i to the prototype j for which a_{ij} is maximum[3]. We see that the fiber trajectory clusters defined by the dictionary are reasonably consistent with prominent neuroanatomical bundles, such as the corpus callosum, cingulum, corticospinal tract and superior cerebellar penduncle. Note, however, that a one-to-one relationship does not always hold between these prototypes and neuroanatomical bundles: complex bundles may be represented using multiple prototypes. Nonetheless, to simplify the presentation, we use the term bundle dictionary when referring to the output of the dictionary learning step.

[3] A separate visualization of each fiber trajectory cluster can be found in the supplementary material.

3.3. Generating the subject fingerprints

The generation of a fingerprint from the fiber trajectory data of a new subject is composed of two steps: sparse coding of fiber trajectories and sparse code pooling.

Sparse coding of fiber trajectories

In the first step, the learned dictionary is used to map the fiber trajectories of a given subject to a common feature space defined by the dictionary's bundles. This encoding process consists of solving the sparse coding problem of Eq. (2), which has been used for dictionary learning. Since each fiber trajectory is represented using at most $S_{\max}$ coefficients, this re-encoding of a subject's fiber trajectory data is very compact.

The fiber trajectory sparse codes of four different subjects, obtained using the dictionary of Figure 2, are illustrated in Figure 3. We represent bundles using the same colors as in Figure 2, and assign each fiber trajectory i to the bundle for which w_{ji} is maximum, where W is the sparse code matrix of a given subject. This hard assignment of fiber trajectories to dictionary bundles corresponds to the fiber trajectory segmentation approach presented in [38]. The strength of the relationship between fiber trajectories and individual bundles can also be visualized by considering the values in each row of W. In Figure 4, the sparse code values (i.e., rows of W) corresponding to the left and right corticospinal bundles are color coded such that fiber trajectories having a high membership to a bundle are red and those having a low membership are green (fiber trajectories with zero membership are not shown). These figures highlight the implicit correspondence of bundles across subjects, as well as the variability in the fiber trajectory geometry of bundles, observed for different subjects.

Sparse code pooling

Because subjects may have a different number of fiber trajectories, to allow comparison across subjects, the sparse codes for fiber trajectories obtained in the previous step must be aggregated in a fixed-size feature vector. This is

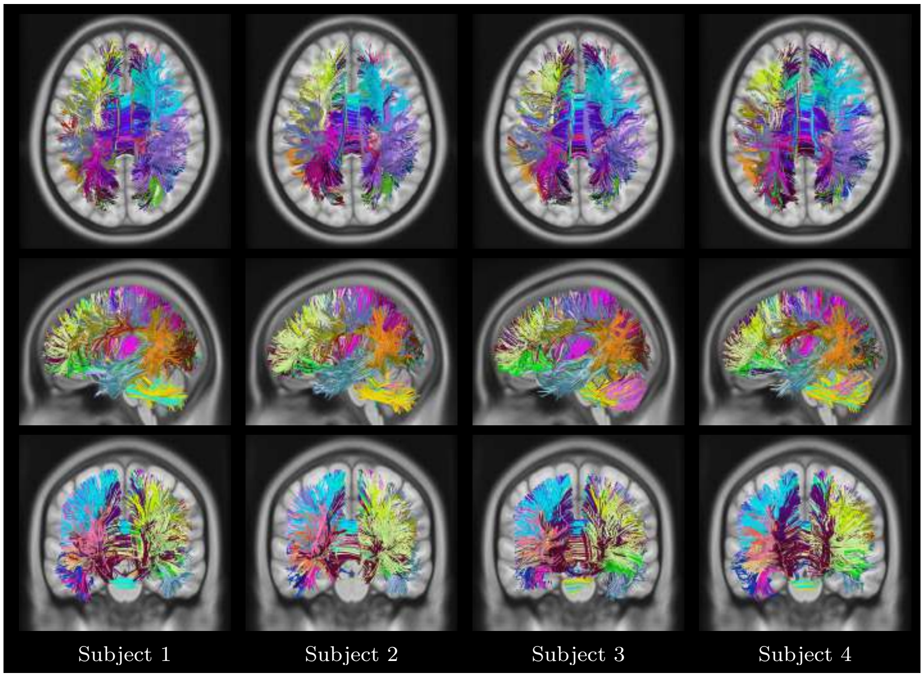


Figure 3: Visualization of sparse code representation of fiber trajectories from four subjects. Each fiber trajectory is assigned to a single bundle by taking the maximum of the sparse code vector. Bundles are represented using the same colors as in Figure 2. (superior axial (top), left sagittal (middle), and anterior coronal (bottom) views respectively)

achieved using a sparse code pooling function [39] that combines, for each dictionary bundle, the relationship between this bundle and all fiber trajectories of 310 a subject into a single value. Let $W \in \mathbb{R}^{m \times n}$ be the sparse code matrix obtained in the previous step, each column corresponding to a different fiber trajectory of the subject to encode. We consider three pooling functions frequently used in the literature, based on the root mean square (RMS), mean and maximum:

$$\big[f_{\mathrm{RMS}}(W)\big]_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n} w_{ji}^2} \tag{4}$$

$$\big[f_{\mathrm{Mean}}(W)\big]_j = \frac{1}{n}\sum_{i=1}^{n} |w_{ji}| \tag{5}$$

$$\big[f_{\mathrm{Max}}(W)\big]_j = \max\big\{|w_{j1}|, |w_{j2}|, \ldots, |w_{jn}|\big\}. \tag{6}$$

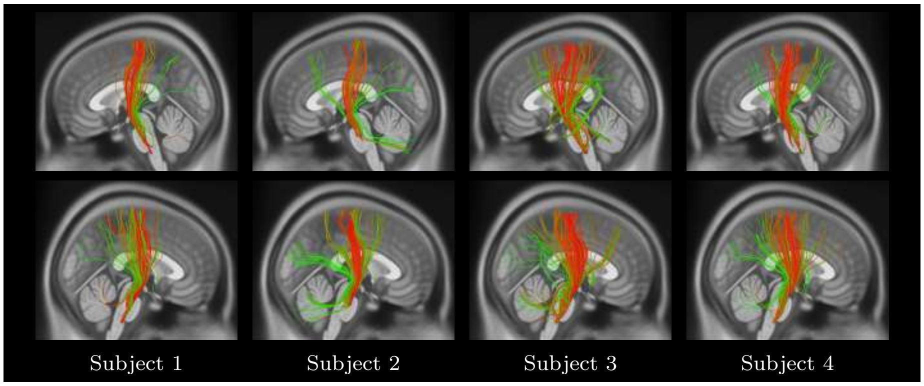
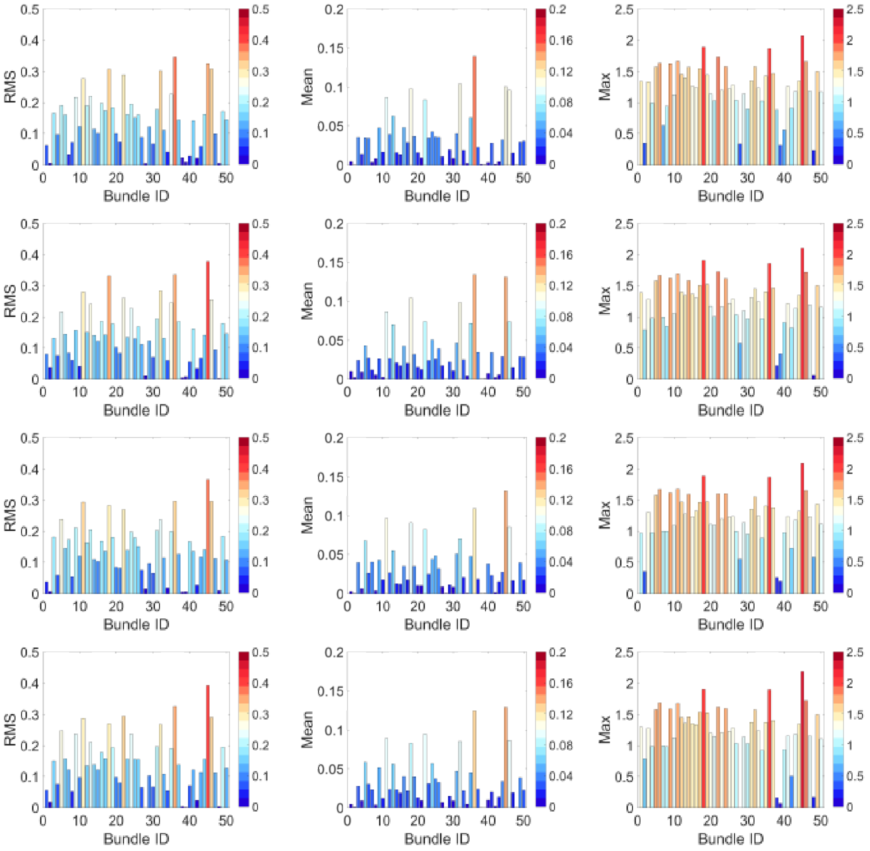


Figure 4: Color coded visualization of sparse code memberships of fiber trajectories for the left (top row) and right (bottom row) corticospinal bundles from four subjects. Green and red represent, a low and a high membership of a fiber trajectory to a bundle, respectively. Fiber trajectories with a zero membership to the bundle are removed for a simplified visualization.

where $[f(W)]_j$ is the pooled feature corresponding to the j-th dictionary bundle.

Each of these pooling functions encodes a different property of a subject's fiber trajectory distribution along the dictionary bundles. Function f_{mean} computes the average sparse code value of fiber trajectories belonging to a bundle, thus giving an estimate of the bundle's density. f_{RMS} is another measure of density, which gives a greater importance to large magnitude values in W. Finally, $f_{\max}$ selects the maximum sparse code value over all fiber trajectories in relationship to a given bundle. In practice, this value will be low for dictionary prototypes which are not useful for encoding a subject's fiber trajectories.

Figure 5 shows a bar plot representation of fingerprints obtained using the three pooling functions, for four different subjects. We observe small but meaningful differences when comparing these fingerprints, supporting the hypothesis that they encode unique characteristics of fiber trajectory geometry. Moreover, we see that the pooling functions capture different properties (in particular the max pooling function) and have varying responses across bundles. The uniqueness of subject fingerprints can be further visualized in Figure 6, which color codes the fiber trajectory bundles of the four subjects based on the magnitude of their corresponding RMS pooling function values. We observe that the bundles showing the highest response are consistent across subjects, although the magnitude of these responses differs from one subject to another.



RMS Mean Max

Figure 5: Subject fingerprint visualization. Color coded bar plot representation for four subjects (rows) and three pooling functions (RMS, Mean, and Max; columns), plotted as a value per bundle ID.

4. Experiments and results

In this section, we test the hypothesis that the proposed subject fingerprint can effectively capture a particular subject's white matter fiber geometry.

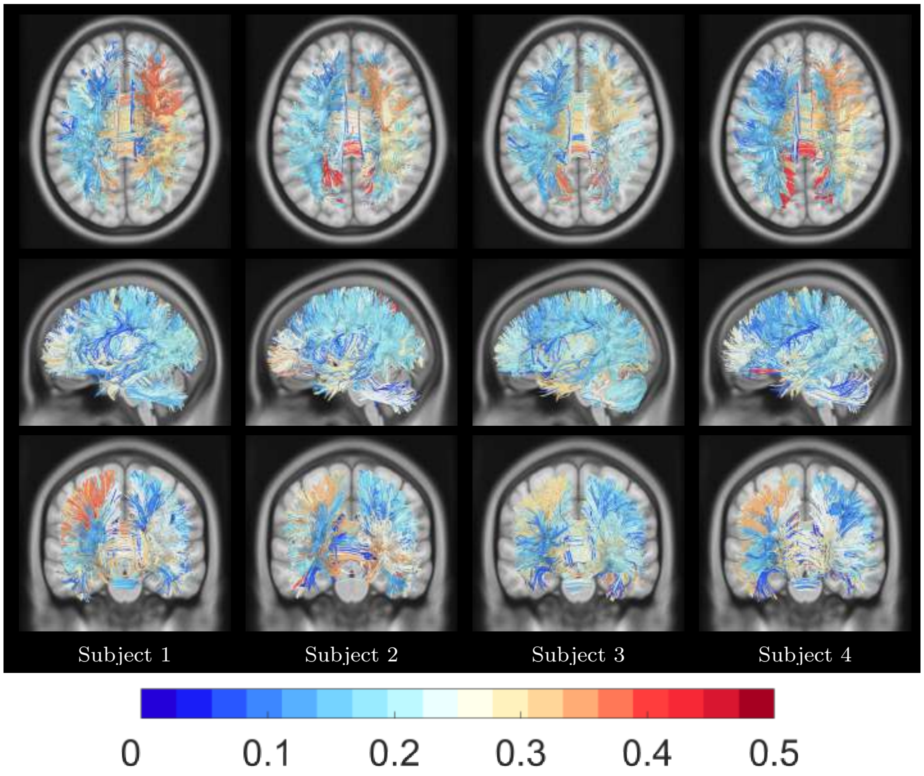


Figure 6: Subject fingerprint visualization. Color coded bundles from four subjects representing the magnitude of their corresponding RMS pooling function values. We use the same color code scheme as in Figure 5. (superior axial (top), left sagittal (middle), and anterior coronal (bottom) views respectively)

Because there are many parameters and factors involved in the generation of fingerprints (e.g., pooling function, dictionary size, and fiber tracking approach), we first perform an analysis to assess the robustness of our fingerprint to these various parameters and factors. We then validate our main hypothesis using the task of subject identification and twin identification. Specifically, we try to determine if an individual can be identified using the proposed fingerprint, and whether this fingerprint can discriminate between twin and non-twin siblings. In the process, we also analyze important properties of our fingerprint, such as the number of fiber trajectories, from the whole brain or individual hemispheres, required to characterize a subject's fiber trajectory geometry. Finally, we conduct a significance testing analysis to identify fiber trajectory bundles which show important differences related to the genetic proximity of siblings (i.e., twins vs non-twins), and subject gender (i.e., males vs females).

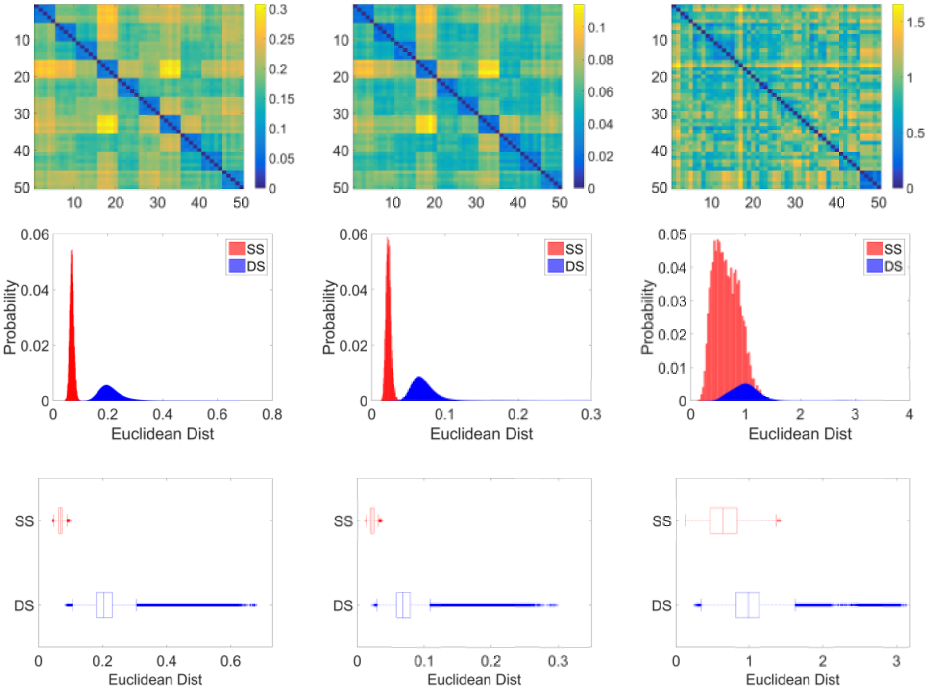
4.1. Impact of method parameters

We first analyze the impact of various parameters on the proposed subject fingerprint's ability to discriminate between subjects. The following parameters are considered in our analysis: the pooling function (i.e., RMS, Mean or Max), the dictionary size (i.e., m), the sets of dictionary learning subjects, the fiber trajectory representation sparsity (i.e., $S_{\max}$), the inclusion/exclusion of cerebellar white matter, the fiber tracking parameters, and the number of fiber trajectories used to generate the fingerprint.

The fingerprint's discriminability is measured quantitatively as follows. First, the 50 000 fiber trajectories of each subject (i.e., the 851 subjects not used for training the dictionary) are randomly divided into 5 instances, each one containing 10 000 fiber trajectories. These instances are then converted to subject fingerprints using the sparse coding and pooling process of Section 3.3, giving a total of $851 \times 5 = 4\,255$ fingerprints. Each of these fingerprints is a vector of m features, one for each dictionary bundle. We use the Euclidean distance between two fingerprints to measure their similarity, and evaluate the separability of the proposed approach by comparing the distribution of distances between same-subject instances and instances obtained from different subjects. The d-prime sensitivity index [82] is used to obtain a quantitative measure of separability:

$$\text{d-prime} \;=\; \frac{\mu_1 - \mu_2}{\sqrt{\frac{1}{2}\big(\sigma_1^2 \;+\; \sigma_2^2\big)}}, \tag{7}$$

where, μ_1, μ_2 are the means and σ_1, σ_2 the standard deviations of the compared distributions. Higher d-prime values indicate better separability. In this work we report absolute value of d-prime.



RMS
(d-prime = 4.261)

Mean
(d-prime = 3.440)

Max
(d-prime = 1.368)

Figure 7: Impact of pooling functions. Euclidean distance between fingerprints of 10 subjects with 5 instances each (top). Probability normalized histogram (middle) and box plot (bottom) for distances between same subject (SS) and different subject (DS) instances for all 851 subjects. Pooling functions: RMS, Mean, and Max (left to right columns respectively)

4.1.1. Pooling function

The impact of the pooling function on the fingerprint's ability to distinguish subjects is analyzed in Figure 7. The top row of this figure shows the Euclidean distance between all pairs of instances from 10 different subjects, where same-subject instances are grouped together. Except for the Max function, we observe a clear pattern where distances between same-subject instances (i.e., 5×5 diagonal blocks) are smaller compared to distances between different-subject instances (off diagonal block elements). Pooling functions are further compared in the middle and bottom rows of the figure, showing the normalized histogram and box plots of distances between same-subject and different-subject instances, computed for all 851 subjects. Once again, we notice a clear separation for the RMS and Mean pooling functions (d-prime of 4.261 and 3.440), but not the Max function (d-prime of 1.368). In an unpaired t-test, the means of same-subject and different-subject distances are significantly different, with $p < 0.01$.

Overall, this analysis shows that fingerprints obtained using the RMS and Mean pooling functions are significantly more similar for same-subject instances than instances from different subjects, and that the RMS function slightly outperforms Mean. As mentioned above, both functions estimate the fiber trajectory density along prominent bundles defined by the dictionary. In contrast, the Max function leads to a poorly discriminative fingerprint. This could be due to the fact that features corresponding to each bundle are estimated using a single fiber trajectory with maximum sparse code magnitude, which does not capture the full variability in bundle geometry across subjects. The RMS pooling function was used for the remaining experiments of this study.

4.1.2. Dictionary size

The size of the dictionary (i.e., parameter m), which reflects the number of different bundles that can be captured by the encoding, can also impact the quality of the fingerprint: a small number of bundles may be insufficient to capture subtle differences between subjects, while having a large number of bundles can affect the performance of the dictionary learning and sparse coding steps.

We tested seven different dictionary sizes, i.e. $m = 10, 25, 50, 75, 100, 125, 150$, while keeping the number of fiber trajectories per subject to 50 000. Note that varying m affects the number of fiber trajectories per bundle, as well as the number of features in subject fingerprints. Figure 8 (left) shows the box plot of Euclidean distances between same-subject (red) and different-subject (blue) instances, for the tested dictionary sizes. We observe that the separation between same-subject and different-subject distance distributions increases slightly with the number of bundles, mostly due to a decrease in variance for distances between different-subject instances. In summary, the separability of our subject

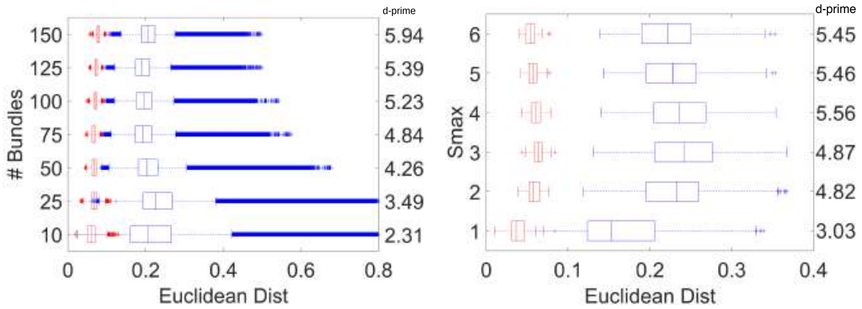


Figure 8: Impact of the size of the dictionary and the level of sparsity $S_{\max}$ on subject fingerprint. Box plot of Euclidean distances between same-subject (red) and different-subject (blue) instances for seven different dictionary sizes using all 851 subjects (left); and for varying level of the sparsity parameter $S_{\max}$ using 10 subjects (right).

fingerprint remains significant for dictionaries sizes of $m \geq 50$, and using a higher number of bundles may improve the consistency of the fingerprint. A dictionary size of $m = 50$ was used for the remaining experiments.

4.1.3. Independent dictionary sets

Since white matter geometry varies across individuals, changing the subjects used for learning the dictionary can also impact our fingerprint. To measure this impact, we created 5 different dictionaries learned from independent sets of 10 subjects, while keeping the sampling strategy and other parameters to their default values ($m = 50$). Figure 9 (top left) shows the box plot of Euclidean distances between same-subject and different-subject instances using each of these dictionaries. We observe no significant difference across dictionaries, demonstrating the robustness of our fingerprint to the choice of dictionary subjects.

4.1.4. Encoding sparsity

In the fiber trajectory encoding process, parameter $S_{\max}$ controls the level of sparsity, i.e., the maximum number of dictionary prototypes used to encode a given fiber trajectory. This parameter can also be interpreted as the maximum number of bundles to which a fiber trajectory can be assigned, thereby providing a soft fiber-to-bundle assignment for $S_{\max} > 1$.

To evaluate the impact of sparsity, we varied parameter $S_{\max}$ from 1 to 6, both for learning the dictionary and encoding new fiber trajectory data. Figure 8 (right) shows the box plots of distances between same-subject and different-subject instances, obtained from 10 subjects. We observe that the separability increases with $S_{\max}$ and saturates around $S_{\max} = 4$ (Box plots for $m = 100$ can be found in the supplementary materials). These results indicate that having a soft fiber-to-bundle assignment is necessary to capture the complex topology of bundles, which may cross or overlap one another. Since a maximum d-prime value was obtained for $S_{\max} = 4$, this sparsity level was kept for the following experiments.

4.1.5. Fiber tracking parameters

We analyzed the robustness of the proposed method to various fiber tracking parameters, for a given QSDR based signal reconstruction (in MNI space) and a fixed dictionary. For this purpose, we generated fingerprints based on the fiber trajectories of 10 subjects, obtained by varying the following parameters: the number of output fiber trajectories (from 30 000 to 150 000), the deterministic fiber tracking approach (Runge-Kutta – RK4 or Euler [40, 73]), the turning angle threshold (from 15 to 75 degrees), and the minimum length of fibers (from 20 to 250 mm). A single parameter was varied at a time, all other ones set to the value used in the previous experiments.

Figure 9 summarizes the results of this analysis, leading us to the following observations. First, we notice that the separation between same-subject (red) and different-subject (blue) instances remains similar for numbers of output fiber trajectories of 30 000 or more. Moreover, the separability of our fingerprint is nearly the same for both the RK4 and Euler fiber tracking approaches. For the turning angle threshold, the separation between the medians of the two distributions decreases as we increase the threshold's value. Increasing this threshold may lead to the generation of fibers with large curvature or very small length, which are significantly different from other fibers in the same bundle. Encoding these fibers can therefore add noise to the sparse code representation

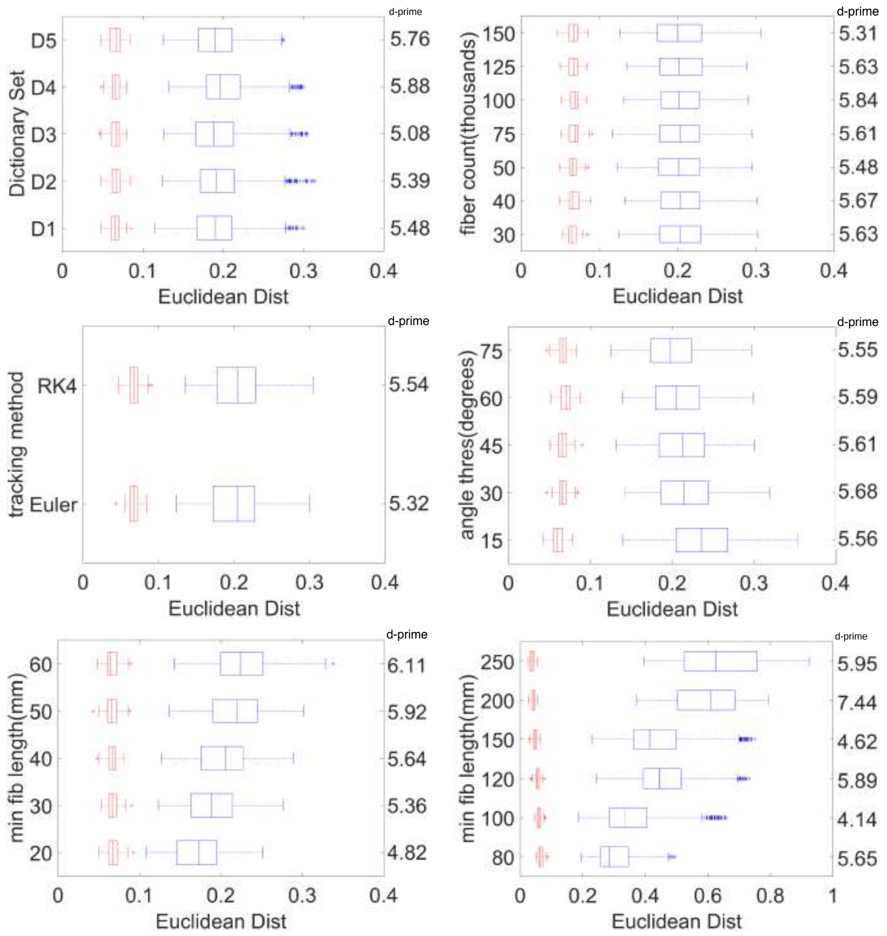


Figure 9: Impact of independent dictionary sets and fiber tracking parameters on subject fingerprints. Box plots of Euclidean distances between same-subject (red) and different-subject (blue) instances using 10 subjects for: independent sets of dictionaries; the number of output fiber trajectories; the fiber tracking approach; the turning angle threshold; and the minimum length of fiber trajectories. (d-prime values are reported along the right axis of each plot)

of subjects, resulting in a reduced separability.

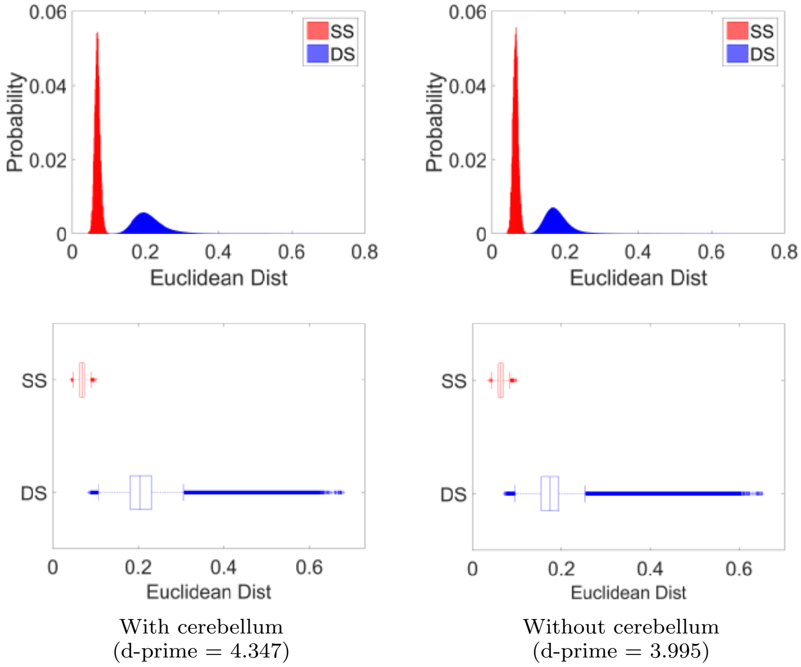
Results also show a higher separation for larger values of minimum fiber trajectory length. As highlighted in several fiber-related studies [77, 76], fiber trajectories below 40 mm in length represent short-range connections, having lower clinical relevance (e.g., surgical planning). In applications like automated fiber grouping, such fiber trajectories may pose a considerable challenge [77]. For long fiber trajectories (i.e., 80 mm to 250 mm), we observe a similar trend where the distance between distribution medians increases with minimum fiber length. However, the separation in terms of d-prime does not increase monotonically due to a higher variance in different-subject distances. Note that this phenomenon could also be explained by the fact that the dictionary used in this experiment was generated with a minimum fiber length of 40 mm. Overall, we observe that the fingerprints are quite separable across a large range of variations in these parameters.

4.1.6. Inclusion of cerebellum

The inclusion of fiber trajectories from cerebellar white matter could also impact the proposed fingerprint, due to the variability in cerebellum slice coverage across subjects. Figure 10 gives the normalized histograms and box plots of distances between same-subject and different-subject instances of all 851 subjects, obtained with and without considering the cerebellum. Fingerprints without cerebellum were obtained from the full fingerprints by removing the features corresponding to fiber trajectory bundles in the cerebellum. These bundles were determined by visual inspection of bundles in the dictionary. These results show a small decrease in separability when excluding cerebellum fiber trajectories (d-prime from 4.347 to 3.995), which could be due to the reduction in the number of bundles from 50 to 44, and also the reduction in total number of fiber trajectories contributing to the fingerprint. Nevertheless, the fingerprints generated without information from the cerebellum still exhibits significant differences across subjects.

4.1.7. Number of fingerprint fiber trajectories

Since the fingerprint (with RMS or Mean pooling) estimates the fiber trajectory density along specific bundles, another relevant question is the impact of the number of fiber trajectories n used to generate the fingerprint. If this number is low, relative to the number of bundles, it may not be possible to get



With cerebellum
(d-prime = 4.347)

Without cerebellum
(d-prime = 3.995)

Figure 10: Impact of cerebellum exclusion on subject fingerprint. Probability normalized histogram (top) and box plot (bottom) for Euclidean distances between same subject (SS) and different subject (DS) instances for all 851 subjects. Note that the fingerprint without cerebellum is obtained by removing the bundles corresponding to cerebellum from the full subject fingerprint.

an accurate measure of fiber trajectory density. To determine how this parameter affects the fingerprint's separability, we generated fingerprints for all 851 subjects using sub-samples of the subject's fiber trajectories. For every subject, five instances were created for fiber trajectory sub-sample sizes ranging from $n = 100$ to $10\,000$.

Figure 11 (left) gives the box plot of distances between same-subject and different-subject instances. We observe that the separability (i.e., d-prime) increases steadily with the number of fiber trajectories n. Moreover, we notice that separability measures increase only slightly after $n = 3\,000$, suggesting this to be the minimum number of fiber trajectories necessary to obtain a discriminative fingerprint (for a dictionary size of $m = 50$). To understand how

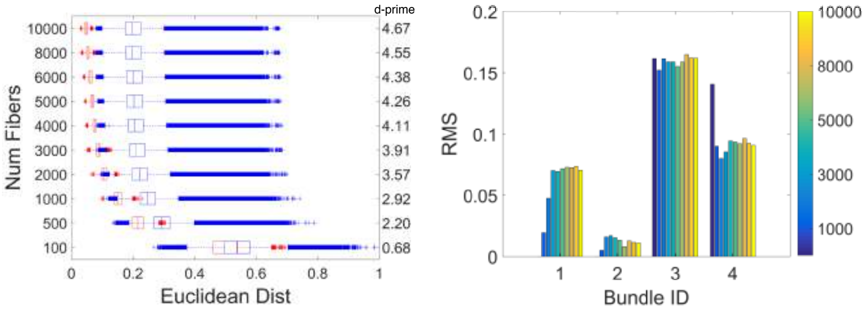


Figure 11: Impact of the number of fiber trajectories used to generate a subject fingerprint. Box plot for Euclidean distances between same-subject (red) and different-subject (blue) instances for all 851 subjects (left). Bar plot of RMS pooled features corresponding to four different bundles of a subject, obtained with varying numbers of fiber trajectories (right).

the number of fiber trajectories affects the fingerprint, Figure 11 (right) shows the RMS pooled features corresponding to four different bundles of a subject, obtained with varying numbers of fiber trajectories. We observe that pooled features stabilize for $n \geq 3\,000$, confirming our previous hypothesis.

4.2. Subject identification

The experiments presented in previous sections showed the robustness of the proposed subject fingerprint to various parameters. In this section, we apply our fingerprint to the task of identifying subjects and pairs of genetically-related subjects (i.e., twins and non-twin siblings). The objective of this analysis is two-fold: to demonstrate that the fingerprint captures characteristics of white matter geometry which can uniquely identify a subject, and to show that some of these characteristics are inheritable.

Toward this goal, we use the fingerprints obtained from each of the 4255 instances of fiber trajectory data (i.e., 851 subjects with 5 instances each), and perform a ranked retrieval analysis based on the k-nearest neighbors of a fingerprint. Given a subject and a target group (i.e., same subject, twins or non-twin siblings), we consider each of the subject's instances individually, and rank the remaining 4254 instances by their similarity to this subject instance (using the Euclidean distance between their fingerprints). Denote as T the set of instances in the target group, and let S_k be the set containing the k most similar instances. We evaluate the retrieval performance of the fingerprint, for a specific value of k, using the measures of precision and recall:

$$\text{precision@}k \ = \ \frac{|T \cap S_k|}{k}, \quad \text{recall@}k \ = \ \frac{|T \cap S_k|}{|T|}. \tag{8}$$

We report the mean precision@k and recall@k, computed over all subjects and instances.

4.2.1. Same subject identification

Table 1 gives the mean precision of the fingerprint for identifying same subject instances, using a single nearest neighbor (i.e., precision@1). In other words, we measure the frequency at which the nearest neighbor of an instance belongs to the same subject. Precision values are reported for a varying number of fiber trajectories used to generate the fingerprints (i.e., parameter n), as well as for fingerprints generated with and without cerebellum fiber trajectories. Furthermore, to evaluate the contribution of fiber trajectories across brain hemispheres, we also report the precision of fingerprints obtained using only fiber trajectories from the left hemisphere (17 bundles) or right hemisphere (15 bundles), as well as those obtained using only inter hemispheric fiber trajectories (12 bundles located mostly in the corpus callosum). Note that we obtained hemisphere-specific fingerprints from the full brain fingerprint by keeping only the features corresponding to bundles within these hemispheres. As mentioned earlier, these bundles were identified by visualization of all dictionary bundles. Finally, to evaluate the chance factor, we also computed the precision obtained from 1 000 random lists of nearest neighbors (i.e., the first k entries in a random permutation), using all $n = 10\,000$ fiber trajectories.

We observe that a mean precision@1 of 100% is achieved, both with and without cerebellum fiber trajectories, when $n = 3\,000$ or more fiber trajectories are used to generate the fingerprints. Below this number, the precision decreases monotonically to 1.0% for $n = 100$. Since a maximum precision@1 of 0.4% was obtained for the randomly generated lists of k-nearest neighbors, we conclude

Table 1: Same-subject instance identification. Mean precision@1 (in %) for a varying number of fiber trajectories using the RMS pooling function and all 851 subjects, in a nearest neighbor analysis. The second column shows results for fingerprints generated from the full brain. The third column shows result for without-cerebellum fingerprints. The right columns evaluate the contribution of fiber trajectories from a specific hemisphere. Note that the without-cerebellum fingerprints are obtained by removing cerebellum bundles from the full brain fingerprint, and the hemisphere specific fingerprints are obtained from the full brain fingerprints by keeping hemisphere-specific bundles only. Also, the first column indicates the number of fiber trajectories used for generation of the full brain fingerprint. Maximum precision@1 of 0.4% was obtained for the randomly generated lists of k-nearest neighbors using the full brain fingerprint.

# **Fibers**	**Cerebellum**		**Hemisphere**		
	Yes	No	Left	Right	Inter
100	1.4	1.0	0.4	0.4	0.4
500	36.9	21.7	5.1	3.9	3.2
1 000	85.7	68.3	17.4	14.0	10.5
2 000	99.7	97.8	54.0	41.5	27.5
3 000	100.0	99.9	77.6	67.4	46.9
4 000	100.0	100.0	88.6	81.5	61.4
5 000	100.0	100.0	94.7	89.5	73.1
6 000	100.0	100.0	97.7	93.6	81.8
8 000	100.0	100.0	99.3	98.3	91.2
10 000	100.0	100.0	99.8	99.3	95.3

that these results are significant. Furthermore, we see that the precision reduces significantly when considering only fiber trajectories from the left or right hemispheres, or just inter-hemispheric fiber trajectories. Once again, this could be due to the smaller number of features in these hemisphere-specific fingerprints, which reduces their ability to differentiate subjects. Nevertheless, for $n = 10\,000$ full-brain fiber trajectories, fingerprints generated using only single-hemisphere or inter-hemispheric fiber trajectories achieve a mean precision@1 above 95%, suggesting that characteristics unique to a subject are located in both hemispheres, as well as in crossing bundles like the corpus callosum. Comparing values across hemispheres, we notice a higher precision in the left hemisphere (e.g., precision@1 of 77.6 for $n = 3\,000$, versus 67.4 for the right hemisphere). To determine whether handedness could be a factor in this difference (i.e., 781 of the 851 subjects are right-handed), we repeated this experiment using 80 left-handed and 80 right-handed subjects. Results obtained with this setup are similar to those observed for the entire set of subjects (see Table 1 of Supplementary materials), indicating that this bilateral asymmetry is independent of subject handedness.

To analyze the robustness of our fiberprint to alignment and signal reconstruction, we generated new fingerprints for two subjects using different methods for these pre-processing steps, and tried to re-identify these two subjects with their original fingerprints. The new fingerprints were obtained by aligning the diffusion data of the subjects to the HCP 842 template [4] (MNI space, 1mm resolution, similar to the QSDR reconstruction output) using FSL [83] flirt with 12 DOF affine transform (first aligning T1w images, and then applying the affine transform to diffusion data using the applyxfm4D option). We then performed DTI signal reconstruction followed by RK4 streamline tracking (FA threshold 0.2, other parameters are kept the same). Five fingerprint instances were generated for each subject, each one obtained by randomly subsampling 5 000 fiber trajectories (see Section 3.3 for details). Note that the same dictionary as in previous experiments was employed for obtaining these fingerprints.

Figure 12 (left) compares the two subjects' tractography output obtained using the different alignment and reconstruction approaches. We can observe clear differences in the produced tractographies, highlighted by the non-overlapping red- and blue-colored fiber trajectories. Examples of fingerprint instances generated using the two processes are shown in Figure 13, the first column corresponding to an instance obtained with QSDR and rigid alignment (QSDR+rigid), and columns two and three showing two fingerprint instances based on DTI and affine alignment (DTI+affine). Although small differences are present, we

[4] http://dsi-studio.labsolver.org/download-images/hcp-842-template)

can see that our fiberprint preserves the location and relative importance of the principal fingerprint values (i.e., “peaks”) across the two different alignment and reconstruction approaches. This can be explained by the fact that the fiberprint models fiber trajectory density along prominent bundles, which is weakly affected by differences in the local geometry of individual fibers.

These results are substantiated in Figure 12 (right), where we report mean recall@k for the task of identifying the DTI+affine fingerprints using the 851 originally generated QSDR+rigid fingerprints. The mean recall@k is computed over 10 identification tasks (two subjects with 5 instance each). We observe that a mean recall@k of 100% is achieved within $k = 10$ nearest neighbors, further demonstrating the robustness of our fiberprint to alignment and signal reconstruction methods.

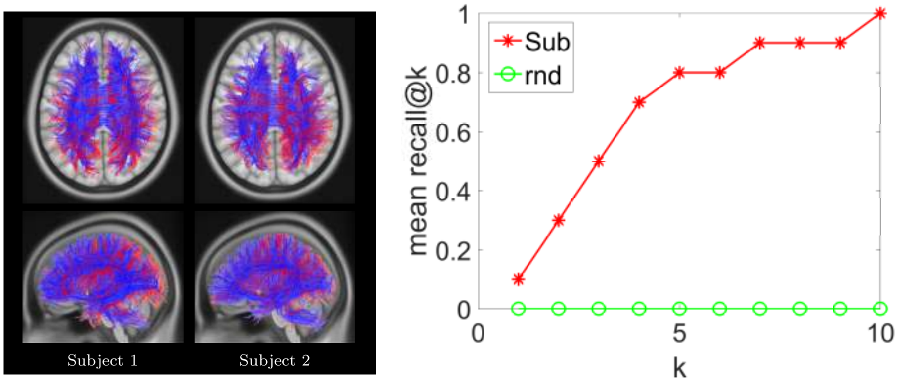


Figure 12: Comparison of QSDR+rigid alignment (blue) and DTI+affine alignment (red) based tractographies for subject 1 and subject 2 (left). Mean recall@k for DTI+affine alignment based fiberprint identification using 851 QSDR+rigid alignment fiberprints (right)

4.2.2. Genetically-related subject identification

A similar analysis was performed to identify genetically-related subjects. For this analysis, we used the Mother ID, Age, Twin stat, and Zygosity fields of the Twin HCP dataset to identify 82 pairs of monozygotic twin (MZ) subjects, 82 pairs of dizygotic twin (DZ) subjects, and 166 pairs of non-twin siblings (NT).

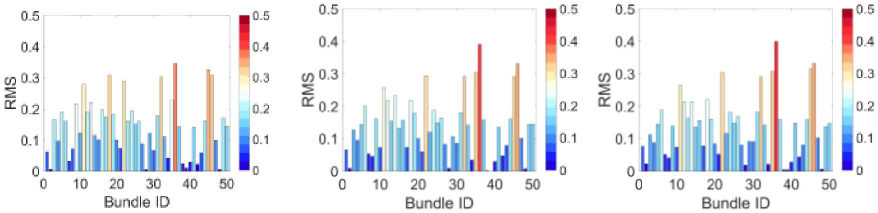


Figure 13: Color-coded bar plot representation of a subject's fiberprint, compared across the different alignment and signal reconstruction methods. Column 1 is a fiberprint based on QSDR and rigid alignment (Figure 5); columns 2 and 3 show fiberprint instances obtained with DTI and affine alignment.

For every subject having a MZ, DZ or NT sibling, we used a single instance, and obtained a measure of recall@k, for $k = 1, \ldots, 30$, by counting the ratio of MZ, DZ or NT sibling subjects within the list of k-nearest neighbors. As in 590 the previous experiment, the chance factor was considered by computing the maximum recall@k value obtained from 1 000 random lists of nearest neighbors.

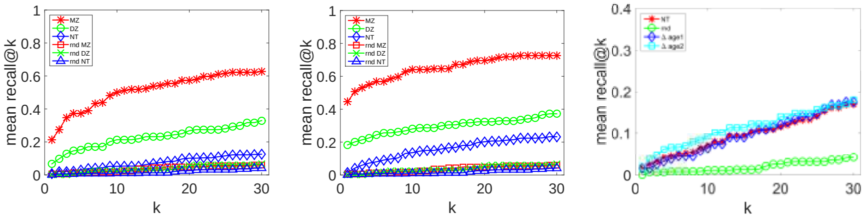


Figure 14: Genetically-related subject identification. The mean recall@k for MZ-twin (82-pairs), DZ-twin (82-pairs), Non-Twin siblings (166 pairs) using Fiberprint (left) and full T1w images rigidly aligned to MNI space as fingerprint (middle). The age difference impact on Non-Twin sibling identification, with $0 \leq \Delta age1 \leq 3$, and $3 < \Delta age2 \leq 11$, 3 being the median age difference (right). In all plots, the chance factor is measured via a random list of nearest neighbors (rnd).

Figure 14 (left) summarizes the results of this analysis. As expected, higher recall values are observed for MZ twins compared to DZ twins and non-twin siblings, reflecting the fact that such subjects have identical genetic material. 595 Moreover, a higher recall is obtained for DZ twins, in comparison to non-twin siblings. Note that, for MZ, DZ and NT pairs, the recall values obtained based on fingerprint similarity are significantly higher than those computed from random lists of nearest neighbors, validating the significance of these results.

Unlike non-twin siblings, DZ twins have the same age, a confound which might bias our analysis. To measure the true impact of this factor, we divided pairs of NT siblings in two groups based on their age difference: $0 \leq \Delta\text{age}_1 \leq 3$ and $3 < \Delta\text{age}_2 \leq 11$. Figure 14 (right) gives the recall@k values obtained for these two groups. It can be seen that NT siblings having greater age differences lead to a slightly higher recall (not statistically significant), and that recall values in both groups are significantly smaller than those observed for DZ twins, thereby eliminating age as a possible bias.

To substantiate these observations, Figure 15 gives the normalized histogram and box plots of Euclidean distances between instances belonging to MZ, DZ and NT siblings. We observe that the mean of distances corresponding to MZ twins is smaller than the mean of DZ twin distances, which is itself less than the mean distance between NT instances (d-prime values of 0.47, 0.64, and 0.26 for BMZ vs BDZ, BMZ vs BNT, and BDZ vs BNT). Note that these differences are significant in an unpaired t-test, with $p < 0.01$. Confidence intervals on the difference of distribution means are $[-0.0190, -0.0158]$, $[-0.0327, -0.0287]$, and $[-0.0154, -0.0113]$, for BMZ vs BDZ, BMZ vs BNT, and BDZ vs BNT, respectively. Overall, this analysis shows that the proposed fingerprint captures genetically-related information on the geometry of white matter.

4.2.3. Comparison with a global fingerprint based on T1-weighted images

To compare our Fiberprint with a standard morphological approach, we used the T1-weighted images (rigidly aligned to MNI space) of subjects as fingerprint and computed nearest neighbors based on the sum of squared differences (SSD) between aligned images. Figure 14 (middle) shows the mean recall@k, for $k = 1, \ldots, 30$, obtained by this fingerprint for identifying MZ, DZ and NT siblings.

We observe higher recall values for the fingerprint using T1-weighted images, compared to our Fiberprint, the most substantial differences obtained for monozygotic twins. This confirms that global brain geometry, as captured by

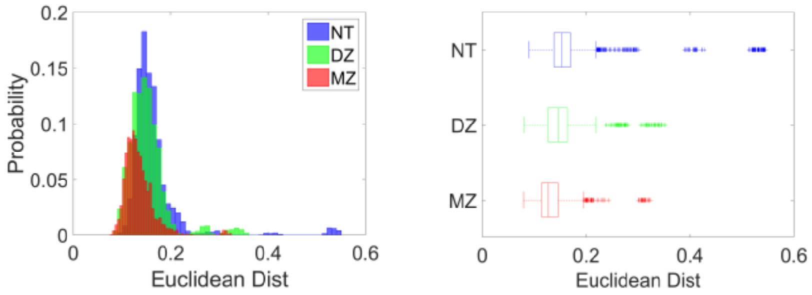


Figure 15: Differences between fingerprints of genetically-related subjects. Probability normalized histogram and box plot of Euclidean distances between instances belonging to MZ, DZ, and Non-Twin siblings

T1-weighted images, is related to genetic proximity and can be used for identifying siblings. However, the fingerprint based on T1-weighted images is much larger than the proposed Fiberprint ($157 \times 189 \times 136 = 4,035,528$ features versus $m = 50$ features for our Fiberprint), and contains a lot of information unrelated to connectivity (e.g., skull, non-white matter brain regions, etc.). In contrast, the proposed Fiberprint is highly compact and thus suitable for large-scale datasets. Moreover, it can be employed to compare subjects specifically on the level of structural connectivity, rather than global geometry.

To further assess the informativeness of our fiberprint compared to a fingerprint based on whole T1-weighted images, we computed the number of distinct and common sibling pairs (MZ/DZ/NT) identified by these two fingerprints. Toward this goal, we used the same lists of nearest neighbors as in Figure 14 and considered the identification of a sibling as successful if this sibling's fingerprint is found within the $k = 30$ nearest neighbors.

Table 2 reports the proportion of subjects for each category (mean over 5 fiberprint instances). It can be seen that the proposed fiberprint provides information complementary to the fingerprint based on raw T1 intensities, finding around 15% of siblings not identified by this fingerprint. Conversely, about 20% of siblings are identified only by the whole-image fingerprint. In summary, both fingerprints capture unique information of the similarity of genetically-related subjects.

Table 2: Informativeness of our fiberprint compared to a fingerprint based on whole T1-weighted images for identifying genetically-related subjects. Column 1 gives the proportion of twins/siblings identified by both fingerprints, Column 2 and 3 the proportion of twins/siblings identified by only one fingerprint, and column 4 the proportion of twins/siblings not identified by any of the fingerprints. A sibling is considered as identified if his/her fingerprint is within the list of $k = 30$ nearest neighbors. Number of identification tasks: 164-MZ, 164-DZ, and 215-NT. We report mean over 5 fiberprint instances.

Sibling	**Both**	**T1w**	**Fiberprint**	**None**
MZ	50.12%	22.44%	15.37%	12.07%
DZ	18.17%	19.02%	15.24%	47.56%
NT	11.35%	19.81%	14.51%	54.33%

4.3. Bundle-wise significance analysis

As mentioned before, the proposed fingerprint encodes fiber trajectory geometry along bundles defined by the dictionary. In this section, we evaluate the significance of individual bundles by comparing the distribution of fingerprint features in instances corresponding to different subject groups (e.g., DZ twins vs non-twin siblings, male vs female, etc.).

This bundle-wise analysis of significance uses the distributions of fingerprint features corresponding to specific bundles, in instances belonging to two different subject groups. For each of the 50 dictionary bundles, we obtain a p-value using a Wilcoxon rank-sum test[5], representing the confidence at which we can reject the hypothesis that the two distributions are equal. To account for multiple comparisons, we correct these p-values using the Holm-Bonferroni method [84] and consider as significant the bundles with corrected $p < 0.05$.

[5] Results obtained using an unpaired t-test can be found in the supplementary materials.

4.3.1. Differences across genetically-related subjects

We first identify the bundles which show a statistically significant difference across two groups of genetically-related subjects. As in the subject identification experiment, we compute the pairwise distances between instances corresponding to MZ twins, DZ twins and non-twin siblings, considering each fingerprint feature (i.e., bundle) individually. The significance of a bundle is measured based on the null hypothesis that the distances in two groups are equally distributed.

Figure 16 shows the Holm-Bonferroni corrected p-values (in -$\log_{10}$ scale) of each bundle, for MZ twins compared to non-twin siblings. The results identify three separate bundles with significant differences (-$\log_{10}$(p-value) > 1.3) corresponding to the corticospinal bundles, with fiber trajectories in the parietal lobe and dorsal regions of the brain. Furthermore, bundle-wise differences between DZ and NT siblings, occurring mainly in frontal cortex areas, can also be seen in Figure 17.

4.3.2. Differences related to gender

A similar analysis was conducted to find bundles showing statistically significant differences between male and female subjects. For this analysis, we used the data from 332 males (age: 28.05 ± 3.65) and 436 females (age: 29.33 ± 3.55), all of them right-handed. While the analysis on genetically-related subjects compared distance distributions, in this case, we compared features directly. That is, for each bundle, we computed the distribution of feature values corresponding to this bundle, and compared the distributions obtained in instances of male and female subjects.

Figure 18 reports the corrected p-values (in -$\log_{10}$ scale) obtained for each bundle. We can see several significant bundles (14 in total), with corrected $p < 0.05$, with the most prominent differences occurring in the frontal cortex. Specifically, significant bundles include fiber trajectories in the pre-frontal area, and around the precuneus.

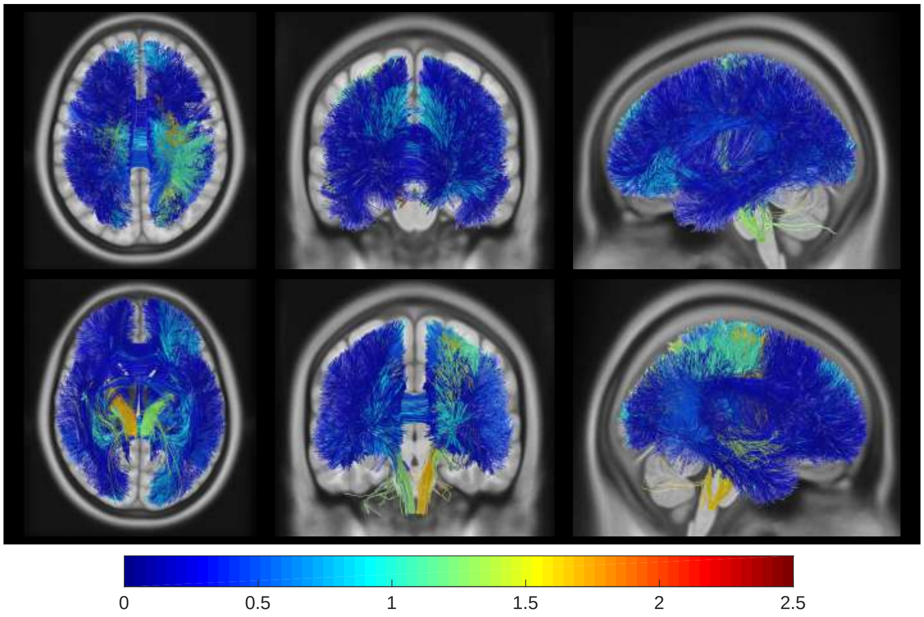


Figure 16: **MZ vs NT**. Differences between MZ-twin and Non-Twin siblings. Color coded bundle visualization for Holm-Bonferroni corrected p-values (in -$\log_{10}$ scale) obtained using a Wilcoxon rank-sum test. (superior axial, anterior coronal, and left sagittal views (top row); inferior axial, posterior coronal, and right sagittal views (bottom row);)

5. Discussion

We now summarize and discuss the findings related to our parameter study, subject identification experiments, and bundle-wise significance tests. We then highlight limitations and additional considerations of this study.

5.1. Findings related to the parameter study

An extensive set of experiments was conducted to determine the impact of various parameters on the fingerprint's ability to uniquely characterize a subject. These experiments showed that pooling functions estimating the fiber trajectory density along dictionary bundles, such as the RMS and Mean functions, provided fingerprints that were significantly more similar for same-subject instances than those from different subjects. Moreover, fingerprints obtained using RMS pooling were found to give significant separability for dictionaries containing

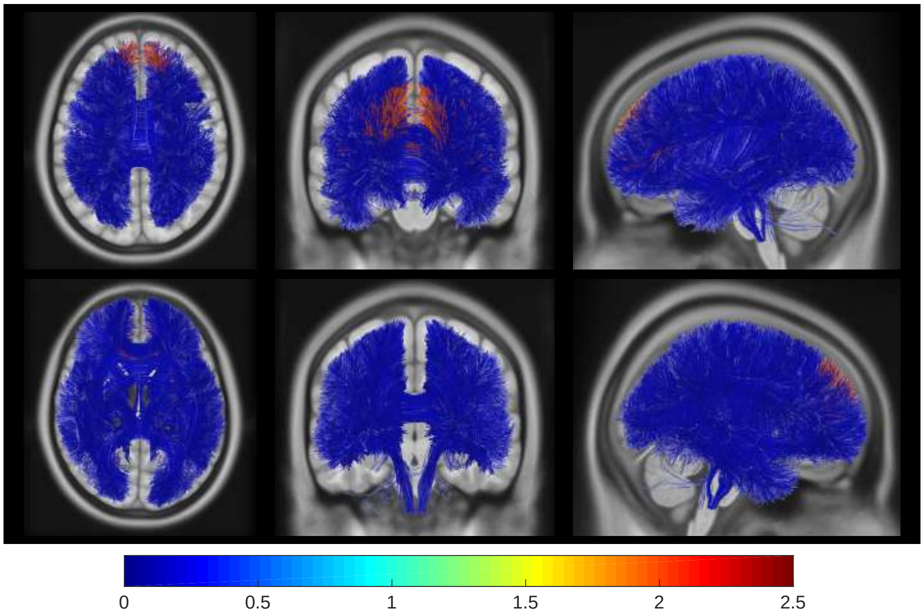


Figure 17: **DZ vs NT**. Differences between DZ-twin and Non-Twin siblings. Color coded bundle visualization for Holm-Bonferroni corrected p-values (in -$\log_{10}$ scale) obtained using a Wilcoxon rank-sum test. (superior axial, anterior coronal, and left sagittal views (top row); inferior axial, posterior coronal, and right sagittal views (bottom row);)

50 bundles or more, a number consistent with previous studies on the topic of fiber trajectory clustering and segmentation [12, 14]. Our experiments have also shown the advantage of using a soft assignment of fiber trajectories to bundles, via our non-negative sparse coding framework, which offers a more precise description of complex bundles that may overlap and cross each other. Specifically, we observed that fiber trajectories can be encoded as a sparse combination of $S_{\max} = 4$ bundle prototypes. This sparsity level was also found to be optimal in our previous work on fiber trajectory segmentation [38].

We evaluated the robustness of the proposed method by varying the fiber tracking parameters. Our method provides separability for 30 000 or more output fiber trajectories, both using the RK4 and Euler fiber tracking approaches. The tracking parameters having the highest impact are the turning angle threshold and minimum fiber trajectory length, although significant separability was

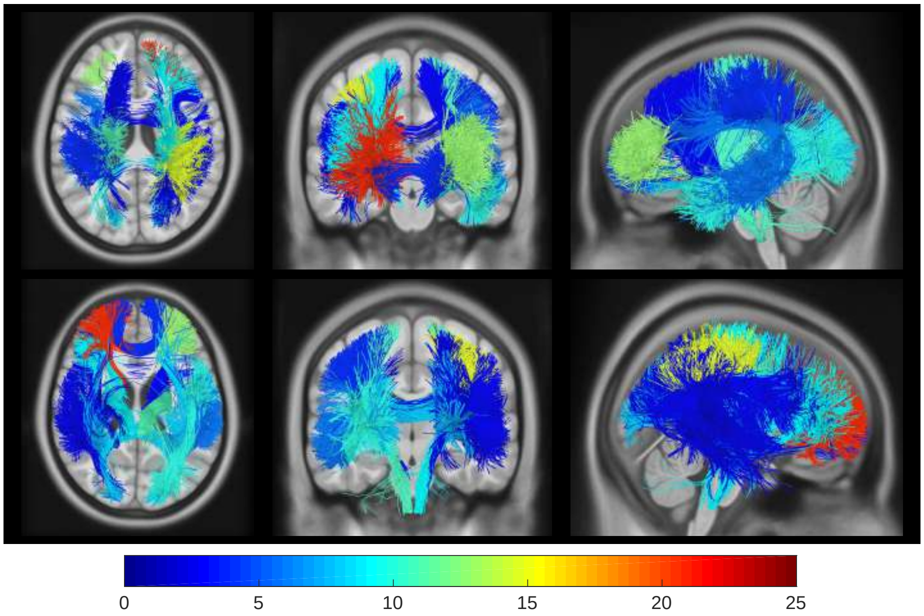


Figure 18: **Male vs Female**. Differences related to gender. Color coded bundle visualization for Holm-Bonferroni corrected p-values (in -$\log_{10}$ scale) obtained using a Wilcoxon rank-sum test. (superior axial, anterior coronal, and left sagittal views (top row); inferior axial, posterior coronal, and right sagittal views (bottom row);) **Note**: for visualization purposes, fibers in non-signifcant bundles are not shown.

achieved for all tested values of these parameters. In another experiment, we found that excluding cerebellum fiber trajectories resulted in a small decrease in separability. However, the fingerprint without information from the cerebellum still exhibit significant differences across subjects.

Varying the number of fiber trajectories used for fingerprint generation, we observed that our fingerprint could uniquely identify a subject with only 3 000 fiber trajectories uniformly sampled over the whole brain. Moreover, we found that fiber trajectories from both hemispheres and inter-hemispheric fiber trajectories contributed in a synergic manner to characterize a subject, the highest separability obtained using left-hemisphere fiber trajectories. This suggests that unique characteristics of a subject, in terms of fiber trajectory distribution, are present in the entire brain. Overall, the small variations found in individual bundles, across subjects, suggest a common blueprint of connectivity, but also an overall pattern that is unique to each individual. This is consistent with previous work in the literature, showing that each individual is unique in terms of brain structure [17], function [19, 21], and white matter micro-structure [25, 26].

5.2. Findings related to subject identification tests

Our experiments on subject identification have also lead to useful observations. Using fingerprint similarity to define the k-nearest neighbors of a subject instance, we obtained results consistent with previous work from the literature, showing that MZ twins are significantly more similar at the fingerprint level than DZ twins, and DZ twins more similar than non-twin siblings [85]. Results also showed a greater similarity between DZ twins than between non-twin siblings, although both types of siblings have the same amount of shared genetic information. A deeper analysis revealed that the higher similarity of DZ twins was not fully explained by age difference. While studies have shown the impact of various environmental factors on white matter development [5], in particular during adolescence, the link between the fetal environment and brain development remains largely unknown. Further investigation is required to determine whether prenatal development factors, like the mother's nutrition and stress levels during pregnancy, could play a role in our observations.

There are many factors to be considered while interpreting these results, for example, environmental factors, gender differences, aging effects, limitations of fiber tracking processes, non-rigid alignment process, etc. Note the twin zygosity labels used in this analysis were self reported (HCP Q3 release). The impact of aging was addressed indirectly by the HCP dataset recruitment policies, which limited the allowed age of subjects to the 22-35 years range, corresponding to a plateau in the FA-aging curve [86, 85, 67]. We also considered the effect of aging for identifying twins and non-twin siblings by dividing pairs of non-twin siblings into two groups, using the median age difference as the separation threshold. No significant difference was observed across age groups, for age differences up to 11 years.

5.3. Findings related to bundle significance tests

Our bundle-wise fingerprint analysis revealed several bundles showing significant differences, when comparing groups of genetically-related subjects, or different sex subjects. For the comparison between MZ twins and Non-Twin siblings, we find three significant bundles ($p < 0.05$ after Holm-Bonferroni correction), corresponding roughly to the corticospinal bundles. The differences between DZ twins and Non-Twin siblings were most prominent in the frontal cortex, suggesting that variations between individuals sharing the same amount of genetic material are linked to higher processing areas. Although a direct comparison is not feasible, these results are consistent with white matter regions in a recent heritability study, based on the voxel-wise analysis of fractional anisotropy (FA) [85].

Moreover, gender-related differences were found to be significant in 14 different bundles, connected mostly to the pre-frontal cortex and precuneus. Again, several of these bundles correspond to regions shown to have significant gender-related effects on FA, in studies using tract based spatial statistics (TBSS) [87, 88] or structural network analysis [89].

5.4. Informativeness of fiberprint compared to fingerprints based on whole T1-weighted images

Comparing the proposed fiberprint with a brain fingerprint generated from intensities in aligned T1w volumes, the two fingerprints yield a similar performance (measured in terms of recall@k) for the task of identifying genetically-related subjects. However, analyzing the list of sibling pairs (MZ/DZ/NT) identified by these two fingerprints indicates that each one provides complimentary information, with 15% to 20% of sibling pairs identified by only one of these fingerprints.

Although using raw intensities as fingerprint also allows to capture both local and global differences in structural or diffusion geometry, the proposed fiberprint provides a more compact and high-level representation of white-matter connectivity. Thus, our fiberprint can effectively encode this information in a vector of about $m = 50$ features, compared to $157 \times 189 \times 136$ features for T1-weighted volumes. This makes our framework particularly attractive for handling large datasets. Moreover, direct voxelwise comparison of diffusion data (e.g., FA maps) could also be challenging, since high-contrast edges in such volumes make them more susceptible to small registration errors and to the variability of local tract geometry [36]. In contrast, our experiments have shown the proposed fiberprint to be robust to differences in the alignment and signal reconstruction process. Lastly, unlike voxelwise fingerprints, our framework allows comparing subjects on the level of structural connectivity (i.e., fiber bundles), rather than unspecified global structure.

5.5. Additional considerations

In this study, we analyzed the impact of various factors on our fingerprint's ability to describe unique characteristics of subjects. However, additional factors could be considered in our analysis. For instance, other distances metrics can be used to measure the similarity between fiber trajectories, such as the Mean of Closest Points (MCP) or the Hausdorff distance. The flexibility of the proposed framework allows its potential extension to various computational models or representations for fiber trajectories, for which a similarity measure can be computed. These measures could help capture additional information on fiber trajectories (e.g., along-tract diffusion signal), which may be not possible to encode with a geometric representation, leading to a more discriminative fingerprint.

Partial volume effects and other tractography-related effects, such as fiber tracking or registration errors, could also impact our fingerprint. Moreover, as highlighted in [74], caution must be used to when interpreting results obtained from diffusion MRI. For instance, since there is no gold standard for calibrating DWI measures, the reliability of tractography outputs cannot be evaluated. However, these factors are in part minimized by the large number of subjects used in our study (i.e., 851 subjects), the pre-processing done by the HCP pipeline and the QSDR signal reconstruction approach.

In our experiments, we have created multiple instances of the same subject using fiber trajectories derived from a single scan. Another aspect could be to test same subject identification using repeat scans of the same subject, as done in [32] for the study of white matter structure. Since we use the same reconstruction approach and toolbox (DSI studio), the results after fiber tracking should extend to repeat scan data. Moreover, because our experiments have demonstrated that fingerprints generated from the scans of identical siblings are more similar than those from other sibling types, we expect repeat scans of the same subject to have highly similar fingerprints.

Although aging effects were considered in our analysis of bundle-wise significance, a deeper study is needed to fully understand the impact of neuroplasticity on fingerprints. This could also be achieved using longitudinal data, by measuring how a subject's fingerprint changes over time. Our bundle-wise significance analysis could also be extended to find differences related to additional phenotypic variables, such as cognitive score.

6. Conclusion

We presented a new subject fingerprint, called Fiberprint, which uses sparse code pooling to characterize the unique properties of subjects at the level of fiber trajectories. The proposed fingerprint measures the fiber trajectory density along specific bundles, which are defined using dictionary learning. Experiments using the dMRI data of 861 subjects from the HCP dataset were conducted to evaluate the impact of our method's parameters, to demonstrate that the proposed fingerprint can be used to identify subjects, pairs of twins, or non-twin siblings, and to find bundles showing significant differences across various subject groups.

Our results show that a fingerprint capable of uniquely identifying subjects can be obtained using only 3 000 fiber trajectories sampled across the brain. Moreover, such a fingerprint is robust to parameters related to fiber tracking, dictionary learning and sparse code pooling. Experiments on the identification of genetically-related subjects demonstrate that the proposed fingerprint can correctly retrieve instances belonging to a given subject. Our experiments also suggest that subjects sharing the same genetic information (i.e., monozygotic twins) have more similar fingerprints than siblings sharing half of their genetic material (i.e., dizygotic twins and non-twin siblings). Furthermore, our bundle-wise analysis of significance showed that corticospinal bundles had significantly different fingerprint features when comparing monozygotic twins with non-twin siblings, and that differences between dizygotic twins and non-twin siblings were most prominent in the pre-frontal cortex. A similar comparison across male and females subjects identified 14 significant bundles, most of them connected to the pre-frontal cortex and precuneus. Several of these results are consistent with recent heritability studies based on the voxel-wise analysis of FA.

This work could be extended by evaluating the impact of additional factors related to the tracking and encoding of fiber trajectories. Likewise, a deeper analysis of aging effects could help better understanding the effect of neuroplasticity on individual characteristics of white matter fiber geometry.

Acknowledgements

Data were provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

References

[1] P. J. Basser, J. Mattiello, D. LeBihan, MR diffusion tensor spectroscopy and imaging., Biophysical journal 66 (1) (1994) 259.

[2] H.-E. Assemlal, D. Tschumperlé, L. Brun, K. Siddiqi, Recent advances in diffusion MRI modeling: Angular and radial reconstruction, Medical image analysis 15 (4) (2011) 369–396.

[3] D. S. Tuch, Q-ball imaging, Magnetic resonance in medicine 52 (6) (2004) 1358–1372.

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

[5] M.-C. Chiang, K. L. McMahon, G. I. de Zubicaray, N. G. Martin, I. Hickie, A. W. Toga, M. J. Wright, P. M. Thompson, Genetics of white matter development: a DTI study of 705 twins and their siblings aged 12 to 29, Neuroimage 54 (3) (2011) 2308–2317.

[6] N. Jahanshad, D. P. Hibar, A. Ryles, A. W. Toga, K. L. McMahon, G. I. De Zubicaray, N. K. Hansell, G. W. Montgomery, N. G. Martin, M. J. Wright, et al., Discovery of genes that affect human brain connectivity: a genome-wide analysis of the connectome, in: 2012 9th IEEE International Symposium on Biomedical Imaging (ISBI), IEEE, 2012, pp. 542–545.

[7] P. M. Thompson, T. Ge, D. C. Glahn, N. Jahanshad, T. E. Nichols, Genetics of the connectome, Neuroimage 80 (2013) 475–488.

[8] H.-W. Chung, M.-C. Chou, C.-Y. Chen, Principles and limitations of computational algorithms in clinical diffusion tensor MR tractography, American Journal of Neuroradiology 32 (1) (2011) 3–13.

[9] A. Griffa, P. S. Baumann, J.-P. Thiran, P. Hagmann, Structural connectomics in brain diseases, Neuroimage 80 (2013) 515–526.

[10] P. Hagmann, L. Jonasson, P. Maeder, J.-P. Thiran, V. J. Wedeen, R. Meuli, Understanding diffusion MR imaging techniques: from scalar diffusion-weighted imaging to diffusion tensor imaging and beyond 1, Radiographics 26 (suppl_1) (2006) S205–S223.

[11] M. E. Thomason, P. M. Thompson, Diffusion imaging, white matter, and psychopathology, Clinical Psychology 7 (1) (2011) 63.

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

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

[14] L. J. O'Donnell, C.-F. Westin, Automatic tractography segmentation using a high-dimensional white matter atlas, IEEE transactions on medical imaging 26 (11) (2007) 1562–1575.

[15] G. Prasad, S. H. Joshi, N. Jahanshad, J. Villalon-Reina, I. Aganj, C. Lenglet, G. Sapiro, K. L. McMahon, G. I. de Zubicaray, N. G. Martin, et al., Automatic clustering and population analysis of white matter tracts using maximum density paths, Neuroimage 97 (2014) 284–295.

[16] K. Amunts, A. Malikovic, H. Mohlberg, T. Schormann, K. Zilles, Brodmann's areas 17 and 18 brought into stereotaxic spacewhere and how variable?, Neuroimage 11 (1) (2000) 66–84.

[17] J.-F. Mangin, D. Riviere, A. Cachia, E. Duchesnay, Y. Cointepas, D. Papadopoulos-Orfanos, P. Scifo, T. Ochiai, F. Brunelle, J. Regis, A framework to study the cortical folding patterns, Neuroimage 23 (2004) S129–S138.

[18] J. Rademacher, U. Bürgel, S. Geyer, T. Schormann, A. Schleicher, H.-J. Freund, K. Zilles, Variability and asymmetry in the human precentral motor system, Brain 124 (11) (2001) 2232–2258.

[19] D. M. Barch, G. C. Burgess, M. P. Harms, S. E. Petersen, B. L. Schlaggar, M. Corbetta, M. F. Glasser, S. Curtiss, S. Dixit, C. Feldt, et al., Function in

the human connectome: task-fMRI and individual differences in behavior, Neuroimage 80 (2013) 169–189.

[20] R. H. Grabner, D. Ansari, G. Reishofer, E. Stern, F. Ebner, C. Neuper, Individual differences in mathematical competence predict parietal brain activation during mental calculation, Neuroimage 38 (2) (2007) 346–356.

[21] S. Mueller, D. Wang, M. D. Fox, B. T. Yeo, J. Sepulcre, M. R. Sabuncu, R. Shafee, J. Lu, H. Liu, Individual variability in functional connectivity architecture of the human brain, Neuron 77 (3) (2013) 586–595.

[22] M. V. Ruiz-Blondet, Z. Jin, S. Laszlo, CEREBRE: A novel method for very high accuracy event-related potential biometric identification, IEEE Transactions on Information Forensics and Security 11 (7) (2016) 1618–1629.

[23] B. Rypma, M. DEsposito, The roles of prefrontal brain regions in components of working memory: effects of memory load and individual differences, Proceedings of the National Academy of Sciences 96 (11) (1999) 6558–6563.

[24] K. Zilles, K. Amunts, Individual variability is not noise, Trends in cognitive sciences 17 (4) (2013) 153–155.

[25] U. Bürgel, K. Amunts, L. Hoemke, H. Mohlberg, J. M. Gilsbach, K. Zilles, White matter fiber tracts of the human brain: three-dimensional mapping at microscopic resolution, topography and intersubject variability, Neuroimage 29 (4) (2006) 1092–1105.

[26] M. T. de Schotten, A. Bizzi, F. Dell'Acqua, M. Allin, M. Walshe, R. Murray, S. C. Williams, D. G. Murphy, M. Catani, et al., Atlasing location, asymmetry and inter-subject variability of white matter tracts in the human brain with mr diffusion tractography, Neuroimage 54 (1) (2011) 49–59.

[27] B. C. Armstrong, M. V. Ruiz-Blondet, N. Khalifian, K. J. Kurtz, Z. Jin, S. Laszlo, Brainprint: Assessing the uniqueness, collectability, and perma-

nence of a novel method for ERP biometrics, Neurocomputing 166 (2015) 59–67.

[28] E. S. Finn, X. Shen, D. Scheinost, M. D. Rosenberg, J. Huang, M. M. Chun, X. Papademetris, R. T. Constable, Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity, Nature neuroscience.

[29] O. Miranda-Dominguez, B. D. Mills, S. D. Carpenter, K. A. Grant, C. D. Kroenke, J. T. Nigg, D. A. Fair, Connectotyping: model based fingerprinting of the functional connectome, PloS one 9 (11) (2014) e111048.

[30] B. Mišić, O. Sporns, From regions to connections and networks: new bridges between brain and behavior, Current opinion in neurobiology 40 (2016) 1–7.

[31] C. Wachinger, P. Golland, W. Kremen, B. Fischl, M. Reuter, A. D. N. Initiative, et al., Brainprint: a discriminative characterization of brain morphology, NeuroImage 109 (2015) 232–248.

[32] F.-C. Yeh, J. Vettel, A. Singh, B. Poczos, S. Grafton, K. Erickson, W.-Y. Tseng, T. Verstynen, Local connectome fingerprinting reveals the uniqueness of individual white matter architecture, bioRxiv (2016) 043778.

[33] F.-C. Yeh, J. M. Vettel, A. Singh, B. Poczos, S. T. Grafton, K. I. Erickson, W.-Y. I. Tseng, T. D. Verstynen, Quantifying differences and similarities in whole-brain white matter architecture using local connectome fingerprints, PLOS Computational Biology 12 (11) (2016) e1005203.

[34] M. Toews, W. M. Wells, How are siblings similar? how similar are siblings? large-scale imaging genetics using local image features, in: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), 2016, pp. 847–850. doi:10.1109/ISBI.2016.7493398.

[35] F.-C. Yeh, D. Badre, T. Verstynen, Connectometry: A statistical approach harnessing the analytical potential of the local connectome, NeuroImage 125 (2016) 162–171.

[36] J. B. Colby, L. Soderberg, C. Lebel, I. D. Dinov, P. M. Thompson, E. R. Sowell, Along-tract statistics allow for enhanced tractography analysis, Neuroimage 59 (4) (2012) 3227–3242.

[37] K. Kumar, C. Desrosiers, K. Siddiqi, Brain fiber clustering using non-negative kernelized matching pursuit, in: International Workshop on Machine Learning in Medical Imaging, Springer, 2015, pp. 144–152.

[38] K. Kumar, C. Desrosiers, A sparse coding approach for the efficient representation and segmentation of white matter fibers, in: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), IEEE, 2016, pp. 915–919.

[39] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, 2009, pp. 1794–1801.

[40] P. J. Basser, S. Pajevic, C. Pierpaoli, J. Duda, A. Aldroubi, In vivo fiber tractography using DT-MRI data, Magnetic resonance in medicine 44 (4) (2000) 625–632.

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

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

[43] L. J. O'Donnell, A. J. Golby, C.-F. Westin, Fiber clustering versus the parcellation-based connectome, NeuroImage 80 (2013) 283–289.

[44] I. Corouge, S. Gouttard, G. Gerig, Towards a shape model of white matter fiber bundles using diffusion tensor MRI, in: Biomedical Imaging: Nano to Macro, 2004. IEEE International Symposium on, IEEE, 2004, pp. 344–347.

[45] G. Gerig, S. Gouttard, I. Corouge, Analysis of brain white matter via fiber tract modeling, in: Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE, Vol. 2, IEEE, 2004, pp. 4421–4424.

[46] A. Brun, H. Knutsson, H.-J. Park, M. E. Shenton, C.-F. Westin, Clustering fiber traces using normalized cuts, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2004, pp. 368–375.

[47] L. Jonasson, P. Hagmann, J. Thiran, V. Wedeen, Fiber tracts of high angular resolution diffusion MRI are easily segmented with spectral clustering., in: Proceedings of 13th Annual Meeting ISMRM, Miami, no. EPFL-CONF-87232, SPIE, 2005, p. 1310.

[48] L. ODonnell, C.-F. Westin, White matter tract clustering and correspondence in populations, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2005, pp. 140–147.

[49] B. Moberts, A. Vilanova, J. J. van Wijk, Evaluation of fiber clustering methods for diffusion tensor imaging, in: VIS 05. IEEE Visualization, 2005., IEEE, 2005, pp. 65–72.

[50] Z. Ding, J. C. Gore, A. W. Anderson, Classification and quantification of neuronal fiber pathways using diffusion tensor MRI, Magnetic Resonance in Medicine 49 (4) (2003) 716–721.

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

[52] L. J. ODonnell, W. M. Wells III, A. J. Golby, C.-F. Westin, Unbiased groupwise registration of white matter tractography, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2012, pp. 123–130.

[53] D. Wassermann, L. Bloy, E. Kanterakis, R. Verma, R. Deriche, Unsupervised white matter fiber clustering and tract probability map generation: Applications of a gaussian process framework for white matter fibers, NeuroImage 51 (1) (2010) 228–241.

[54] M. Maddah, W. E. L. Grimson, S. K. Warfield, W. M. Wells, A unified framework for clustering and quantitative analysis of white matter fiber tracts, Medical image analysis 12 (2) (2008) 191–202.

[55] X. Wang, W. E. L. Grimson, C.-F. Westin, Tractography segmentation using a hierarchical dirichlet processes mixture model, NeuroImage 54 (1) (2011) 290–302.

[56] S. Durrleman, P. Fillard, X. Pennec, A. Trouvé, N. Ayache, A statistical model of white matter fiber bundles based on currents, in: International Conference on Information Processing in Medical Imaging, Springer, 2009, pp. 114–125.

[57] P. Gori, O. Colliot, L. Marrakchi-Kacem, Y. Worbe, F. D. V. Fallani, M. Chavez, C. Poupon, A. Hartmann, N. Ayache, S. Durrleman, Parsimonious approximation of streamline trajectories in white matter fiber bundles, IEEE Transactions on Medical Imaging 35 (12) (2016) 2609–2619.

[58] M. Elad, M. A. Figueiredo, Y. Ma, On the role of sparse and redundant representations in image processing, Proceedings of the IEEE 98 (6) (2010) 972–982.

[59] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE transactions on pattern analysis and machine intelligence 31 (2) (2009) 210–227.

[60] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, S. Yan, Sparse representation for computer vision and pattern recognition, Proceedings of the IEEE 98 (6) (2010) 1031–1044.

[61] M. Lustig, D. L. Donoho, J. M. Santos, J. M. Pauly, Compressed sensing MRI, IEEE Signal Processing Magazine 25 (2) (2008) 72–82.

[62] T. Tong, R. Wolz, P. Coupé, J. V. Hajnal, D. Rueckert, A. D. N. Initiative, et al., Segmentation of MR images via discriminative dictionary learning and sparse coding: Application to hippocampus labeling, NeuroImage 76 (2013) 11–23.

[63] K. Lee, J.-M. Lina, J. Gotman, C. Grova, SPARK: Sparsity-based analysis of reliable k-hubness and overlapping network structure in brain functional connectivity, NeuroImage 134 (2016) 434–449.

[64] Y.-B. Lee, J. Lee, S. Tak, K. Lee, D. L. Na, S. W. Seo, Y. Jeong, J. C. Ye, A. D. N. Initiative, et al., Sparse SPM: Group sparse-dictionary learning in SPM framework for resting-state functional connectivity MRI analysis, NeuroImage 125 (2016) 1032–1045.

[65] H. E. Çetingül, M. J. Wright, P. M. Thompson, R. Vidal, Segmentation of high angular resolution diffusion MRI using sparse riemannian manifold clustering, IEEE transactions on medical imaging 33 (2) (2014) 301–317.

[66] D. Zhu, N. Jahanshad, B. C. Riedel, L. Zhan, J. Faskowitz, G. Prasad, P. M. Thompson, Population learning of structural connectivity by white matter encoding and decoding, in: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), IEEE, 2016, pp. 554–558.

[67] D. C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. W. Curtiss, et al., The human

connectome project: a data acquisition perspective, Neuroimage 62 (4) (2012) 2222–2231.

[68] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium, et al., The WU-minn human connectome project: an overview, Neuroimage 80 (2013) 62–79.

[69] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, et al., The minimal preprocessing pipelines for the human connectome project, Neuroimage 80 (2013) 105–124.

[70] F.-C. Yeh, W.-Y. I. Tseng, NTU-90: a high angular resolution brain atlas constructed by q-space diffeomorphic reconstruction, Neuroimage 58 (1) (2011) 91–99.

[71] F.-C. Yeh, V. J. Wedeen, W.-Y. I. Tseng, Generalized-sampling imaging, IEEE transactions on medical imaging 29 (9) (2010) 1626–1635.

[72] J. Ashburner, Computational neuroanatomy, Ph.D. thesis, University College London (2000).
URL `/spm/doc/theses/john/`

[73] F.-C. Yeh, T. D. Verstynen, Y. Wang, J. C. Fernández-Miranda, W.-Y. I. Tseng, Deterministic diffusion fiber tracking improved by quantitative anisotropy, PloS one 8 (11) (2013) e80713.

[74] D. K. Jones, T. R. Knösche, R. Turner, White matter integrity, fiber count, and other fallacies: the do's and don'ts of diffusion mri, Neuroimage 73 (2013) 239–254.

[75] C. Thomas, Q. Y. Frank, M. O. Irfanoglu, P. Modi, K. S. Saleem, D. A. Leopold, C. Pierpaoli, Anatomical accuracy of brain connections derived from diffusion mri tractography is inherently limited, Proceedings of the National Academy of Sciences 111 (46) (2014) 16574–16579.

[76] L. J. ODonnell, Y. Suter, L. Rigolo, P. Kahali, F. Zhang, I. Norton, A. Albi, O. Olubiyi, A. Meola, W. I. Essayed, et al., Automated white matter fiber tract identification in patients with brain tumors, NeuroImage: Clinical 13 (2017) 138–153.

[77] E. Garyfallidis, M. Brett, M. M. Correia, G. B. Williams, I. Nimmo-Smith, Quickbundles, a method for tractography simplification, Frontiers in neuroscience 6 (2012) 175.

[78] H. Nguyen, V. M. Patel, N. M. Nasrabadi, R. Chellappa, Kernel dictionary learning, in: ICASSP 2012, IEEE, 2012, pp. 2021–2024.

[79] I. S. Dhillon, Y. Guan, B. Kulis, Kernel k-means: spectral clustering and normalized cuts, in: SIGKDD 2004, ACM, 2004, pp. 551–556.

[80] C. Ding, T. Li, W. Peng, H. Park, Orthogonal nonnegative matrix t-factorizations for clustering, in: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2006, pp. 126–135.

[81] C. Fowlkes, S. Belongie, F. Chung, J. Malik, Spectral grouping using the nystrom method, IEEE transactions on pattern analysis and machine intelligence 26 (2) (2004) 214–225.

[82] S. D. Gale, D. J. Perkel, A basal ganglia pathway drives selective auditory responses in songbird dopaminergic neurons via disinhibition, The Journal of Neuroscience 30 (3) (2010) 1027–1037.

[83] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, S. M. Smith, Fsl, Neuroimage 62 (2) (2012) 782–790.

[84] S. Holm, A simple sequentially rejective multiple test procedure, Scandinavian journal of statistics (1979) 65–70.

[85] P. Kochunov, N. Jahanshad, D. Marcus, A. Winkler, E. Sprooten, T. E. Nichols, S. N. Wright, L. E. Hong, B. Patel, T. Behrens, et al., Heritability

of fractional anisotropy in human white matter: a comparison of human connectome project and ENIGMA-DTI data, Neuroimage 111 (2015) 300–311.

[86] P. Kochunov, D. C. Glahn, J. Lancaster, P. M. Thompson, V. Kochunov, B. Rogers, P. Fox, J. Blangero, D. Williamson, Fractional anisotropy of cerebral white matter and thickness of cortical gray matter across the lifespan, Neuroimage 58 (1) (2011) 41–49.

[87] K.-H. Chou, Y. Cheng, I.-Y. Chen, C.-P. Lin, W.-C. Chu, Sex-linked white matter microstructure of the social and analytic brain, Neuroimage 54 (1) (2011) 725–733.

[88] G. Gong, Y. He, A. C. Evans, Brain connectivity gender makes a difference, The Neuroscientist 17 (5) (2011) 575–591.

[89] C. Yan, G. Gong, J. Wang, D. Wang, D. Liu, C. Zhu, Z. J. Chen, A. Evans, Y. Zang, Y. He, Sex-and brain size–related small-world structural cortical networks in young adults: a DTI tractography study, Cerebral cortex 21 (2) (2011) 449–458.