# Data-driven methods for characterizing individual differences in brain MRI

by

Kuldeep KUMAR

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE
TECHNOLOGIE SUPÉRIEURE IN PARTIAL FULFILLMENT FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY
Ph.D.

MONTREAL, JUNE 11, 2018

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

# BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Prof. Christian Desrosiers, Ph.D, Thesis Supervisor
Software and IT Engineering Department at École de technologie supérieure

Prof. Jean Marc Lina, Ph.D, Chair, Board of Examiners
Electrical Engineering Department at École de technologie supérieure

Prof. Matthew Toews, Ph.D, Member of the jury
Automated Production Engineering Department at École de technologie supérieure

Prof. Maxime Descoteaux, Ph.D, External Evaluator
Département d'informatique, Faculté des Sciences at Université de Sherbrooke

THIS THESIS WAS PRESENTED AND DEFENDED

BEFORE A BOARD OF EXAMINERS AND PUBLIC

MAY 31 2018

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

# FOREWORD

This Ph.D. dissertation presents my research work carried out between 2013 and 2018 at École de technologie supérieure, under the supervision of professor Christian Desrosiers. The main objective of this research is to develop methods based on machine learning for human brain fingerprinting, an important step toward a fully-personalized analysis of brain characteristics.

This work resulted in a total of 3 journal papers, and 6 conference/workshop papers, published or under peer review, for which I am the first author. In addition, abstracts and posters based on this work were presented at leading venues. This dissertation focuses on the content of the three journal papers, presented in Chapters 2, 3 and 4. Other publications are listed in Appendix II. The Introduction section presents background information on human brain fingerprinting, as well as the main problem statement, motivations and objectives of this research. A review of relevant literature and key concepts follows in Chapter 1. After presenting the three journal papers (Chapters 2 to 4), Chapter 5 draws a brief summary of contributions and highlights some recommendations for further research.

# ACKNOWLEDGEMENTS

First and foremost, I would like to express my profound and sincere gratitude to my Ph.D. supervisor, Prof. Christian Desrosiers. It all started with a MITACS Globalink internship during summer 2011. The interaction during that brief internship was the main reason I returned for Ph.D at ÉTS. I received immeasurable guidance and support during my Ph.D. study at École de technologie superiéure (ÉTS). I feel fortunate to have worked with him for the past years. His vision and passion for research, his curiosity to details and intense commitment to his work, have all inspired me. During this important period in my career and life, he provided encouragement, sound advice, and fruitful ideas. He always encouraged me to expand my horizons, giving me several opportunities to attend international conferences and support to do an international internship.

I am grateful to Prof. Kaleem Siddiqi, for instilling many research practices during the Shape Analysis Course, and for constant help and support in building my course project into a fruitful research. He has been a great source of inspiration and learning for me. I am glad to have an opportunity to work with Prof. Matthew Toews, who has been like a second supervisor for my Ph.D. I am always inspired by his enthusiasm and excitement for research ideas. I learned a lot from great discussions on my project, his feedback and constant support. My sincere thanks also go to Prof. Olivier Colliot, who provided me an opportunity to visit their team as an intern, and provided precious support and constructive feedback on my research. I thank the great support from Pietro Gori, Benjamin Charlier and researchers at Aramis Lab.

I also want to acknowledge and thank the jury members, Prof. Jean Marc Lina, and Prof. Maxime Descoteaux, for accepting to review my thesis, and sharing meaningful and interesting comments with me.

The smooth completion of this project was made possible by the financial support from MITACS, through their Globalink Graduate Fellowship program and their Globalink Research

Internship Award-Inria. I also appreciate the funding support from the FRQNT-REPARTI International training scholarship, and kind support and help from Prof. Denis Laurendeau.

I would also convey my heartfelt thanks to researchers contributing to the Human Connectome Project (HCP) and ADNI projects. The datasets from HCP and ADNI were a great resource for my research. I am also grateful to Sherbrooke Connectivity Imaging Lab (SCIL) for providing a labeled data used in one of my projects.

Though only my name appears on the cover of this thesis, many partners and collaborators contributed to these works. I would like to acknowledge my colleagues that have graduated or are still at ÉTS: Atefeh, Mellie, Faten, Rémi, Xavier, Houda, Fariba, Jihen, Veronica, Edgar, Laura, Gerardo, Ruth, Helmi, Arlen, Pierre-Luc, Bashar, Ruben, Alpa, Erick, Otilia, and others. Many thanks go to friends and colleagues in LIVIA and other groups: Smarjeet Sharma, Laurent Chauvin, and Jose Dolz. I am grateful to Prof. Luc Duong for great suggestions and encouraging talks, and Prof. Sylvie Ratté for making the lab lively. A warm thanks to Samuel, and Patrice for the servers. My sincere gratitude also goes to many collaborators for helping me in both living and research during my Ph.D. studies.

I would also like to thank my colleagues at Space Applications Center, Ahmadabad. In particular, Santanu Chowdhury Sir, Meenakshi madam, Krishna Sir, and the AIPD. I am also grateful to the support and encouragement from my college friends, the KF's, and my friends from school.

Last but not the least, I would give my special thanks to my wife, Saleem Yadav, for the constant support and encouragement over the last year of my Ph.D. My grandparents and my brother Sandeep for their unconditional support. Finally, my sincere and heartfelt thanks to my parents Ranbir Singh and Sumitra Devi. You are the source of inspiration for me. *This thesis is dedicated to both of you.*

*“Success is the sum of small efforts,*

*repeated day in and day out”*

Robert Collier

# MÉTHODES BASÉES SUR LES DONNÉES POUR CARACTÉRISER LES DIFFÉRENCES INDIVIDUELLES DANS L'IRM CÉRÉBRALE

Kuldeep KUMAR

## RÉSUMÉ

Comprendre la structure et la fonction du cerveau humain est un problème de taille, essentiel au développement de traitements efficaces pour les maladies neurologiques comme la maladie d'Alzheimer et la maladie de Parkinson. Alors que la plupart des études font des inférences au niveau de groupes d'individus, les chercheurs ont établi que la structure et la fonction du cerveau varient d'un individu à un autre. Motivés par cette observation, plusieurs travaux récents se sont consacrés au développement de caractérisations compactes de cerveaux individuels, appelées empreintes du cerveau. Jusqu'à présent, ces études se sont principalement concentrées sur des empreintes à base d'une seule modalité, en particulier, utilisant la connectivité fonctionnelle. Cependant, plusieurs aspects importants n'ont pas été abordés dans ces études. Premièrement, le potentiel des empreintes cérébrales basées sur la connectivité structurelle n'a pas été exploré à fond. Ceci est en partie dû aux défis découlant des données de tractographie, y compris l'absence d'un étalon de comparaison et la variabilité des faisceaux de fibres. Deuxièmement, en raison des difficultés liées à la combinaison de plusieurs modalités dans un modèle unique, le développement d'une empreinte cérébrale multimodale demeure à ce jour un problème non-résolu. Cependant, puisque chaque modalité capture des propriétés uniques du cerveau, la combinaison de plusieurs modalités pourrait fournir une information plus riche et plus discriminante. Cette thèse aborde ces défis à travers trois contributions distinctes.

La première contribution consiste en des approches efficaces, basées sur l'apprentissage de dictionnaires par noyaux et les aprioris de parcimonie, pour la segmentation des fibres de la matière blanche et la caractérisation de leur variabilité inter-sujet. Le principe général des approches proposées est d'apprendre un dictionnaire compact de fibres capable de décrire l'ensemble des données, et d'encoder les faisceaux comme une combinaison parcimonieuse de plusieurs prototypes du dictionnaire. Ces approches permettent d'assigner des fibres à plus d'un faisceau, ce qui les rend mieux adaptées aux scénarios où les fibres ne sont pas clairement séparées, les faisceaux se chevauchent, ou lorsqu'il existe une variabilité inter-individuelle importante. De plus, ces approches n'exigent pas une représentation explicite des fibres et peuvent être adaptées à n'importe quelle représentation de fibres ou mesure de distance / similarité. Des expériences sur un jeu de données étiquetées et des données du Human Connectome Project (HCP) mettent en évidence la capacité de ces approches à regrouper les fibres en faisceaux plausibles, et illustrent les avantages de l'utilisation d'aprioris de parcimonie.

La deuxième contribution est une nouvelle empreinte cérébrale, appelée Fiberprint, la première à capturer la géométrie des fibres de matière blanche dans les individus. Cette empreinte utilise les approches d'apprentissage de dictionnaire provenant de la première contribution pour représenter les fibres de sujets dans un espace commun correspondant aux faisceaux proéminents. Des empreintes cérébrales compactes sont générées en appliquant une fonction de regroupement pour chaque faisceau, encodant les propriétés uniques des fibres telles que leur densité le long des faisceaux. Dans une analyse à grande échelle utilisant les données de 861 sujets du HCP, l'empreinte proposée est capable d'identifier des exemplaires du même individu ou des sujets génétiquement liés en utilisant un nombre limité de fibres.

Enfin, la troisième contribution de cette thèse est une première méthode automatique pour générer des empreintes cérébrales à partir de données multimodales. L'idée de base de cette méthode est de représenter chaque image comme un ensemble de caractéristiques locales, et d'utiliser ces caractéristiques multimodales pour représenter des sujets dans un sous-espace de basse dimension appelé variété. Des expériences utilisant l'IRM pondérée T1/T2, l'IRM de diffusion et l'IRM fonctionnelle au repos de 945 sujets du HCP démontrent l'avantage de combiner plusieurs modalités, les empreintes multimodales étant plus discriminantes que celles générées par les modalités individuelles. Les résultats mettent également en évidence le lien entre la similarité des empreintes et la proximité génétique, les jumeaux monozygotes ayant des empreintes plus similaires que les jumeaux dizygotes ou non-jumeaux.

Le travail présenté dans cette thèse peut être bénéfique à diverses études en neuroscience. Ainsi, les approches de segmentation présentées dans la thèse fournissent un moyen flexible et efficace pour analyser des courbes 3D comme celles provenant de la tractographie, et conviennent aux analyses à grande échelle de la connectivité structurelle. Fiberprint, qui est la première empreinte cérébrale caractérisant la géométrie des fibres de la matière blanche, offre une technique puissante pour explorer les différences individuelles en termes de connectivité de la matière blanche et leurs relations avec la génétique. En incluant des informations sur la microstructure le long de faisceaux, l'empreinte proposée pourrait également être utilisée pour définir de nouveaux biomarqueurs pouvant détecter et suivre la progression de maladies neurologiques comme la maladie de Parkinson. Enfin, l'empreinte cérébrale multimodale issue de cette recherche complémente les efforts en cours pour analyser les caractéristiques individuelles du cerveau en permettant de comparer et de contraster la contribution de différentes modalités d'imagerie. Elle peut ainsi mener à de nouvelles connaissances sur la variabilité de la structure et de la fonction du cerveau, ce qui pourrait aider au développement de stratégies de traitement personnalisées.

**Mots clés:** Empreintes cérébrales, Différences individuelles, IRM, Données de jumeaux, HCP, Représentation parcimonieuse, Apprentissage de dictionnaire, Méthodes à noyaux, Fibres de matière blanche, Imagerie multi-modale, Variété

# DATA-DRIVEN METHODS FOR CHARACTERIZING INDIVIDUAL DIFFERENCES IN BRAIN MRI

Kuldeep KUMAR

## ABSTRACT

Understanding the structure and function of the human brain is an outstanding problem that is critical to the development of efficient treatments for neurological diseases like Alzheimer's and Parkinson's. While most studies make group level inferences, researchers have established that structure and function show variability across individuals. Motivated by these, recent studies have focused on defining compact characterizations of individual brains, called brain fingerprints. So far, these studies have mostly focused on single modalities, with functional connectivity based fingerprints gaining considerable research interest. However, there are certain aspects which have not been addressed. First, the potential of fingerprints based on structural connectivity has not been fully explored. This is in part due to the challenges arising from fiber tracking data, including lack of gold standard and bundle variability. Second, due to the challenges of combining multiple modalities in a single framework, defining a multi-modal brain fingerprint remains to this day an elusive task. Yet, since each modality captures unique properties of the brain, combining multiple modalities could provide a richer, more discriminative information. This thesis addresses these challenges through three distinct contributions.

The first contribution consists of efficient approaches, based on kernel dictionary learning and sparsity priors, for segmenting white matter fibers and characterizing their inter-subject variability. The general principle of the proposed approaches is to learn a compact dictionary of training streamlines capable of describing the whole dataset, and to encode bundles as a sparse combination of multiple dictionary prototypes. These approaches allow streamlines to be assigned to more than one bundle, making them more suitable for scenarios where streamlines are not clearly separated, bundles overlap, or when there is important inter-individual variability. Additionally, they do not require an explicit representation of the streamlines and can extend to any streamline representation or distance/similarity measure. Experiments on a labeled set and data from HCP highlight the ability of our approaches to group streamlines into plausible bundles, and illustrate the benefits of employing sparsity priors.

The second contribution is a novel brain fingerprint, called Fiberprint, which is the first to capture white matter fiber geometry in individual subjects. This fingerprint leverages the sparse dictionary learning approaches of the first contribution to map streamlines of subjects to a common space representing prominent bundles. Compact fingerprints are generated by applying a pooling function for each bundle, encoding unique properties of streamlines such as their density along bundles. In a large-scale analysis using data from 861 HCP subjects, the proposed Fiberprint is shown capable of identifying exemplars from the same individual or genetically-related subjects, with only a small number of streamlines.

Lastly, the third contribution of this thesis is a first data-driven framework to generate brain fingerprints from multi-modal data. The key idea is to represent each image as a bag of local features, and use these multi-modal features to map subjects in a low-dimension subspace called manifold. Experiments using the T1/T2-weighted MRI, diffusion MRI, and resting state fMRI data of 945 HCP subjects demonstrate the benefit of combining multiple modalities, with multi-modal fingerprints more discriminative than those generated from individual modalities. Results also highlight the link between fingerprint similarity and genetic proximity, with monozygotic twins having more similar fingerprints than dizygotic or non-twin siblings.

The work described in this thesis can be of benefit to various neuroscience studies. The segmentation approaches presented in the thesis provide a flexible and efficient way to analyze 3D curves like tractography streamlines, and is suitable for large-scale analyses of structural connectivity. The proposed Fiberprint, which is the first brain fingerprint characterizing white matter fiber geometry, offers a powerful technique to explore individual differences in terms of white matter connectivity and its relationship to genetics. By including along-tract information on microstructure, it could also be used to define novel biomarkers for detecting and tracking the progression of neurological diseases like Parkinson's. Finally, the multi-modal brain fingerprint stemming from this research complements ongoing efforts to analyze individual brains characteristics by allowing to compare and contrast the contribution of different imaging modalities. It can thus lead to new insights on the variability of both brain structure and function, which could help the development of personalized treatment strategies.

**Keywords:** Brain fingerprinting, Individual differences, MRI, HCP twin data, Sparse coding, Kernel dictionary learning, White matter fibers, Multi-modal, Manifold

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

Page

# LIST OF ALGORITHMS

# LIST OF ABBREVIATIONS

| | |
|---|---|
| AD | Axial diffusivity |
| ADMM | Alternating direction method of multipliers |
| ADNI | Alzheimer’s disease neuroimaging initiative |
| ARI | Adjusted rand index |
| CC | Corpus callosum |
| CST | Corticospinal tract |
| dMRI | Diffusion MRI |
| DTI | Diffusion tensor imaging |
| DZ | Di zygotic |
| EP | End points |
| FA | Fractional anisotropy |
| fMRI | Functional magnetic resonance imaging |
| FS | Full siblings |
| GFA | Generalized fractional anisotropy |
| GQI | Generalized q-sampling imaging |
| HARDI | High angular resolution diffusion-weighted imaging |
| HCP | Human connectome project |
| IFO | Inferior fronto-occiptal fasciculus |
| MCP | Mean closest points |

| | |
|---|---|
| MD | Mean diffusivity |
| MDF | Minimum average direct-flip |
| MR | Magnetic resonance |
| MRI | Magnetic resonance imaging |
| MZ | Mono zygotic |
| NARI | Normalized adjusted rand index |
| NT | Not twin |
| PPMI | Parkinson’s progression markers initiative |
| RD | Radial diffusivity |
| rfMRI | Resting state functional magnetic resonance imaging |
| QA | Quantitative anisotropy |
| QB | Quickbundles |
| QBI | Q-ball imaging |
| QSDR | Q-space diffeomorphic reconstruction |
| RI | Rand index |
| SI | Silhouette |

# INTRODUCTION

> *“The greatest challenge to any thinker is stating the problem in a way that will allow a solution”*
>
> Bertrand Russell

The human brain is one of the most remarkable and complex structures known to man. Understanding the functional and structural properties of this organ is an outstanding research challenge that can have a significant impact on the advancement of science and the well-being of the population. Among many other benefits, a better understanding of the brain would be a major step toward the effective treatment of neuro-degenerative diseases like Alzheimer’s. According to the World Alzheimer Report 2016, such diseases present a global healthcare crisis, with an estimated 47 million people suffering dementia worldwide and costing over $800 billion per year.

Researchers have established that brain structure (Mangin *et al.*, 2004), function (Barch *et al.*, 2013; Gordon *et al.*, 2017a; Mueller *et al.*, 2013) and white matter architecture (Bürgel *et al.*, 2006; de Schotten *et al.*, 2011) show variability across individuals (Gordon *et al.*, 2017b). While most brain studies focus on group level inferences, recent advances in imaging hardware, as well as initiatives in precision medicine, have spawned considerable interest in the analysis of individual brains differences, a topic known as brain fingerprinting. Such analyses can help assess and optimize personalized treatment strategies, thereby reducing healthcare costs.

## 0.1 Problem statement and motivation

Fingerprinting studies are motivated by the fact that brain characteristics are largely determined by genetic factors that are often unique to individuals (Thompson *et al.*, 2013). Various neurological disorders like Parkinson’s (Geevarghese *et al.*, 2015) and Autism (Goldman *et al.*, 2013) have also been linked to specific brain abnormalities that are difficult to describe at the population level. With the rapid improvements in MRI acquisition hardware and analysis tools, and thanks to large brain-related initiatives like the Human Connectome Project (HCP) (Van Essen *et al.*, 2013) and UK Biobank (Sudlow *et al.*, 2015), researchers are better poised to study individual subjects in addition to groups (Dubois and Adolphs, 2016; Gordon *et al.*, 2017c), thus taking a step towards precision medicine (Hampel *et al.*, 2017) and precision psychiatry (Finn and Constable, 2016).

The importance of characterizing individual differences in brain MRI is evident from the recent surge in studies on this topic. For example, Yeh et al. (Yeh *et al.*, 2016a) built a local connectome fingerprint using dMRI data and applied this fingerprint to the analysis of genetically-related subjects. Finn et al. (Finn *et al.*, 2015) considered the correlation between time courses of atlas-defined nodes to generate a functional connectivity profile and used this profile to identify individuals across scan sessions, both for task and rest conditions. Liu et al. (Liu *et al.*, 2018) use dynamic brain connectivity patterns for identifying individuals and predicting higher cognitive functions. Moreover, Miranda et al. (Miranda-Dominguez *et al.*, 2014) proposed a linear model to describe the activity of brain regions in resting-state fMRI as a weighted sum of its functional neighboring regions. Their functional fingerprint, derived from the model's coefficients, has the ability to predict individuals using a limited number of non-sequential frames. Various morphometry-based fingerprints have also been proposed for structural MRI modalities like T1- or T2-weighted images. For example, Wachinger et al. (Wachinger *et al.*, 2015a) quantify the shape of cortical and subcortical structures via the spectrum of the Laplace-Beltrami operator. The resulting representation, called Brainprint, is used for subject identification and analyzing potential genetic influences on brain morphology.

So far, studies on characterizing individual differences have focused on single modalities, with functional connectivity based approaches gaining considerable research interest (Finn and Constable, 2016; Poldrack *et al.*, 2015; Gordon *et al.*, 2017c). While applications to understanding individual differences in behavior, personality, and clinical utility have been investigated (Dubois and Adolphs, 2016), various aspects of this research area remain uncharted. First, the potential of fingerprints based on structural connectivity has not been fully explored. This is in part due to the challenges arising from fiber tracking data, including lack of gold standard, bundle variability, differences in fiber length, and variability in fiber tracking output. Since white matter organization is an important factor underlying neurological function and disease, there is thus a critical need for novel tools characterizing the individual differences in structural connectivity.

Second, due to the complexity of combining multiple modalities in a single framework, defining a multi-modal brain fingerprint remains to this day an elusive task. Recent morphometry-based approaches, such as Brainprint (Wachinger *et al.*, 2015a), could potentially be extended to other modalities like diffusion MRI. However, this requires solving non-trivial problems such as the cross-modality alignment of images with different resolutions, the segmentation and correspondence of neuroanatomical structures, etc. Computational efficiency is another important issue when dealing with large-scale, multi-subject and multi-modal datasets like HCP. Nonetheless, since each modality captures unique properties of the brain, combining multiple modalities could provide a richer, more discriminative information.

## 0.2 Research objectives and contributions

Following the challenges and limitations highlighted above, the objective of this research is to **develop data-driven methods for characterizing individual differences in brain MRI**. Toward this goal, we propose a framework based on kernel sparse dictionary learning for the unsupervised clustering and segmentation of tractography streamlines. This framework enables to compare streamlines of different length and capture the complexity and variability of streamline bundles. We then use this framework to define a brain fingerprint, called Fiberprint, which characterizes white matter fiber geometry in individual subjects. Finally, we present a powerful approach using a bag-of-features image representation and manifold embedding to derive multi-modal brain fingerprints. The main contributions of this thesis can be further detailed as follows:

1) **Kernel sparse dictionary learning for white matter fiber segmentation**: Existing streamline clustering approaches assume crisp streamline-to-bundle membership. Due to lack of gold standard, variations in fiber tracking output, and different streamline lengths, this can be challenging in scenarios where streamlines are not clearly separated, bundles overlap, or when there is important inter-individual variability. Moreover, in various pattern recognition and neuroimaging applications, sparsity priors have shown robustness to variations due to noise or other factors. Based on this idea, the first contribution proposes novel frameworks based on kernel dictionary learning and sparsity priors for white matter fiber segmentation. The general idea of these approaches is to learn a compact dictionary of training streamlines capable of describing the whole dataset and to encode bundles as a sparse combination of multiple dictionary prototypes. Kernel dictionary learning allows capturing the non-linear relationship between streamlines and bundles, without the requirement for explicit embedding. The proposed methods allow streamlines to be assigned to more than one bundle, making it more suitable for above-mentioned scenarios. By exploiting group sparsity and manifold regularization, these methods provide robustness to the input number of clusters and allows incorporating anatomical constraints in the clustering.

   The findings related to this contribution are presented in the following papers:

   - **Kumar Kuldeep**, Siddiqi Kaleem, Desrosiers Christian. "White matter fiber analysis using kernel dictionary learning and sparsity priors". *Pattern Recognition*, Elsevier. *Submitted*.
   - **Kumar Kuldeep**, Gori Pietro, Charlier Benjamin, Durrleman Stanley, Colliot Olivier, Desrosiers Christian. "White matter fiber segmentation using functional varifolds". International Workshop on *Mathematical Foundations of Computational Anatomy, MICCAI 2017*. Graphs in Biomedical Image Analysis, Computational Anatomy and Imaging Genetics. Springer, Cham. 92-100.

- **Kumar Kuldeep**, Desrosiers Christian. "A sparse coding approach for the efficient representation and segmentation of white matter fibers". *IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, 2016. (pp. 915-919).
- **Kumar Kuldeep**, Desrosiers Christian, Siddiqi Kaleem. "Brain Fiber Clustering Using Non-negative Kernelized Matching Pursuit". International Workshop on *Machine Learning in Medical Imaging, MICCAI 2015*. (pp. 144-152). Springer, Cham.
- **Kumar Kuldeep**, Desrosiers Christian. "Group Sparse Kernelized Dictionary Learning for the Clustering of White Matter Fibers". Workshop on *Sparsity Techniques in Medical Imaging, MICCAI 2014*.

2) **Brain fingerprint modeling white matter fiber geometry**: Most fingerprint studies in the literature analyze individual brain differences from functional or structural MRI. While recent works have also investigated diffusion MRI for this task, they have done so at the level of single voxels, not brain fibers. Considering fibers instead of voxels provides additional information on white matter organization and structural connectivity. The second contribution of this thesis is a brain fingerprint called Fiberprint, which is the first to characterize white matter fiber geometry. To achieve this goal, we exploit the concept of feature pooling that plays a key role in numerous pattern recognition techniques like deep learning. Leveraging the sparse dictionary learning framework proposed in the first contribution, we map streamlines of subjects to a common space representing prominent bundles and generate compact fingerprints by applying a pooling function for each bundle. Obtained fingerprints encode unique properties of streamlines along dictionary bundles, such as their density. A large-scale analysis using data from 861 HCP subjects is conducted to measure the impact on the fingerprint of various parameters, including pooling function, dictionary size, sparsity and fiber tracking method/parameters. Furthermore, zygosity and siblingship information from the HCP dataset is used

to analyze the relationship between genetic proximity and fingerprint similarity across different bundles.

The second contribution resulted in the following paper and poster:

- **Kumar Kuldeep**, Desrosiers Christian, Siddiqi Kaleem, Colliot Olivier, Toews Matthew. "Fiberprint: A subject fingerprint based on sparse code pooling for white matter fiber analysis". *NeuroImage*. 158: 242-259. 2017.
- **Kumar Kuldeep**, Desrosiers Christian, Siddiqi Kaleem, Colliot Olivier, Toews Matthew. "Fiberprint: Identifying subjects and twins using fiber geometry based brain fingerprint". *Medical Imaging meets NIPS, NIPS 2017* (poster).

3) **Multi-modal brain fingerprint using bag-of-features and manifold approximation**: Current fingerprints encode brain characteristics from a single MRI modality. However, different modalities capture unique properties of the brain, and combining them can provide a richer information on individual differences. The third contribution of this thesis proposes a first data-driven framework, based on a bag-of-features representation and manifold approximation, to generate brain fingerprints from multi-modal data. In this framework, images are represented as a bag of local features, and manifold approximation is employed to map subjects in a common low-dimensional subspace. Experiments using the T1/T2-weighted MRI, diffusion MRI, and resting state fMRI data of 945 HCP subjects demonstrate the benefit of combining multiple modalities, and highlight the link between fingerprint similarity and genetic proximity.

The findings led to the following contributions:

- **Kumar Kuldeep**, Toews Matthew, Chauvin Laurent, Colliot Olivier, Desrosiers Christian. "Multi-modal brain fingerprinting: a manifold approximation based framework". *NeuroImage*, Elsevier. *Under review*
- **Kumar Kuldeep**, Chauvin Laurent, Toews Matthew, Colliot Olivier, Desrosiers Christian. "Multi-modal analysis of genetically-related subjects using SIFT de-

scriptors in brain MRI". Workshop on *Computational Diffusion MRI, MICCAI 2017*. Book Series: Mathematics and Visualization. (pp. 219-228). Springer.

- **Kumar Kuldeep**, Chauvin Laurent, Toews Matthew, Colliot Olivier, Desrosiers Christian. "Multi-modal brain fingerprinting: a bag of features and manifold approximation based twin analysis". Montreal Artificial Intelligence & Neuroscience (*MAIN 2017*, poster).
- **Kumar Kuldeep**, Chauvin Laurent, Toews Matthew, Colliot Olivier, Desrosiers Christian. "Analysis of genetically related subjects using multi-modal brain fingerprints". 1st Symposium on Applications of Artificial Intelligence in Medicine (*IAM 2018*). (**Best poster award**)

The full list of publications that resulted from this research can be found in **Appendix I**.

### 0.3 Thesis outline

The work presented in this thesis is organized as follows (Figure 0.1). In **Chapter 1**, we present basic concepts of magnetic resonance imaging and give a review of relevant works on the genetic basis of brain structure and function, individual-based approaches for neurological disorders, characterizing individual differences using brain MRI, and other concepts relevant to the thesis: streamline clustering, sparse coding in neuroimaging, and multi-modal studies in neuroimaging. **Chapter 2** then introduces the proposed streamline segmentation approaches, based on kernel dictionary learning and sparsity priors. The work presented in this chapter corresponds to the paper "White matter fiber analysis using kernel dictionary learning and sparsity priors", which was submitted to the *Pattern Recognition* journal. Following this, **Chapter 3** presents our Fiberprint framework that combines kernel dictionary learning and sparse code pooling to generate compact characterizations of white matter fiber geometry in individual subjects. This chapter corresponds to the paper entitled "Fiberprint: a subject fingerprint based on sparse code pooling for white matter fiber analysis", published in the *NeuroImage* journal. In

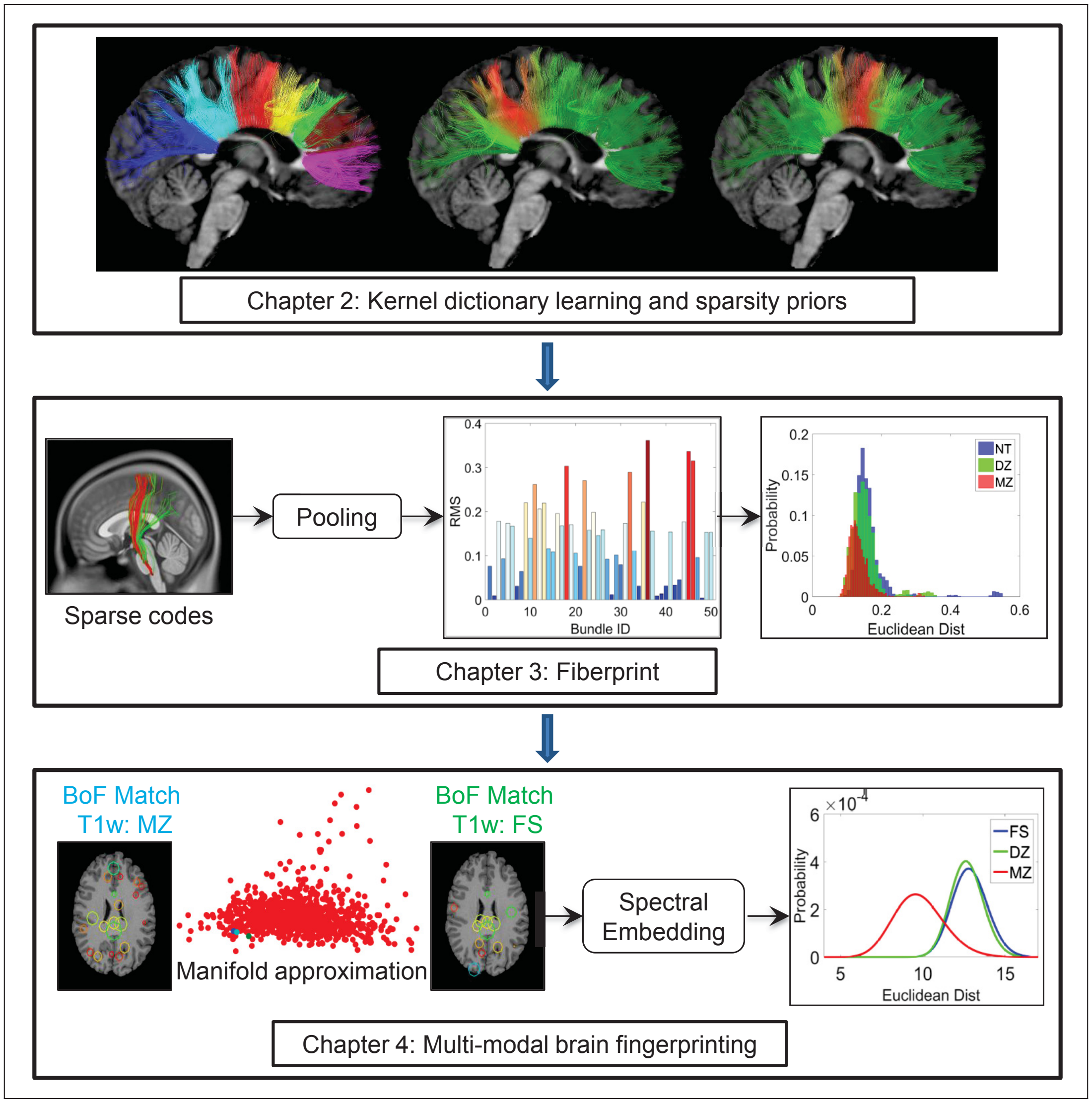


Figure 0.1 Sketch of the chapters of the thesis.

**Chapter 4**, we introduce our multi-modal brain fingerprinting framework, analyzing the contribution of various MRI modalities and their combinations in characterization of individual differences. The content of this Chapter corresponds to the paper "Multi-modal brain fingerprinting: a manifold approximation based framework", submitted to the *NeuroImage* journal. **Chapter 5** summarizes the main contributions of this dissertation and discusses its limitations as well as possible extensions. Finally, **Appendix I** provides a complete list of papers resulting from this Ph.D. study, and **Appendix II** provides link to codes.

# CHAPTER 1

# LITERATURE REVIEW

> *"If I have seen further it is by standing*
> *on the shoulders of giants"*
>
> Isaac Newton

## 1.1 Basics of the human brain

The human brain is one of the most complex structures known to man. It exerts centralized control over other organs of the body and is responsible for cognition, perception, emotion, thought, memory, and behavior. *It makes us who we are* (Seung, 2012).

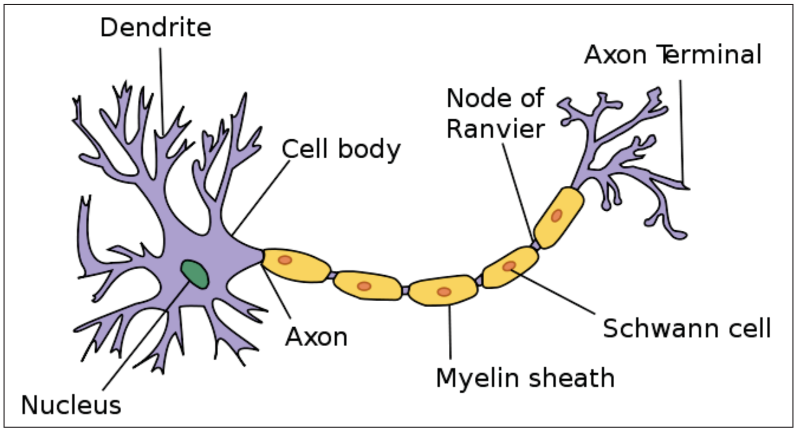


Figure 1.1 Diagram of a neuron (Source: https://en.wikipedia.org/wiki/Neuron)

The fundamental unit of the brain is the neuron (Fig. 1.1), a specialized cell that transmits nerve impulses and is vital to brain function. It is estimated that a typical healthy human brain contains around 100 billion neurons (Herculano-Houzel, 2009), linked to one another via trillions of tiny contacts called synapse. This intricate wiring of the brain is responsible for receiving and processing signals encoding the various functions of the brain.

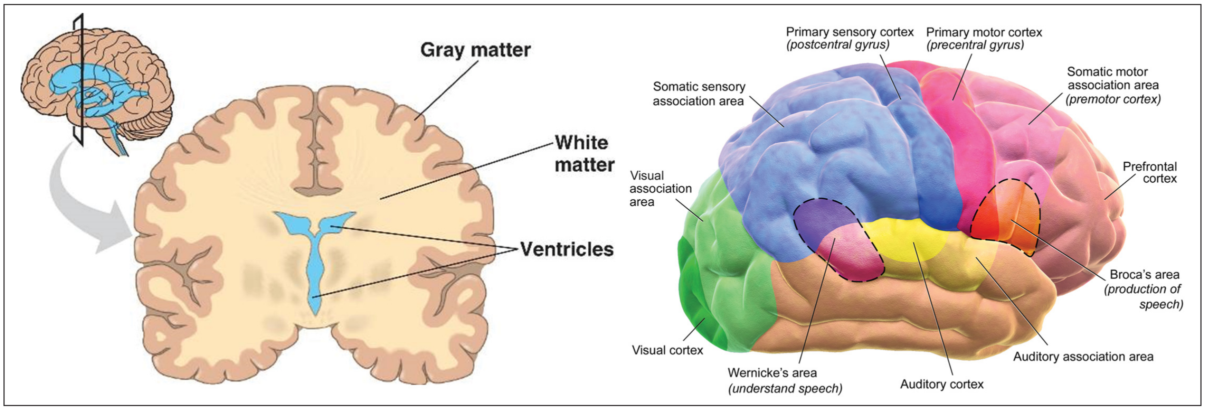


Figure 1.2 Left: Coronal slice of a healthy human brain with tissue types (Source: http://elearningbiology.weebly.com); Right: Functional areas of human brain (Source: https://en.wikipedia.org/wiki/Human_brain)

Anatomically, the human brain is made of three main tissue classes: gray matter, white matter, and cerebrospinal fluid (CSF, filling ventricles). Figure 1.2 (left) shows a typical coronal slice of human brain, highlighting tissue classes. Gray matter contains the cell bodies, dendrites, and axon terminals, and has a pinkish gray color in the living brain (hence the name). White matter is made of axons, which connect different parts of the brain to each other. Lastly, the CSF occupies the ventricular system around and inside the brain, acting as a cushion for the cortex.

The cerebral cortex of the brain can also be divided into different functional areas (Fig. 1.2, right). The frontal areas play an important role in reasoning, planning, language, memory, and motor control. The occipital lobe, rearmost part of the cortex, processes visual stimuli. Similarly, major functional areas and their functions can be inferred from the figure. A detailed description of functional areas can be found in (Damasio, 1995; Gray, 1878; Standring, 2015).

## 1.2 Imaging the brain: basics of MRI

As shown in Figure 1.3, the brain can be analyzed at different spatial and temporal scales (Frackowiak and Markram, 2015). Magnetic resonance imaging (MRI) provides a non-invasive approach to study the structure and function of the human brain at a macroscale, with a typical

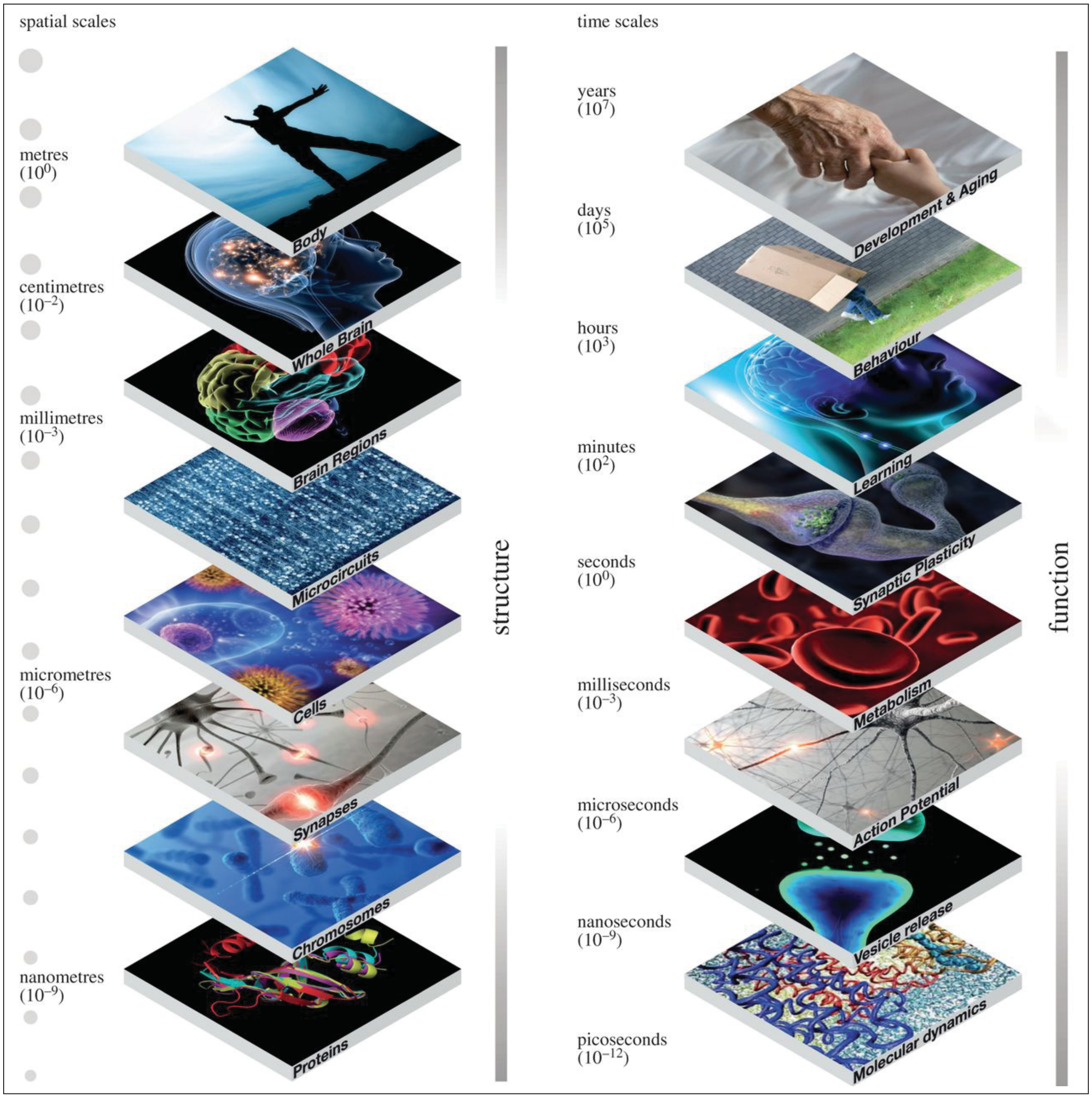


Figure 1.3 Illustration of brain as a multi-scale (spatial and temporal) organ. (Source: (Frackowiak and Markram, 2015))

spatial resolution of millimeters and temporal resolution of seconds (Lauterbur *et al.*, 1973). The underlying principle of MRI is nuclear magnetic resonance (NMR) (Bloch, 1946; Purcell *et al.*, 1946), a physical phenomenon in which nuclei (e.g., hydrogen atoms) in a magnetic field absorb and re-emit electromagnetic radiation. NMR-based imaging involves the alignment of the magnetic nuclear spins in a constant magnetic field and the perturbation of this alignment by an electromagnetic pulse (Fig. 1.4). After the withdrawal of the pulse, excited nuclei return to their equilibrium state and emit a radio-frequency (RF) signal. This signal is measured by a receiving coil and processed to form an image. The contrast between different tissues is determined by the rate at which excited nuclei return to the equilibrium state. A more detailed description of MRI acquisition and applications can be found in (Jacobs *et al.*, 2007; Pooley, 2005).

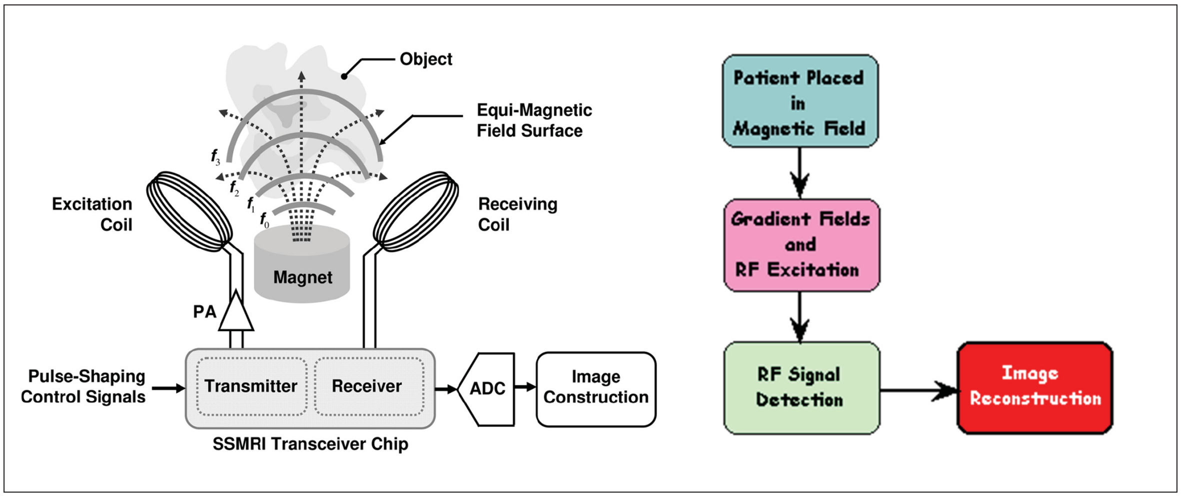


Figure 1.4 Illustration of the basic principle of MRI. (Source: (Source: (Hassibi *et al.*, 2009))

### 1.2.1 sMRI: imaging the anatomy

Structural MRI (sMRI) is used to obtain describe, both qualitatively and quantitatively, the shape, size, and integrity of gray and white matter structures in the brain. Different types of sMRI images, such as T1-weighted (T1w) and T2-weighted (T2w), can be generated by changing pulse sequence parameters corresponding to repetition time (TR) and echo time (TE). T1w images are obtained with a short TR and short TE, and emphasize the contrast between gray and white matter. Converesly, T2w images result from a long TR and long TE, and offer a high contrast between brain tissue and CSF. This type of images is often used to visualize fluids in brain tissues, for instance, resulting from cerebral edema.

sMRI modalities play a key role in standard processing pipelines for tasks like image registration, skull stripping and tissue segmentation. They are also essential to a wide range of applications, for instance, related to guided surgery, detecting and grading brain tumors, or the diagnosis and progression assessment of neuro-degenerative diseases.

### 1.2.2 dMRI: imaging white matter architecture

Diffusion magnetic resonance imaging (dMRI) (Basser *et al.*, 1994; Le Bihan and Breton, 1985; Le Bihan *et al.*, 1986; Pierpaoli and Basser, 1996) uses the diffusion of water molecules to generate contrast in MR images. It is based on the basic principle of Brownian motion of these molecules, in which their random motion is restricted upon encountering white matter fibers (tubular structures). Since directional information about diffusion is needed, dMRI commonly requires acquisitions along multiple directions.

DMRI is typically used for non-invasive inference of the underlying white matter structure at a macroscale. Compared to sMRI, it also provides information about the underlying microscopic structure and may indicate early pathologic changes (e.g., dMRI is more sensitive to early changes after a stroke than T1/T2-weighted MRI). The analysis of dMRI data often requires various processing steps (Garyfallidis *et al.*, 2014; Sotiropoulos *et al.*, 2013) including noise removal, signal reconstruction (Assemlal *et al.*, 2011; Ning *et al.*, 2015), tractography (Côté *et al.*, 2013; Neher *et al.*, 2015), parcellation-based connectome (Sotiropoulos and Zalesky, 2017), and streamline clustering (O'Donnell *et al.*, 2013). Additional details on dMRI and its analysis can be found in (Descoteaux, 2008, 2015; Hagmann *et al.*, 2006; Jbabdi *et al.*, 2015; Wedeen *et al.*, 2005).

### 1.2.3 fMRI: imaging brain activity

Functional magnetic resonance imaging (fMRI) is a non-invasive imaging technique which is widely used to probe brain function (Logothetis, 2008; Pike, 2012). It relies on the fact that cerebral blood flow and neuronal activation are coupled. Hence, the primary form of fMRI uses the blood-oxygen-level-dependent (BOLD) contrast (Ogawa *et al.*, 1990), and depicts changes in deoxyhemoglobin concentration consequent to task-induced or spontaneous modulation of neural metabolism. fMRI can localize activity to within millimeters and provides temporal resolution of the order of seconds.

There are two paradigms for studying the function of the brain: resting state fMRI (rfMRI) and task fMRI (tfMRI). Resting state fMRI is used to evaluate regional interactions that occur when a subject is not performing an explicit task (Biswal, 2012; Smith *et al.*, 2013a). It is a popular tool to explore functional organization and how this organization is altered in neurological diseases (Lee *et al.*, 2013). Conversely, task fMRI (Le Bihan *et al.*, 1995) infers brain activity when a subject performs a particular task compared to another moment when that task is not executed. Task fMRI serves for the study of cognitive behaviors related to motor, sensory, cognitive and emotional functions (Huettel, 2012; Liu, 2012; Barch *et al.*, 2013). A review of perspectives and applications of fMRI in cognitive neuroscience are provided in (Bandettini, 2012; Poldrack, 2012; Uğurbil, 2012).

## 1.3 Individual differences in neuroimaging

To date, most brain MRI studies have been concerned primarily with evincing population-level characteristics (Dubois and Adolphs, 2016). However, recent studies demonstrate that, despite having gross similarities, the brains of different individuals are unique (Barch *et al.*, 2013; Gordon *et al.*, 2017b; Seung, 2012). This has motivated researchers to explore the potential of building a science of individual differences (Dubois and Adolphs, 2016), and to address the specific challenges of interpreting inter-subject and intra-subject variability (Zilles and Amunts, 2013).

Numerous studies have established that individual differences in brain exist in terms of structure (Durrleman, 2010; Mangin *et al.*, 2004), function (Barch *et al.*, 2013; Gordon *et al.*, 2017a; Mueller *et al.*, 2013), and white matter architecture (Bürgel *et al.*, 2006; de Schotten *et al.*,

2011) (see Fig. 1.5). This discovery raises important questions, which are being explored by researchers and form the basis of this thesis:

- What is the source of these individual differences?
- Do individual differences impact neurological disorders?
- Can we understand these individual differences using brain MRI?

The following subsections present key research aiming at answering these questions, in particular, with respect to genetic basis of individual differences, individual differences in brain disorders, and brain fingerprinting.

### 1.3.1 Genetic basis of individual differences

Studies suggest that an individual's brain architecture is determined predominantly by genetic and environmental influences (Thompson *et al.*, 2013; Gu and Kanai, 2014), the main source of evidence coming from twin studies. Monozygotic (MZ) and dizygotic (DZ) twins share 100% and 50% of their genes, respectively, and typically live in the same environment. Phenotype differences between MZ twins thus reveal environmental or gene-environment interplay effects, while differences between DZ twins are also driven by genetics (Peper *et al.*, 2007).

While the specific genes involved in brain structure or function are largely unknown, heritability studies show that genetic effects vary regionally within the brain (Elliott *et al.*, 2017). For instance, the study in (Kochunov *et al.*, 2014) showed genetic effects on fractional anisotropy (FA), a measure of white matter microstructure. In (Kohannim *et al.*, 2012), FA is used to predict white matter integrity from multiple common genetic variants. Significant advances toward understanding the relationship between genetics and brain characteristics are expected with the creation of large initiatives like ENIGMA consortium (Thompson *et al.*, 2014) and UK Bio-bank (Sudlow *et al.*, 2015). These advancements will provide new insights into the genetic basis of individual differences and their impact on neurological disorders.

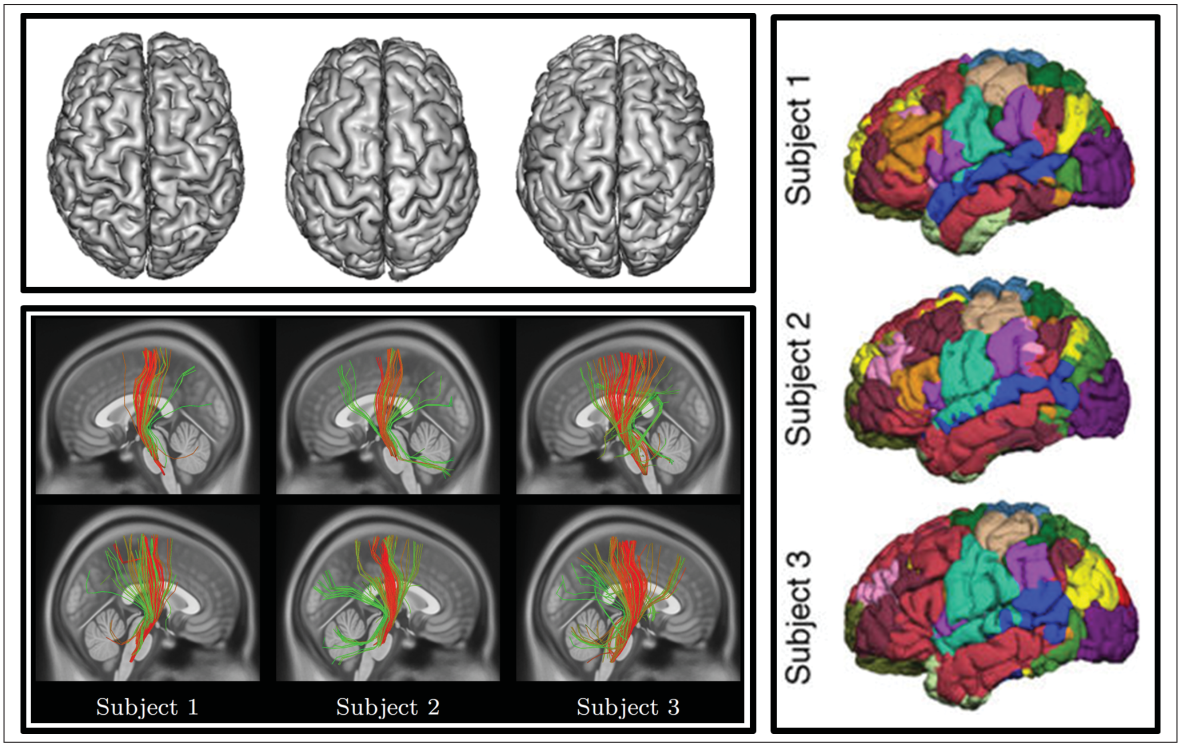


Figure 1.5 Illustration of individual differences in structure, function, and white matter architecture. Top left: folding patterns of the cortex surface (structure) (Source: (Durrleman, 2010)); Right: resting state fMRI based functional areas (Source: (Wang *et al.*, 2015)); Bottom left: corticospinal tract bundles (green and red represent, a low and a high membership of a streamline to the bundle, respectively.) (Source: (Kumar *et al.*, 2017c))

### 1.3.2 Individual differences in brain disorders

Various neurological disorders like Parkinson's (Geevarghese *et al.*, 2015) and autism (Goldman *et al.*, 2013) have been linked to specific brain abnormalities that are difficult to describe at the population level. Taking into account multiple factors that contribute to a specific disease, including genetic, biomarker, phenotypic and psycho-social characteristics, can therefore improve our understanding of disease onset and progression, as well as response to treatment. Moreover, considering an individual's specific makeup, instead of using a "one-size-fits-all" approach, can also optimize the effectiveness of disease prevention or treatment, and minimize side effects for patients less likely to respond to a particular therapeutic (Bu *et al.*, 2016; Reitz, 2016; Zou *et al.*, 2016).

The study in (Reitz, 2016) indicates that considering Alzheimer's as a homogeneous disease is one of the major reasons for the failure to identify effective treatments for this disease, and that better preventive or therapeutic interventions can be developed by employing a more personalized approach. Likewise, Bu et al. (Bu *et al.*, 2016) argue that personalized approaches based on multi-dimensional information can help achieve minimal side effects and maximal benefits in patients suffering from Parkinson's disease (PD). Neuroimaging studies on the single-subject prediction of brain disorders, including schizophrenia, depressive disorders, autism spectrum disease (ASD) and attention-deficit hyperactivity disorder (ADHD), highlight the need for individual-based approaches for better clinical diagnostic/prognostic adoption (Arbabshirani *et al.*, 2017; Calhoun *et al.*, 2017). Also, psychiatric disorders like bipolar disorder (BD) and major depression (MD) cannot be fully disambiguated at group level as they have considerable overlap in clinical observations (Frangou *et al.*, 2017).

### 1.3.3 Brain fingerprinting

The importance of quantifying and interpreting individual differences, for a fully-personalized investigation of brain structure and function, has been recognized for many years (Barch *et al.*, 2013; Bürgel *et al.*, 2006; de Schotten *et al.*, 2011; Mueller *et al.*, 2013; Mangin *et al.*, 2004). However, this has only been made possible recently with technological advances such as higher field strength, faster acquisition, and substantially improved resolution Dubois and Adolphs (2016); Glasser *et al.* (2013). With these improvements in MRI acquisition and analysis tools, and thanks to large brain-related initiatives like the Human Connectome Project (HCP) (Van Essen *et al.*, 2013) and UK Biobank (Sudlow *et al.*, 2015), researchers are better poised to study individual subjects (Gordon *et al.*, 2017c), thus taking a step towards fully-personalized investigations (Dubois and Adolphs, 2016; Hampel *et al.*, 2017). A critical aspect of such investigations is the development of techniques to characterize individual differences in the brain, called *brain fingerprints* (Fig. 1.6).

The recent surge in studies on brain fingerprinting reflects the high importance of this topic in neuroscience. Initial studies focused on building fingerprints based on functional MRI. For example, Finn et al. (Finn *et al.*, 2015) considered the correlation between time courses of functional regions (parcellation) to generate a functional connectivity profile, and used this profile to identify individuals across scan sessions, as well as between task and rest conditions. This study establishes that the individual variability in functional organization of brain is robust and can act as a fingerprint. This functional connectome fingerprint is also shown to predict levels of fluid intelligence. In (Liu *et al.*, 2018), Liu et al. investigate whether the variation in coupling among brain regions over time (i.e., the dynamic brain connectivity patterns) can characterize individual uniqueness. This characterization, termed as chronnectome fingerprinting, is used for identifying individuals and predicting higher cognitive functions. Similarly, Miranda-Dominguez et al. (Miranda-Dominguez *et al.*, 2014) propose a model-based approach toward characterizing resting state functional connectivity MRI in individual participants. A linear model is used to describe the activity of brain regions in resting-state fMRI as a weighted sum of its functional neighboring regions. Their functional fingerprint, derived from the model's coefficients, has the ability to predict individuals at a later date using a limited number of non-sequential frames.

Brain fingerprints using structural or diffusion MRI have also been proposed. For instance, Yeh et al. (Yeh *et al.*, 2016a,c) build a local connectome fingerprint using voxel-wise diffusion information from dMRI data, and apply this fingerprint to the analysis of genetically-related subjects and neuroplasticity. Moreover, Powell et al. (Powell *et al.*, 2018) show that the local connectome fingerprint derived phenotype maps could predict social, health, and cognitive factors. In (Wachinger *et al.*, 2015a), Wachinger et al. introduce a fingerprint called Brainprint for the discriminative representation of brain morphology from T1w and T2w MRI. Brainprint, which quantifies the shape of cortical and subcortical structures via the spectrum of the Laplace-Beltrami operator, is used for subject identification, age and sex prediction, brain asymmetry analysis, and analyzing potential genetic influences on brain morphology.

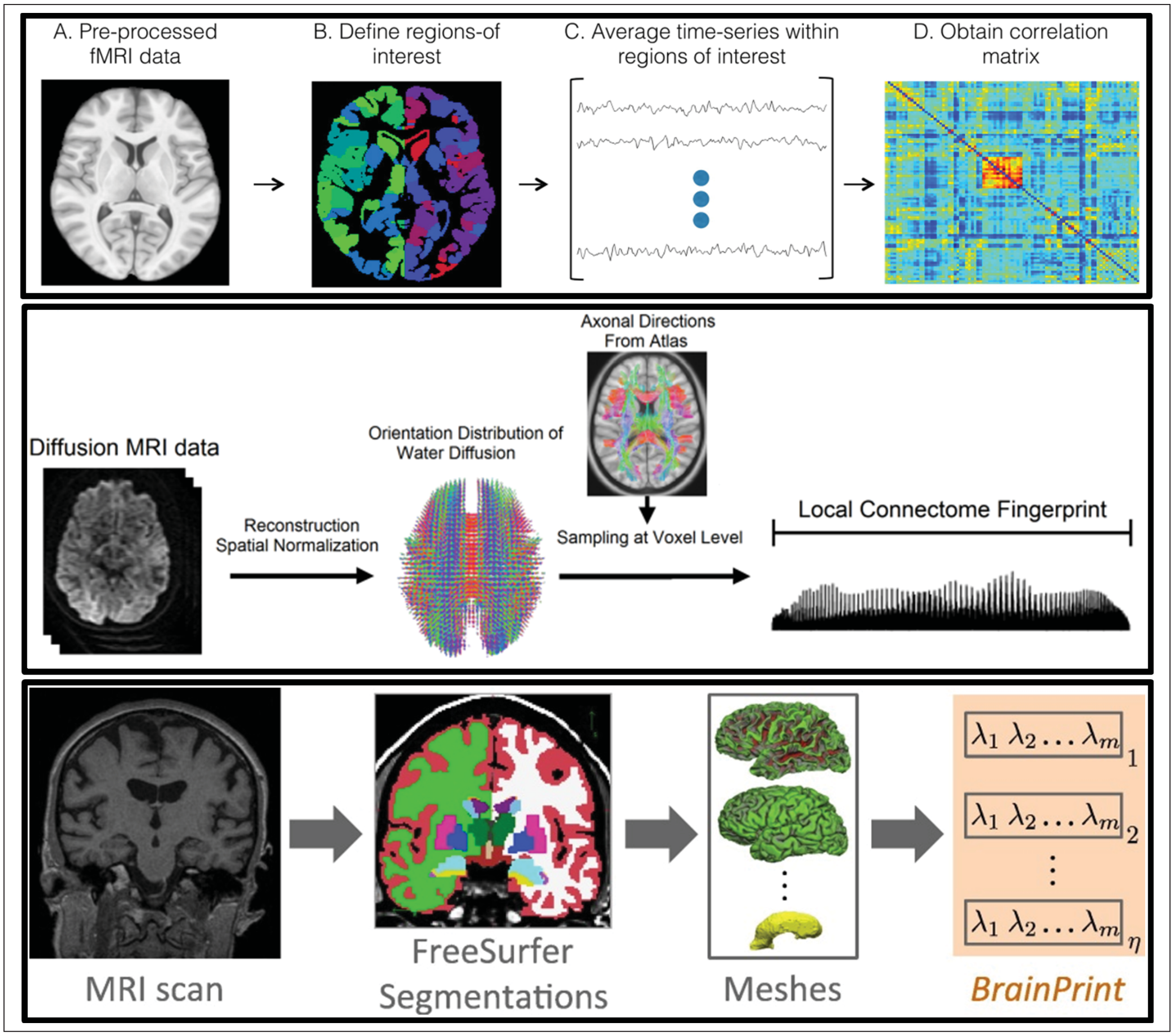


Figure 1.6 Overview of brain fingerprinting approaches. Top row: functional connectome fingerprint generation summary (Source: (Pedersen *et al.*, 2015)); Middle row: diffusion MRI based Local connectome fingerprint (Source: (Yeh *et al.*, 2016c)); and Bottom row: structural MRI based Brainprint (Source: (Wachinger *et al.*, 2015a))

While the field of brain fingerprinting is still in its infancy, researchers are investigating the potential and challenges of its application in a clinical setting. For example, Waller et al. (Waller *et al.*, 2017b) highlight the need for establishing the replicability, specificity, and generalizability of connectome fingerprints. Moreover, (Finn and Constable, 2016) argue that functional connectome fingerprint could help develop personalized approaches to psychiatric illness, and (Powell *et al.*, 2018; Liu *et al.*, 2018) utilize fingerprints for prediction of cognitive factors.

These studies underline that brain fingerprinting can lead to a better understanding of brain function and structure from a single subject's perspective.

## 1.4 Streamline clustering and analysis

White matter fiber tracts can be virtually reconstructed or traced throughout the brain as 3D space curves using tractography (Behrens *et al.*, 2007; Côté *et al.*, 2013; Daducci *et al.*, 2015; Neher *et al.*, 2015; Mori *et al.*, 1999). These 3D space curves, called streamlines, are generated by following most probable tract orientations at each voxel. Each streamline is an estimate of part of the course of underlying anatomical fiber tract, and has no direct correspondence with individual axons (Jones *et al.*, 2013). Despite this, tractography output provides a powerful tool to generate a macroscopic description of the white matter connections.

Recent advances in dMRI acquisition hardware and software have increased the spatial and angular resolution, yielding large tractography datasets of the order of thousands or millions of streamlines. These datasets are difficult to parse or interpret manually, and clustering approaches are often employed to group streamlines into anatomically meaningful and easier to visualize bundles. Clustering streamlines is also important for the creation of white matter atlases and the statistical analysis of microstructure measures along tracts (Guevara *et al.*, 2012; O'Donnell and Westin, 2007b; Siless *et al.*, 2018).

Over the years, several approaches have been proposed to cluster streamlines. These approaches can be roughly grouped into two categories: approaches requiring an explicit representation of streamlines and those based on streamline similarity/distance measures. Methods in the first category encode individual streamlines using a fixed set of features, for instance, the distribution parameters (mean and covariance) of points along the streamline (Brun *et al.*, 2004) or B-splines (Maddah *et al.*, 2006). Such method typically suffer from two problems: they are sensitive to the length and endpoint positions of the streamlines and/or are unable to capture their full shape. Instead of using explicit features, streamlines can also be compared using specialized distance measures, for example some function of the streamline coordinates

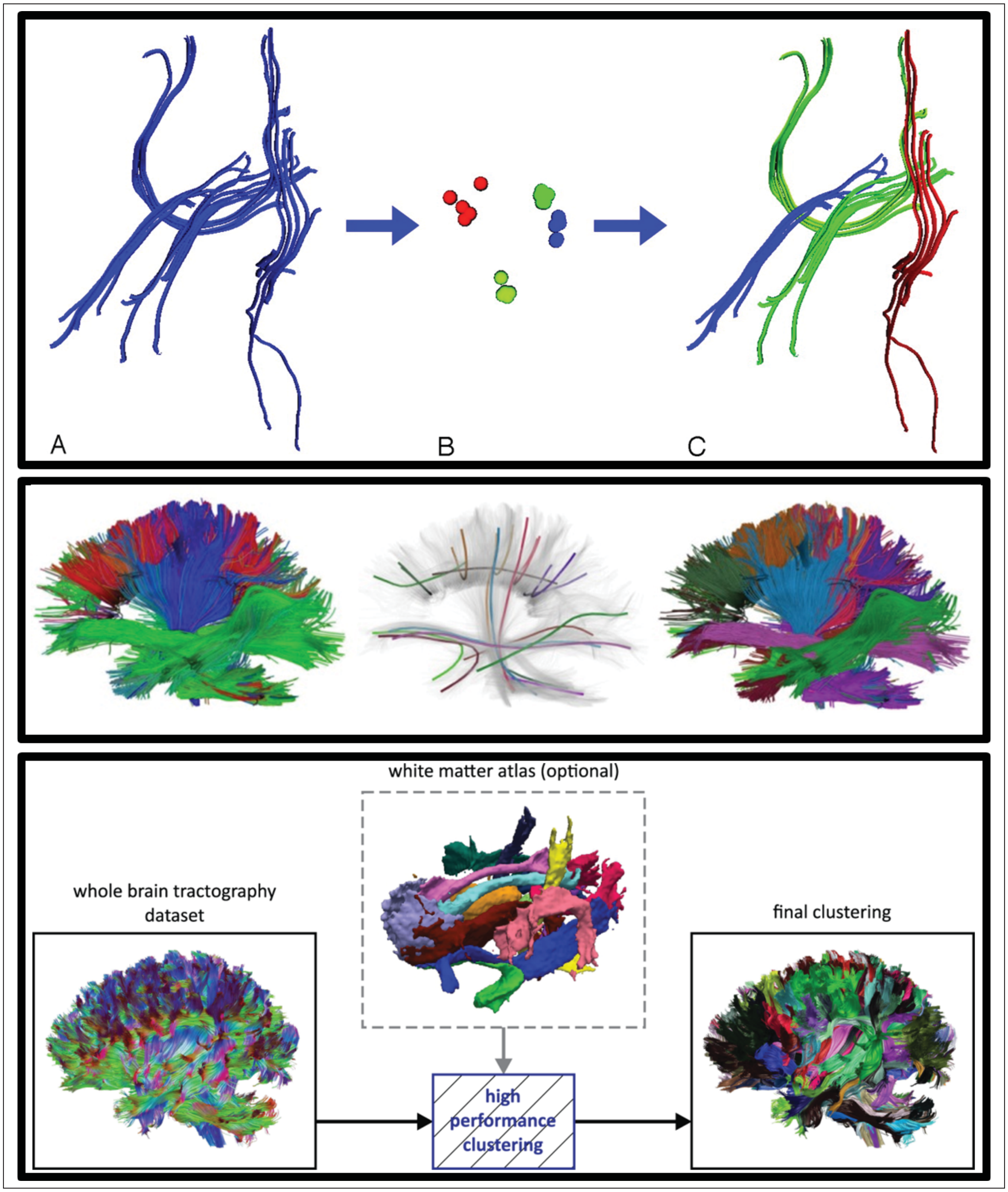


Figure 1.7 Illustrative examples of clustering approaches. Top row: Spectral embedding based streamline clustering (Source: O'Donnell *et al.* (2006)), Middle row: QuickBundles, bundle centroid estimation based approach (Source: (Garyfallidis *et al.*, 2012)). Bottom row: Atlas based approach (Source: (Ros *et al.*, 2013))

in Euclidean space. Popular distance measures for this task include the Hausdorff distance (Moberts *et al.*, 2005), the Minimum Direct Flip (MDF) distance (Garyfallidis *et al.*, 2012) and the Mean Closest Points (MCP) distance (Corouge *et al.*, 2004; Moberts *et al.*, 2005).

Streamline clustering algorithms include manifold embedding techniques such as spectral clustering (O'Donnell and Westin, 2007a) and normalized cuts (Brun *et al.*, 2004), agglomerative approaches like hierarchical clustering (Moberts *et al.*, 2005), k-means (Li *et al.*, 2010), and Dirichlet processes (Wassermann *et al.*, 2010; Wang *et al.*, 2011c). Recently, studies have also focused on incorporating anatomical features into the clustering (Siless *et al.*, 2018; Wassermann *et al.*, 2016), or on clustering large multi-subject datasets (Jin *et al.*, 2014; Prasad *et al.*, 2014). A detailed description and comparison of several distances and clustering approaches can be found in (Moberts *et al.*, 2005; Olivetti *et al.*, 2017; Siless *et al.*, 2013).

Various studies have also focused on the streamlines segmentation for drawing cross-population inferences (Guevara *et al.*, 2012; Jin *et al.*, 2014; Prasad *et al.*, 2014; Zhang *et al.*, 2018). Most of these studies either follow an atlas-based approach (Guevara *et al.*, 2012; Jin *et al.*, 2014; O'Donnell and Westin, 2007a; Ros *et al.*, 2013) or align specific tracts directly across subjects (Garyfallidis *et al.*, 2015; O'Donnell *et al.*, 2012). Multi-step or multi-level approaches have also been proposed to segment streamlines, for example, by combining both voxel and streamline groupings (Guevara *et al.*, 2012), fusing labels from multiple hand-labeled atlases (Jin *et al.*, 2014), or using a bundle representation based on maximum density paths (Prasad *et al.*, 2014). A few studies have also investigated the representation of specific streamline bundles using different techniques such as gamma mixture models (Maddah *et al.*, 2008), the computational model of rectifiable currents (Durrleman *et al.*, 2009; Gori *et al.*, 2016), and functional varifolds (Kumar *et al.*, 2017d). Figure 1.7 provides an illustration of three different clustering approaches including spectral embedding (top row), a greedy approach based on MDF distance and bundle centroid estimation (middle row), and an atlas based approach (last row). For a detailed review of white matter clustering approaches, we refer the reader to (O'Donnell *et al.*, 2013).

Despite the many works on this topic, streamline clustering and segmentation are still open problems. Thus, clustering streamlines into anatomically meaningful bundles is challenging due to lack of a gold standard. As shown in Figure 1.7 (top row), even for a single subject, streamlines within the same bundle can have different lengths and endpoints. Thus, using standard geometric distance measures can often lead to poor results. Another challenge comes from the weak separability of certain bundles, which can result in low-quality (e.g., too small or too large) clusters (Fig. 1.7, middle and bottom row). Streamline bundles may also overlap and intersect each other, making their extraction and analysis difficult. Finally, since there can be up to millions of streamlines to consider, clustering these streamlines is computationally complex.

## 1.5 Methodological concepts explored in the thesis

In this section we explore the basics of concepts explored in this thesis, including, sparse coding, multi-modal analysis of brain MRI, manifold embedding and bag of feature based image representation.

### 1.5.1 Sparse coding

Sparse coding is a well-known technique for encoding and analyzing signals like images. It has been used in numerous image processing and pattern recognition applications, including compression, denoising, deblurring, inpainting, and super-resolution and classification (Elad *et al.*, 2010; Wright *et al.*, 2009, 2010; Yang *et al.*, 2009; Rubinstein *et al.*, 2010). The basic principle of sparse coding is to learn a basis, called dictionary, which can effectively represent examples from a given dataset using only a few basis elements (Fig. 1.8). Various priors can be employed to impose sparsity when encoding examples using the dictionary (Fig. 1.9). Minimizing the $L_0$ norm of the encoding limits the number of non-zero elements in this encoding, however this results in a non-convex optimization problem. A popular alternative is to use the $L_1$ norm as sparsity prior, defined as the sum of absolute encoding coefficients.

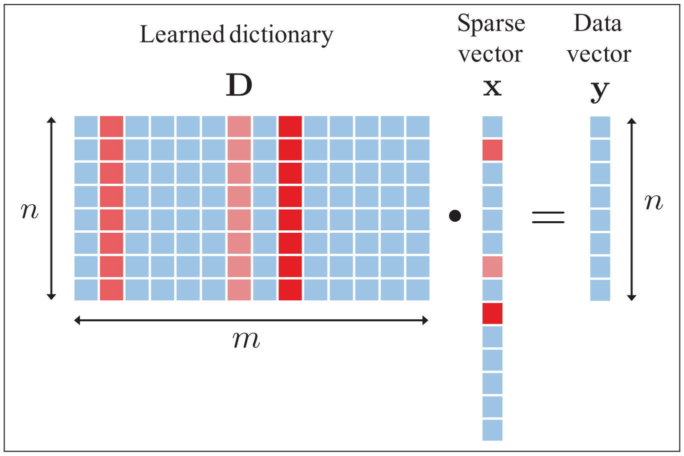


Figure 1.8 Illustrative example of sparse coding.

Dictionary learning and sparse coding have been investigated for various neuroimaging applications, for instance, compressed sensing MRI (CS-MRI) (Lustig *et al.*, 2008) or segmentation of MRI data (Tong *et al.*, 2013). In functional connectivity analysis, Lee et al. (Lee *et al.*, 2016a) proposed a sparsity based analysis of k-hubness for overlapping network structures. For diffusion MRI data, sparse coding has been used successfully for clustering white matter voxels from Orientation Density Function (ODF) data (Çetingül *et al.*, 2014), for finding a population-level dictionary of key white matter tracts (Zhu *et al.*, 2016), for higher-order tensor (HOT) based diffusion MRI reconstruction (Feng *et al.*, 2015), and for denoising and reconstruction for diffusion spectrum imaging (Aranda *et al.*, 2015; Bilgic *et al.*, 2013; Gramfort *et al.*, 2014; Merlet *et al.*, 2013). Moreover, Daducci et al. (Daducci *et al.*, 2014) showed that using $L_0$ norm priors improves the reconstruction of fiber orientation distribution functions (ODF). Likewise, Auria et al. (Auría *et al.*, 2015) use voxel-wise sparsity regularization to obtain a more accurate reconstruction of fiber orientation distribution functions (FOD).

### 1.5.2 Manifold embedding

Manifold learning is an approach to non-linear dimensionality reduction (Tenenbaum *et al.*, 2000) based on the assumption that the dimensionality of many datasets is only artificially high. Techniques based on this principle, like Isomap (Tenenbaum *et al.*, 2000), Locally Linear

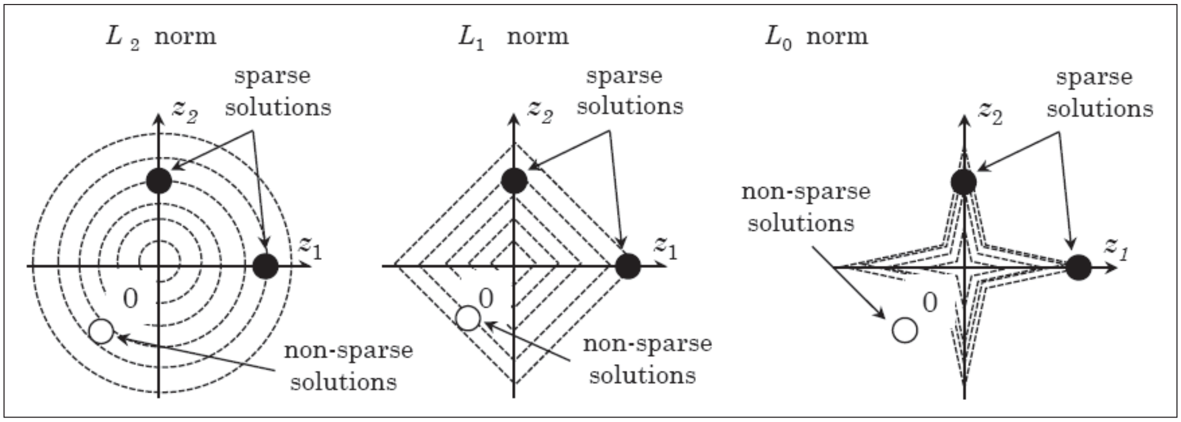


Figure 1.9 Illustrative example of $L_0$, $L_1$ and $L_2$ norms in the 2D plane. Source: (Kudo *et al.*, 2013).

Embedding (LLE) (Roweis and Saul, 2000) and Spectral Embedding (SE) (Belkin and Niyogi, 2003), aim to extract a low-dimensional manifold that can best describe the high-dimensional data (Bengio *et al.*, 2013). Each of these techniques preserve certain properties of data, for example Isomap maintains geodesic distance between all points, LLE preserves distances within local neighborhoods, and SE ensures that points close to each other on the manifold are mapped close to each other in the low dimensional space.

Manifold learning has also played a crucial role in medical imaging studies (Aljabar *et al.*, 2012), with applications in a wide variety of problems including registration (Ye *et al.*, 2012), segmentation (Li *et al.*, 2015), and classification (Ye *et al.*, 2014). For example, Gerber et al. (Gerber *et al.*, 2010) use manifold learning to perform a population analysis of brain images. Likewise, Brosch et al. (Brosch *et al.*, 2013) explore a deep learning based approach to learn the manifold of brain MRIs. Finally, Aljabar et al. performed a morphological analysis of brain MRI using spectral methods (Aljabar *et al.*, 2008), with application to neonatal MRI (Aljabar *et al.*, 2010).

### 1.5.3 Bag of features

Representations based on local features, often referred to as bag of features (BoF), offer an efficient alternative for encoding and matching image structures without the stringent requirement of one-to-one correspondence (Lowe, 2004, 1999). This technique is inspired by how human visual cortex works, where an object can be recognized using only a few salients points on this object (i.e., the local features). A popular algorithm for detecting and encoding local features is the Scale-Invariant Feature Transform (SIFT) (Lowe, 2004, 1999), which has the ability to handle intensity, rotation, scale and affine variations (Mikolajczyk and Schmid, 2005). BoFs have been used in a wide range of computer vision problems, including scene classification (Yang *et al.*, 2007), object categorization, semantic video retrieval (Jiang *et al.*, 2007), and image annotation (Tsai, 2012).

In brain imaging, BoFs have been shown to automatically identify known structural differences between healthy controls and Alzheimer's subjects in a fully data-driven fashion (Toews *et al.*, 2010). They have also been used successfully to model the development of infant brains (Toews *et al.*, 2012) and align images of different modalities (Toews and Wells, 2013). Despite their numerous advantages, BoFs have thus far not been explored for brain fingerprinting. This is mainly due to their large and variable size, which makes comparing two fingerprints non-trivial.

## 1.6 Summary

While there are multiple facets and fascinating opportunities, this thesis focuses on individual differences. As asserted before, “we are our connectome” (Seung, 2012). If so, how do we differ? Where do these differences come from? What is the impact of these differences? How can we study them? These questions form the motivation and basis of this research.

### 1.6.1 Studies on individual differences: potential and utility

As highlighted in (Dubois and Adolphs, 2016), there is growing interest in the interpretation of fMRI data at the level of individual brains. For example, Dosenbach et al. (Dosenbach *et al.*, 2010) predict individual brain maturity using fMRI. Individual differences in relation to personality (Yarkoni, 2015), intelligence (Finn *et al.*, 2015; van den Heuvel *et al.*, 2009), and mood (Smith *et al.*, 2015) have also been investigated in the literature. Similarly, the comprehensive review of Arbabshirani et al. (Arbabshirani *et al.*, 2017) highlights the potential of neuroimaging data for single subject prediction of brain disorders including schizophrenia, depressive disorders, autism spectrum disease (ASD), and attention-deficit hyperactivity disorder (ADHD). Personalized approaches have also been advocated for Alzheimer's (Hampel *et al.*, 2017; Reitz, 2016) and Parkinson's (Bu *et al.*, 2016).

One of the recent developments in the study of individual differences is brain fingerprinting. While studies on this topic are still in infancy, the potential of brain fingerprinting for different applications has been explored in the literature (see Table 1.1). For example, Finn et al. (Finn and Constable, 2016) argue that functional connectome fingerprint could help develop personalized approaches to psychiatric illness, Powell et al. (Powell *et al.*, 2018) show that local connectome phenotypes can predict social, health, and cognitive factors, and Liu et al. (Liu *et al.*, 2018) report that dynamic brain connectivity patterns can predict individual higher cognitive performance (e.g., fluid intelligence and executive function).

### 1.6.2 Brain fingerprinting: challenges and opportunities

The growing interest in brain fingerprinting highlights various challenges and opportunities that need to be explored. For example, the study in (Chamberland *et al.*, 2017), which explores the role of white matter architecture in the origin of individual differences in functional connectivity, suggests that further research is needed to understand the role of anatomical pathways in supporting vascular-based measures of functional connectivity. In studying long term

Table 1.1 Summary of brain fingerprinting studies

| Study | Modalities | Fingerprint |
|---|---|---|
| Finn *et al.* (2015) | rfMRI & tfMRI | Functional connectome fingerprint: defined using correlation between time courses of functional regions (parcellation) |
| Miranda-Dominguez *et al.* (2014) | rfMRI | Connectotyping: model's coefficients describing the activity of brain regions as a weighted sum of its functional neighboring regions |
| Liu *et al.* (2018) | rfMRI | Chronnectome fingerprinting: dynamic network analysis of functional connectivity |
| Yeh *et al.* (2016c) | dMRI | Local connectome fingerprint: a histogram encoding voxel-wise density of diffusing water along a set of atlas defined directions in white matter |
| Wachinger *et al.* (2015a) | T1w | Brainprint: the shape of cortical and subcortical structures quantified via the spectrum of the Laplace-Beltrami operator (shape-DNA) |

neural and physiological phenotyping of a single human, Poldrack et al. (Poldrack *et al.*, 2015) motivate the necessity of larger efforts to characterize psychological and brain function longitudinally. Likewise, Wang et al. (Wang *et al.*, 2015) highlight the need for cortical parcellation approaches that can accurately map functional organization at the level of individuals, and Laumann et al. (Laumann *et al.*, 2015) underline the importance of reproducibility and validity of single subject areal parcellation. Finally, evaluating multi-site reliability of functional connectivity, Noble et al. (Noble *et al.*, 2017a) claims that aggregation of data across longer scan durations is necessary to increase the reliability of functional connectivity estimates at the single-subject level.

Other aspects of fingerprints that require further investigation include improving the interpretability of existing fingerprints (Poldrack *et al.*, 2015; Vanderwal *et al.*, 2017), establishing the influence of genetics and environment factors (Miranda-Domínguez *et al.*, 2017), observing changes across life-span (Brown, 2017; Chan *et al.*, 2017) and neurocognitive changes during adolescence (Foulkes and Blakemore, 2018), as well as addressing the challenges on the applicability of fingerprints (Finn and Constable, 2016; Horien *et al.*, 2018; Noble *et al.*, 2017b; Shen *et al.*, 2017; Waller *et al.*, 2017b).

# CHAPTER 2

# WHITE MATTER FIBER ANALYSIS USING KERNEL DICTIONARY LEARNING AND SPARSITY PRIORS

Kuldeep Kumar[a,*], Kaleem Siddiqi[b], Christian Desrosiers[a]
[a]LIVIA, École de technologie supérieure, Montreal, Canada
[b]CIM, McGill University, Montreal, Canada
Email: kkumar@livia.etsmtl.ca, siddiqi@cim.mcgill.ca, christian.desrosiers@etsmtl.ca

This article was submitted to *Pattern Recognition*, Elsevier, on March 31, 2018

## 2.1 Abstract

Diffusion magnetic resonance imaging, a non-invasive tool to infer white matter fiber connections, produces a large number of streamlines containing a wealth of information on structural connectivity. The size of these tractography outputs makes further analyses complex, creating a need for methods to group streamlines into meaningful bundles. In this work, we address this by proposing a set of kernel dictionary learning and sparsity priors based methods. Proposed frameworks include $L_0$ norm, group sparsity, as well as manifold regularization prior. The proposed methods allow streamlines to be assigned to more than one bundle, making it more robust to overlapping bundles and inter-subject variations. We evaluate the performance of our method on a labeled set and data from Human Connectome Project. Results highlight the ability of our method to group streamlines into plausible bundles and illustrate the impact of sparsity priors on the performance of the proposed methods.

## 2.2 Introduction

Since its development in the 1980s, diffusion tensor imaging (DTI) has become an essential tool to study white matter connectivity in the human brain. Its ability to infer the orientation of white matter fibers, in-vivo and non-invasively, is key to understanding brain connectivity and associated neurological diseases (Hagmann *et al.*, 2006; de Schotten *et al.*, 2011). Since the macroscopic inference of underlying fibers from dMRI data, known as tractography, typically produces a large number of streamlines, it is common to group these streamlines into anatomically meaningful clusters called *bundles* (O'Donnell *et al.*, 2013). Clustering streamlines is also essential for the creation of white matter atlases, visualization, and statistical analysis of microstructure measures along tracts (Guevara *et al.*, 2012; O'Donnell and Westin, 2007b; Siless *et al.*, 2018). Furthermore, clinical applications of tractography analysis are also numerous and include identifying major bundles for neurological planning in patients with tumors (O'Donnell *et al.*, 2017), understanding difference between white matter connectivity in typically developing controls versus children with autism (Zhang *et al.*, 2017), and uncovering white matter bundles as bio-markers for the diagnosis of Parkinson's disease (Cousineau *et al.*, 2017).

Clustering streamlines into anatomically meaningful bundles is a challenging task in part due to lack of gold standard. There can be several hundreds of thousands of streamlines to consider, making the clustering problem computationally complex. As illustrated in Fig. 2.1, streamlines within the same bundle can have different lengths and endpoints. Thus, using standard geometric distance measures often leads to poor results. Another challenge comes from the weak separability of certain bundles, which can result in low-quality (e.g., too small or too large) clusters. Also, while many clustering approaches assume a crisp membership of streamlines to bundles, as shown in Fig. 2.1, such a separation of streamlines into hard clusters is often arbitrary. In practice, streamline bundles may overlap and intersect each other, making their extraction and analysis difficult. Moreover, when used to label the streamlines of a new subject, the clusters learned using crisp methods often give unsatisfactory results due to the variability across individual brains.

In this paper, we propose a set of flexible and efficient streamline clustering approaches based on kernel dictionary learning and sparsity priors. The general idea of these approaches is to learn a compact dictionary of training streamlines capable of describing the whole dataset, and

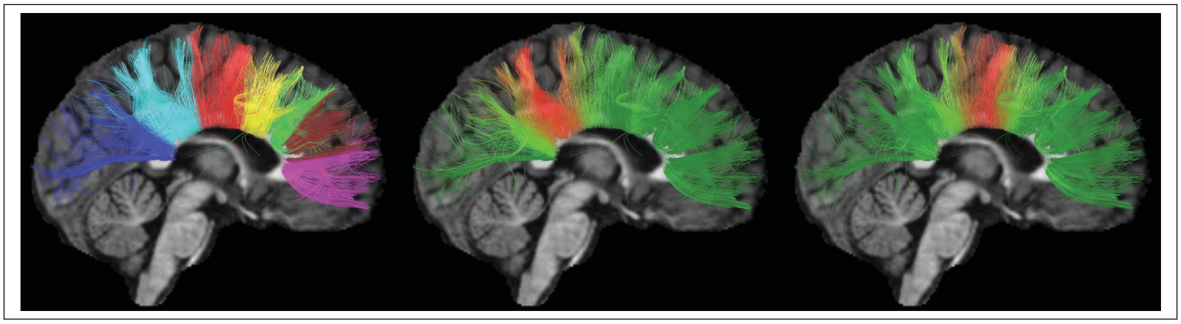


Figure 2.1 Illustrative example. Clustering of the corpus callosum by our method: hard clustering (**left**), and membership of each streamline to two bundles (**center** and **right**). Dark green represents a zero membership and bright red a maximum membership to the bundles.

to encode bundles as a sparse non-negative combination of multiple dictionary prototypes. In contrast to spectral embedding methods (e.g., (Brun *et al.*, 2004; O'Donnell and Westin, 2005)) which perform the embedding and clustering in two separate steps, our approaches find clusters in the kernel space without having to explicitly compute an embedding.

The proposed streamline clustering approaches have several advantages over existing methods for this task. First, they do not require an explicit representation of the streamlines and can extend to any streamline representation or distance/similarity measure. Second, they use a non-linear kernel mapping which facilitates the separation of clusters in a manifold space. Third, unlike hard-clustering methods like the k-means algorithm and its variants (e.g. spectral clustering), they can distribute the membership of streamlines across multiple bundles, making them more robust to overlapping bundles and outliers, as well as to variability across subjects.

Our specific contributions include:

a. We propose three different streamline clustering models based on kernel k-means, non-negative factorization and sparse coding, and demonstrate the advantages of these models with respect to the state of the art;

b. We provide a flexible platform to integrate and evaluate streamline distance measures, and compare the performance of three popular measures using two different datasets;

c. Whereas dictionary learning and sparsity have shown promise in various pattern recognition and neuroimaging applications, to our knowledge, the present article is the first account of their use for streamline clustering in a peer-reviewed indexed publication. Our results on the streamline clustering problem show the potential of this approach for other imaging applications.

The rest of the paper is structured as follows. Section 2.3 provides a brief survey of relevant literature on streamline clustering. In Section 2.4, we present our kernel dictionary learning based methods. Section 2.5 evaluates the methods on the task of clustering streamlines using real data. Finally, we conclude with a summary of our main contributions, and discuss potential extensions.

## 2.3 Related works

Our presentation of relevant work is divided into two parts, focusing respectively on the various approaches for representation and analysis of streamlines, and the application of sparse coding techniques in neuroimaging.

### 2.3.1 White matter fiber analysis

Over the years, several approaches have been proposed to cluster streamlines and provide a simplified quantitative description of white matter connections, including cross-population inferences (Guevara *et al.*, 2012; Jin *et al.*, 2014; O'Donnell and Westin, 2007a; Prasad *et al.*, 2014). These studies could be vaguely classified into two categories: representation of streamlines or streamline similarity, and clustering approaches. Features proposed to represent streamlines include the distribution parameters (mean and covariance) of points along the streamline (Brun *et al.*, 2004) and B-splines (Maddah *et al.*, 2006). Approaches using such explicit features typically suffer from two problems: they are sensitive to the length and endpoint positions of the streamlines and/or are unable to capture their full shape. Instead of using explicit features, streamlines can also be compared using specialized distance measures. Popular distance measures for this task include the Hausdorff distance, the Minimum Direct Flip (MDF) distance and the Mean Closest Points (MCP) distance (Corouge *et al.*, 2004; Moberts *et al.*, 2005).

Fiber clustering approaches include manifold embedding techniques such as spectral clustering and normalized cuts (Brun *et al.*, 2004), agglomerative approaches like hierarchical clustering (O'Donnell and Westin, 2007b; Corouge *et al.*, 2004), k-means (Li *et al.*, 2010), and Dirichlet processes (Wassermann *et al.*, 2010; Wang *et al.*, 2011c). Several studies have also focused on incorporating anatomical features into the clustering (Siless *et al.*, 2018; O'Donnell and Westin, 2007a), or on clustering large multi-subject datasets (Guevara *et al.*, 2012). A detailed description and comparison of several distances and clustering approaches can be found in (Moberts *et al.*, 2005; Olivetti *et al.*, 2017; Siless *et al.*, 2013).

Various studies have also focused on the segmentation of streamlines, toward the goal of drawing cross-population inferences (Guevara *et al.*, 2012; Jin *et al.*, 2014; O'Donnell and Westin, 2007a; Prasad *et al.*, 2014). These studies either follow an atlas based approach (Guevara *et al.*, 2012; Jin *et al.*, 2014; O'Donnell and Westin, 2007a) or align specific tracts directly across subjects (Garyfallidis *et al.*, 2015; O'Donnell *et al.*, 2012). Multi-step or multi-level approaches have also been proposed to segment streamlines, for example, by combining both voxel and streamline groupings (Guevara *et al.*, 2012), fusing labels from multiple hand-labeled atlases (Jin *et al.*, 2014), or using a bundle representation based on maximum density paths (Prasad *et al.*, 2014). A few studies have also investigated the representation of specific streamline bundles using different techniques such as gamma mixture models (Maddah *et al.*, 2008), the computational model of rectifiable currents (Durrleman *et al.*, 2009; Gori *et al.*, 2016), and functional varifolds (Kumar *et al.*, 2017d). For detailed review of white matter clustering approaches, we refer the reader to (O'Donnell *et al.*, 2013).

### 2.3.2 Sparse coding for neuroimaging

Sparse coding, with an objective of 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 (Elad *et al.*, 2010; Wright *et al.*, 2009, 2010; Yang *et al.*, 2009). Various neuroimaging applications have also utilized concepts from this technique, such as the reconstruction (Lustig *et al.*, 2008) or segmentation (Tong *et al.*, 2013) of MRI data, and for functional connectivity analysis (Lee *et al.*, 2016a,b). For diffusion data, sparse coding has been used successfully for clustering white matter voxels from Orientation Density Function (ODF) data (Çetingül *et al.*, 2014), for finding a population-level dictionary of key white matter tracts (Zhu *et al.*, 2016), for higher-order tensor (HOT) based diffusion MRI reconstruction (Feng *et al.*, 2015), and for denoising and reconstruction for diffusion spectrum imaging (Aranda *et al.*, 2015; Bilgic *et al.*, 2013; Gramfort *et al.*, 2014; Merlet *et al.*, 2013).

Recently, several studies have outlined the connection between clustering and factorization problems, such as dictionary learning (Aharon *et al.*, 2006; Sprechmann and Sapiro, 2010) and non-negative matrix factorization (Kim and Park, 2007). Thus, dictionary learning can be seen as a soft clustering, where objects can be linked to more than one cluster. Researchers have also recognized the advantages of applying kernels to existing clustering methods, like the k-means algorithm (Dhillon *et al.*, 2004), as well as dictionary learning approaches (Nguyen *et al.*, 2012). Such "kernel" methods have been shown to better learn the non-linear relations in the data (Hofmann *et al.*, 2008).

Sparse coding and dictionary learning were used in (Moreno *et al.*, 2016; Alexandroni *et al.*, 2017) to obtain a compressed representation of streamlines. In our previous work (Kumar *et al.*, 2015; Kumar and Desrosiers, 2016), we applied these concepts to learn an multi-subject streamline atlas for labelling the streamlines of a new subject. In recent studies, we showed how this idea can be used to derive a brain fingerprint capturing genetically-related information on streamline geometry (Kumar *et al.*, 2017c), and to incorporate along-tract measures of microstructure in the representation (Kumar *et al.*, 2017d).

The present study extends our preliminary work in (Kumar *et al.*, 2017d, 2015; Kumar and Desrosiers, 2016; Kumar *et al.*, 2017c) by providing an in-depth analysis that compares different sparsity priors and evaluates the impact of various parameters. As algorithmic contributions, we present two extensions of the model in (Kumar and Desrosiers, 2016), based on group sparsity and manifold regularization, that provide more meaningful bundles and can incorporate information on streamline geometry, such as the proximity of streamline endpoints, to constrain the clustering process.

## 2.4 Kernel dictionary learning for streamline clustering

In this section we propose kernel dictionary learning and sparsity priors based frameworks for white matter fiber analysis. We start with a brief review of dictionary learning and the k-means algorithm, followed by proposed methods based on various sparsity priors, and algorithm complexity analysis.

### 2.4.1 Dictionary learning and the k-means algorithm

Let $\boldsymbol{X}$ be the set of $n$ streamlines, each represented as a set of 3D coordinates. For the purpose of explanation, we suppose that each streamline $i$ is encoded as a feature vector $\boldsymbol{x}_i \in \mathbb{R}^d$, and that $\boldsymbol{X}$ 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 streamlines having a different number of 3D coordinates could be compared with a suitable similarity measure (i.e., the kernel function).

The traditional (hard) clustering problem can be defined as assigning each streamline to a bundle from a set of $m$ bundles, such that streamlines are as close as possible to their assigned bundle's prototype (i.e., cluster center). Let $\Psi^{m \times n}$ be the set of all $m \times n$ cluster assignment matrices (i.e., matrices in which each row has a single non-zero value equal to one), this problem can be expressed as finding the matrix $\boldsymbol{D}$ of $m$ bundle prototypes and the streamline-to-bundle assignment matrix $\boldsymbol{W}$ that minimize $\|\boldsymbol{X} - \boldsymbol{DW}\|_F^2$. This formulation of the clustering problem can be seen as a special case of dictionary learning, where $\boldsymbol{D}$ is the dictionary and $\boldsymbol{W}$ is constrained to be a cluster assignment matrix, instead of enforcing its sparsity.

While solving this clustering problem is NP-hard, optimizing $\boldsymbol{W}$ or $\boldsymbol{D}$ individually is easy. For a given dictionary $\boldsymbol{D}$, the optimal $\boldsymbol{W}$ assigns each streamline $i$ to the prototype $m$ closest to its feature vector. Likewise, for a fixed $\boldsymbol{W}$, the optimal dictionary is found by solving a simple linear regression problem. This simple heuristic correspond to the well-known k-means algorithm.

### 2.4.2 Kernel k-means

In our streamline clustering problem, the k-means approach described in the previous section has two important disadvantages. First, it requires to encode streamlines as a set of features, which is problematic due to the variation in their length and endpoints. Also, it assumes linear relations between the streamlines and bundle prototypes, while these relations could be better defined in a non-linear subspace (*manifold*).

These problems can be avoided by using a kernel version of k-means for the streamline clustering problem. In this approach, each streamline is projected to a $q$-dimensional space using a mapping function $\boldsymbol{\phi} : \mathbb{R}^d \rightarrow \mathbb{R}^q$, where $q \gg d$. We denote by $\boldsymbol{\Phi}$ the $\mathbb{R}^{q \times n}$ matrix containing the tracts of $\boldsymbol{X}$ mapped with $\boldsymbol{\phi}$. The inner product of two streamlines in this space corresponds to a kernel function $k$, i.e. $k(\boldsymbol{x}_i, \boldsymbol{x}_j) \; = \; \boldsymbol{\phi}(\boldsymbol{x}_i)^\top \boldsymbol{\phi}(\boldsymbol{x}_j)$. With $\boldsymbol{K} \; = \; \boldsymbol{\Phi}^\top \boldsymbol{\Phi}$, the kernel matrix, the *kernel* clustering problem can be expressed as:

$$\underset{\substack{\boldsymbol{D} \in \mathbb{R}^{q \times k} \\ \boldsymbol{W} \in \{0,1\}^{m \times n}}}{\arg\min} \quad \|\boldsymbol{\Phi} - \boldsymbol{DW}\|_F^2 \quad \text{subject to } \; \boldsymbol{W}^\top \mathbf{1}_m \; = \; \mathbf{1}_n. \tag{2.1}$$

Since the dictionary prototypes are defined in the kernel space, $\boldsymbol{D}$ cannot be computed explicitly. To overcome this problem, we follow the strategy proposed in (Nguyen *et al.*, 2012; Rubinstein *et al.*, 2010) and define the dictionary as $\boldsymbol{D} = \boldsymbol{\Phi}\boldsymbol{A}$, where $\boldsymbol{A} \in \mathbb{R}^{n \times m}$.

Using a similar optimization approach as in k-means, we alternate between updating matrix $\boldsymbol{W}$ and $\boldsymbol{A}$. Thus, we update $\boldsymbol{W}$ by assigning each streamline $i$ to the prototype $m$ whose features in the kernel space are the closest:

$$w_{mi} \;=\; \begin{cases} 1: & \text{if } m = \arg\min_{m'} \, [\boldsymbol{A}^\top \boldsymbol{K} \boldsymbol{A}]_{m'm'} \;-\; 2[\boldsymbol{A}^\top \boldsymbol{k}_i]_{m'}, \\ 0: & \text{otherwise.} \end{cases}, \tag{2.2}$$

where $\boldsymbol{k}_i$ corresponds to the $i$-th column of $\boldsymbol{K}$. Recomputing $\boldsymbol{A}$ corresponds once again to solving a linear regression problem with optimal solution:

$$\boldsymbol{A} \;=\; \boldsymbol{W}^\top \big(\boldsymbol{W}\boldsymbol{W}^\top\big)^{-1}. \tag{2.3}$$

We initialize matrix $\boldsymbol{A}$ as a random selection matrix (i.e., random subset of columns in the identity matrix). This is equivalent to using a random subset of the transformed streamlines (i.e., subset of columns in $\boldsymbol{\Phi}$) as the initial dictionary. This optimization process is known as *kernel k-means* (Dhillon *et al.*, 2004).

### 2.4.3 Non-negative kernel sparse clustering

Because they map each streamline to a single bundle, hard clustering approaches like (kernel) k-means can be sensitive to poorly separated bundles and streamlines which do not fit in any bundle (outliers). This section describes a new clustering model that allows one to control the hardness or softness of the clustering.

In the proposed model, the hard assignment constraints are replaced with non-negativity and $L_0$-norm constraints on the columns of $\boldsymbol{W}$. Imposing non-negativity is necessary because the values of $\boldsymbol{W}$ represent the membership level of streamlines to bundles. Moreover, since the $L_0$-

norm counts the number of non-zero elements, streamlines can be expressed as a combination of a small number of prototypes, instead of a single one. When updating the streamline-to-bundle assignments, the columns $\boldsymbol{w}_i$ of $\boldsymbol{W}$ can be optimized independently, by solving the following sub-problem:

$$\underset{\boldsymbol{w}_i \in \mathbb{R}_+^m}{\arg\min} \; \|\phi(\boldsymbol{x}_i) - \boldsymbol{\Phi}\boldsymbol{A}\boldsymbol{w}_i\|_2^2 \quad \text{subject to} \;\; \|\boldsymbol{w}_i\|_0 \leq S_{\max}. \tag{2.4}$$

Parameter $S_{\max}$ defines the maximum number of non-zero elements in $\boldsymbol{w}_i$ (i.e., the sparsity level), and is provided by the user as input to the clustering method.

The algorithm summary and computational complexity is reported in Supplement material, Algorithm 1. To compute non-negative weights $\boldsymbol{w}_i$, we modify the kernel orthogonal matching pursuit (kOMP) approach of (Nguyen *et al.*, 2012) to include non-negativity constrains of sparse weights (Supplement material, Algorithm 2). Unlike kOMP, the most *positively* correlated atom is selected at each iteration, and the sparse weights $\boldsymbol{w}_s$ are obtained by solving a non-negative regression problem. Note that, since the size of $\boldsymbol{w}_s$ is bounded by $S_{\max}$, computing $\boldsymbol{w}_s$ is fast.

In the case of a soft clustering (i.e., when $S_{\max} \geq 2$), updating $\boldsymbol{A}$ with (2.3) can lead to negative values in the matrix. As a result, the bundle prototypes may lie outside the convex hull of their respective streamlines. To overcome this problem, we adapt a strategy proposed for non-negative tri-factorization (Ding *et al.*, 2006) to our kernel model. In this strategy, $\boldsymbol{A}$ is recomputed by applying the following update scheme, until convergence:

$$[\boldsymbol{A}]_{ij} \;\leftarrow\; [\boldsymbol{A}]_{ij} \cdot \frac{\left[\boldsymbol{K}\boldsymbol{W}^\top\right]_{ij}}{\left[\boldsymbol{K}\boldsymbol{A}\boldsymbol{W}\boldsymbol{W}^\top\right]_{ij}}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m. \tag{2.5}$$

The above update scheme produces small positive values instead of zero entries in $\boldsymbol{A}$. To resolve this problem, we apply a small threshold in post-processing. In terms of computational complexity, the bottleneck of the method lies in computing the kernel matrix. For large datasets, we could reduce this computational complexity by approximating the kernel matrix with the Nyström method (Fowlkes *et al.*, 2004; O'Donnell and Westin, 2007b) (Supplement material, Section 1.5).

### 2.4.4 Extension 1: group sparse kernel dictionary learning

The methods proposed above may find insignificant bundles (e.g., bundles containing only a few streamlines) when the parameter controlling the number of clusters is not properly set. Due to the lack of gold standard in tractography analysis, finding a suitable value for this parameter is challenging.

To overcome this problem, we present a new clustering method based on group sparse kernel dictionary learning. We reformulate the clustering problem as finding the dictionary $\boldsymbol{D}$ and non-negative weight matrix $\boldsymbol{W}$ minimizing the following problem:

$$\underset{\substack{\boldsymbol{A}\in\mathbb{R}^{n\times m}\\ \boldsymbol{W}\in\mathbb{R}_{+}^{m\times n}}}{\arg\min}\ \frac{1}{2}\|\boldsymbol{\Phi}-\boldsymbol{\Phi}\boldsymbol{A}\boldsymbol{W}\|_F^2 + \lambda_1\|\boldsymbol{W}\|_1 + \lambda_2\|\boldsymbol{W}\|_{2,1}. \tag{2.6}$$

In this formulation, $\|\boldsymbol{W}\|_1 = \sum_{i=1}^{K}\sum_{j=1}^{N}|w_{ij}|$ is an $L_1$ norm prior which enforces global sparsity of $\boldsymbol{W}$, and $\|\boldsymbol{W}\|_{2,1} = \sum_{i=1}^{K}\|\boldsymbol{w}_{i\boldsymbol{\cdot}}\|_2$ is a mixed $L_{2,1}$ norm prior imposing the vector of row norms to be sparse. Concretely, the $L_1$ norm prior limits the "membership" of streamlines to a small number of bundles, while the $L_{2,1}$ prior penalizes the clusters containing only a few streamlines. Parameters $\lambda_1, \lambda_2 \geq 0$ control the trade-off between these three properties and the reconstruction error (i.e., the first term of the cost function).

We solve this problem using an Alternating Direction Method of Multipliers (ADMM) algorithm (Boyd *et al.*, 2011). First, we introduce ancillary matrix $\boldsymbol{Z}$ and reformulate the problem as:

$$\underset{\substack{\boldsymbol{A}\in\mathbb{R}_{+}^{n\times m}\\ \boldsymbol{W},\boldsymbol{Z}\in\mathbb{R}_{+}^{m\times n}}}{\arg\min}\ \frac{1}{2}\|\boldsymbol{\Phi}-\boldsymbol{\Phi}\boldsymbol{A}\boldsymbol{W}\|_F^2 + \lambda_1\|\boldsymbol{Z}\|_1 + \lambda_2\|\boldsymbol{Z}\|_{2,1}\ \text{subject to}\ \boldsymbol{W}=\boldsymbol{Z}. \tag{2.7}$$

We then convert this an unconstrained problem using an Augmented Lagrangian formulation with multipliers $\boldsymbol{U}$:

$$\underset{\substack{\boldsymbol{A}\in\mathbb{R}_+^{n\times m}\\ \boldsymbol{W},\boldsymbol{Z}\in\mathbb{R}_+^{m\times n}}}{\arg\min}\ \frac{1}{2}\|\boldsymbol{\Phi}-\boldsymbol{\Phi}\boldsymbol{A}\boldsymbol{W}\|_F^2\ +\ \lambda_1\|\boldsymbol{Z}\|_1\ +\ \lambda_2\|\boldsymbol{Z}\|_{2,1}\ +\ \frac{\mu}{2}\|\boldsymbol{W}-\boldsymbol{Z}+\boldsymbol{U}\|_F^2. \tag{2.8}$$

Parameters $\boldsymbol{W}$, $\boldsymbol{Z}$ and $\boldsymbol{U}$ are updated alternatively until convergence. In this work, we use primal feasibility as convergence criteria and stop the optimization once $\|\boldsymbol{W}-\boldsymbol{Z}\|_F^2$ is below a small epsilon.

Dictionary matrix  is updated as (2.5). To update $\boldsymbol{W}$, we derive the objective function with respect to this matrix and set the result to $0$, yielding:

$$\boldsymbol{W}\ =\ \big(\boldsymbol{A}^\top\boldsymbol{K}\boldsymbol{A}+\mu\boldsymbol{I}\big)^{-1}\big(\boldsymbol{A}^\top\boldsymbol{K}+\mu(\boldsymbol{Z}-\boldsymbol{U})\big). \tag{2.9}$$

Note that imposing non-negativity on $\boldsymbol{W}$ is not required since we ensure this property for $\boldsymbol{Z}$ and have $\boldsymbol{W}\approx\boldsymbol{Z}$ at convergence.

Optimizing $\boldsymbol{Z}$ corresponds to solving a group sparse proximal problem (Friedman *et al.*, 2010). This can be done in two steps. First, we do a $L_1$-norm shrinkage by applying the non-negative soft-thresholding operator to each element of $\boldsymbol{W}+\boldsymbol{U}$:

$$\hat{z}_{ij}\ =\ S^+_{\lambda_1/\mu}\big(w_{ij}+u_{ij}\big)\ =\ \max\Big\{w_{ij}+u_{ij}-{}^{\lambda_1}\!/_{\mu},\,0\Big\},\quad i\leq K,\ j\leq N. \tag{2.10}$$

Then, $\boldsymbol{Z}$ is obtained by applying a group shrinkage on each row of $\hat{\boldsymbol{Z}}$:

$$\boldsymbol{z}_{i\boldsymbol{\cdot}}\ =\ \max\Big\{\|\hat{\boldsymbol{z}}_{i\boldsymbol{\cdot}}\|_2-{}^{\lambda_2}\!/_{\mu},\,0\Big\}\cdot\frac{\hat{\boldsymbol{z}}_{i\boldsymbol{\cdot}}}{\|\hat{\boldsymbol{z}}_{i\boldsymbol{\cdot}}\|_2},\quad i\leq K. \tag{2.11}$$

Finally, the Lagrangian multipliers are updated as in standard ADMM methods: $\boldsymbol{U}\ :=\ \boldsymbol{U}+(\boldsymbol{W}-\boldsymbol{Z})$. The overall optimization procedure and its computational complexity are reported in Supplement material, Algorithm 3.

### 2.4.5 Extension 2: kernel dictionary learning with manifold prior

Another challenge in streamline clustering is to generate anatomically meaningful groupings. This may require incorporating prior information into the clustering process, for example, to impose streamlines ending in the same anatomical region to be grouped together. In this work, we address this challenge by proposing a manifold-regularized kernel dictionary learning method.

In the proposed method, we define the manifold as a graph with adjacency matrix $\boldsymbol{G} \in \mathbb{R}^{n \times n}$. In this matrix, $g_{i,i'} = 1$ if streamlines $i$ and $i'$ should be grouped in the same bundle, otherwise $g_{i,i'} = 0$. The manifold regularization prior on the streamline-to-bundle assignments can be formulated as

$$\begin{aligned} \mathcal{R}_{\text{man}}(\boldsymbol{W}) &= \lambda_L \sum_{i=1}^{n} \sum_{i'=1}^{n} g_{i,i'} \, \|\boldsymbol{w}_i - \boldsymbol{w}_{i'}\|_2^2 \\ &= \lambda_L \, \text{tr}(\boldsymbol{W}\boldsymbol{L}\boldsymbol{W}^\top), \end{aligned} \tag{2.12}$$

where $\boldsymbol{L} \in \mathbb{R}^{n \times n}$ is the Laplacian of $\boldsymbol{G}$ and $\lambda_L$ is a parameter controlling the strength of constraints on streamlines.

Our manifold-regularized formulation is obtained by replacing the $L_{2,1}$ prior on $\boldsymbol{W}$ with $\mathcal{R}_{\text{man}}(\boldsymbol{W})$. This new formulation can be solved, as the previous one, with an ADMM algorithm. The main difference occurs when updating $\boldsymbol{W}$, which corresponds to the following problem:

$$\underset{\boldsymbol{W} \in \mathbb{R}^{k \times n}}{\arg\min} \; \|\boldsymbol{\Phi} - \boldsymbol{\Phi}\boldsymbol{A}\boldsymbol{W}\|_F^2 \, + \, \lambda_L \, \text{tr}(\boldsymbol{W}\boldsymbol{L}\boldsymbol{W}^\top) \, + \, \mu\|\boldsymbol{W} - \boldsymbol{Z} + \boldsymbol{U}\|_F^2. \tag{2.13}$$

Derive this objective function with respect to $\boldsymbol{W}$ and setting the result to $0$ gives a Sylvester equation of the form $\boldsymbol{P}\boldsymbol{W} + \boldsymbol{W}\boldsymbol{Q} = \boldsymbol{R}$ where, $\boldsymbol{P} = \boldsymbol{A}^\top\boldsymbol{K}\boldsymbol{A} + \mu\boldsymbol{I}$, $\boldsymbol{Q} = \lambda_L\boldsymbol{L}$, and $\boldsymbol{R} = \boldsymbol{A}^\top\boldsymbol{K} + \mu(\boldsymbol{Z} - \boldsymbol{U})$. This equation can be solved using Bartels-Stewart algorithm (Bartels and Stewart, 1972), which requires transforming $\boldsymbol{P}$ and $\boldsymbol{Q}$ into Schur form with a QR algorithm, and solving the resulting triangular system via back-substitution (Supplement material, Algorithm 4). The computational complexity is $O(n^3)$, $n$ being the size of $\boldsymbol{Q}$. However, this can be drastically reduced by pre-computing once the Schur form of $\boldsymbol{Q}$.

## 2.5 Experimental results and analysis

In this section, we evaluate our proposed methods on a labeled dataset, followed by parameter impact analysis, and concluding with Human Connectome Project data results on clustering and automated segmentation of new subjects.

### 2.5.1 Data and pre-processing

In the first experiment, we compared the proposed methods on a dataset of manually/expert labeled streamline bundles provided by the Sherbrooke Connectivity Imaging Laboratory (SCIL). The source dMRI data was acquired from a 25 year old healthy right-handed volunteer and is described in (Fortin *et al.*, 2012). We used 10 of the largest bundles, consisting of 4449 streamlines identified from the cingulum, corticospinal tract, superior cerebellar penduncle and other prominent regions. Figure 2.2 (left) shows the coronal and sagittal plane view of the ground truth set. Fibernavigator tool (Chamberland *et al.*, 2014) was used for visualizations of this dataset.

To evaluate the performance of our method across a population of subjects, we used two datasets. First, consisting of 12 healthy volunteers (6 males and 6 females, between 19 to 35 years of age) from the freely available MIDAS dataset (Bullitt *et al.*, 2005). For streamline tractography, we used the tensor deflection method (Lazar *et al.*, 2003) with the following parameters: minimum fractional anisotropy of 0.1, minimum streamline length of 100 mm, threshold for streamline deviation angle of 70 degrees. A mean number of 9124 streamlines was generated for the 12 subjects.

Second, we used the pre-processed dMRI data of 10 unrelated subjects (age 22–35) from the Q3 release of the Human Connectome Project (Glasser *et al.*, 2013; Van Essen *et al.*, 2012, 2013), 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 $^{s}/_{mm^2}$, 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[1].

For signal reconstruction and tractography, we used the freely available DSI Studio toolbox (Yeh *et al.*, 2010). All subjects were reconstructed in MNI space using the Q-space diffeomorphic reconstruction (QSDR) (Yeh and Tseng, 2011) option in DSI Studio. We set output resolution to 1 mm. For skull stripping, we used the masks provided with pre-processed diffusion HCP data. Other parameters were set to the default DSI Studio values. Deterministic tractography was performed with the Runge-Kutta method of DSI Studio (Basser *et al.*, 2000; Yeh *et al.*, 2013), 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 quantitative anisotropy (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 streamlines for each subject.

As a note, whether the streamlines, generated from tractography, represent the actual white matter pathways remains a topic of debate (Jones *et al.*, 2013; Thomas *et al.*, 2014). Streamlines 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 streamlines.

---

[1] http://www.humanconnectome.org/documentation/Q3/

### 2.5.2 Experimental methodology

We tested three distance measures used in the literature for the streamline clustering problem: 1) the Hausdorff distance (Haus) (Corouge *et al.*, 2004; O'Donnell and Westin, 2005) which measures the maximum distance between any point on a streamline and its closest point on the other streamline, 2) the mean of closest points distance (MCP) (Corouge *et al.*, 2004) that computes the mean distance between any point on a streamline and its closest point on the other streamline, and 3) the end points distance (EP) (Moberts *et al.*, 2005) measuring the mean distance between the endpoints of a streamline and the closest endpoint on the other streamline.

Fiber distances were converted into similarities by applying a radial basis function (RBF) kernel: $k_{i,i'} = \exp\left(-\gamma \, \mathrm{dist}^2_{i,i'}\right)$. Parameter $\gamma$ was adjusted separately for each distance measure, using the distribution of values in the corresponding distance matrix. Since the tested distance measures are not all metrics, we applied spectrum shift to make kernels positive semi-definite: $\boldsymbol{K}_{\text{psd}} = \boldsymbol{K} + |\lambda_{\min}|\,\boldsymbol{I}$, where $\lambda_{\min}$ is the minimum eigenvalue of $\boldsymbol{K}$. This technique only modifies self similarities and is well adapted to clustering (Chen *et al.*, 2009).

We initialized $\boldsymbol{W}$ using the output of a spectral clustering method (O'Donnell and Westin, 2005), which applies the k-means algorithm on the $10$ first eigenvectors of the normalized Laplacian matrix of $\boldsymbol{K}$. To avoid inversion problems when $\boldsymbol{W}\boldsymbol{W}^\top$ is close to singular, we used a small regularization value of 1e-8. Finally, to compare our method with hard clustering approaches, we converted its soft clustering output to a hard clustering by mapping each streamline $i$ to the bundle $j$ for which $w_{ji}$ is maximum.

We compared our kernel sparse clustering (KSC) approach to four other methods: kernel k-means (KKM) using the same $\boldsymbol{K}$ and initial clustering, the spectral clustering (Spect) approach described above, single linkage hierarchical clustering (HSL) (Moberts *et al.*, 2005), and QuickBundles (QB) (Garyfallidis *et al.*, 2012). The performance of these methods was evaluated using four clustering metrics: the Rand Index (RI) which measures the consistency of the clustering output with respect to the ground truth, the Adjusted Rand Index (ARI) adjusting ARI values by removing the chance agreement, the Normalized Adjusted Rand Index (NARI) that further normalizes the values by considering the cluster sizes, and the Silhouette (SI) measure which does not use the ground truth and measures the ratio between the intra-cluster and inter-cluster distances (Rousseeuw, 1987). While RI, ARI and NARI values range from $0.0$ to $1.0$, SI values are between $-1.0$ and $1.0$. In practice, SI values are generally much lower than $1.0$ due to the intrinsic intra-cluster variance. More information about these metrics can be found in (Moberts *et al.*, 2005; Siless *et al.*, 2013).

Table 2.1 Clustering accuracy of our KSC method ($S_{\text{max}} = 3$), kernel k-means (KKM), spectral clustering (Spect), and hierarchical clustering (HSL), using the Hausdorff, MCP and EP distances, on the SCIL dataset. For KSC, KKM and Spect, the mean accuracy over $10$ initializations with $m$=10 is reported. The best results for a distance and accuracy metric are shown in boldface type.

| **Dist** | **Method** | **RI** | **ARI** | **NARI** | **SI** |
|---|---|---|---|---|---|
| | | mean (std) | mean (std) | mean (std) | mean (std) |
| MCP | KSC | **0.948 (0.012)** | **0.780 (0.051)** | **0.716 (0.047)** | **0.543 (0.032)** |
| | KKM | 0.947 (0.011) | 0.777 (0.049) | **0.716 (0.046)** | 0.541 (0.028) |
| | Spect | 0.942 (0.014) | 0.752 (0.058) | 0.701 (0.047) | 0.515 (0.059) |
| | HSL | 0.915 (0.000) | 0.704 (0.000) | 0.612 (0.000) | 0.474 (0.000) |
| | QB | 0.943 (0.000) | **0.780 (0.000)** | 0.696 (0.000) | 0.486 (0.000) |
| Haus | KSC | **0.924 (0.013)** | **0.658 (0.068)** | **0.634 (0.030)** | **0.425 (0.022)** |
| | KKM | 0.904 (0.020) | 0.589 (0.082) | 0.573 (0.068) | 0.365 (0.054) |
| | Spect | 0.884 (0.018) | 0.517 (0.041) | 0.538 (0.054) | 0.317 (0.069) |
| | HSL | 0.891 (0.000) | 0.640 (0.000) | 0.609 (0.000) | 0.221 (0.000) |
| | QB | 0.851 (0.000) | 0.468 (0.000) | 0.485 (0.000) | 0.143 (0.000) |
| EP | KSC | **0.919 (0.005)** | **0.634 (0.026)** | **0.641 (0.006)** | **0.422 (0.020)** |
| | KKM | 0.915 (0.013) | 0.621 (0.052) | 0.634 (0.034) | 0.410 (0.032) |
| | Spect | 0.911 (0.014) | 0.603 (0.053) | 0.616 (0.040) | 0.408 (0.031) |
| | HSL | 0.842 (0.000) | 0.539 (0.000) | 0.445 (0.000) | 0.197 (0.000) |
| | QB | 0.885 (0.000) | 0.534 (0.000) | 0.550 (0.000) | 0.129 (0.000) |

### 2.5.3 Comparison of methods and distance measures

Table 2.1 gives the accuracy obtained by KSC ($S_{\max}$=3) and the four other tested methods on the SCIL dataset, for the same number of clusters as the ground truth ($m$=10). Since the output of spectral clustering depends on the initialization of its k-means clustering step, for Spect, KSC and KKM, we report the mean performance and standard deviation obtained using $10$ different random seeds. We see that our KSC method improves the initial solution provided by spectral clustering, and gives in most cases a higher accuracy than other clustering methods. We also observe that KSC is more robust to the choice of distance measure than other methods and, as reported in (Moberts *et al.*, 2005), that MCP is consistently better than other distance measures.

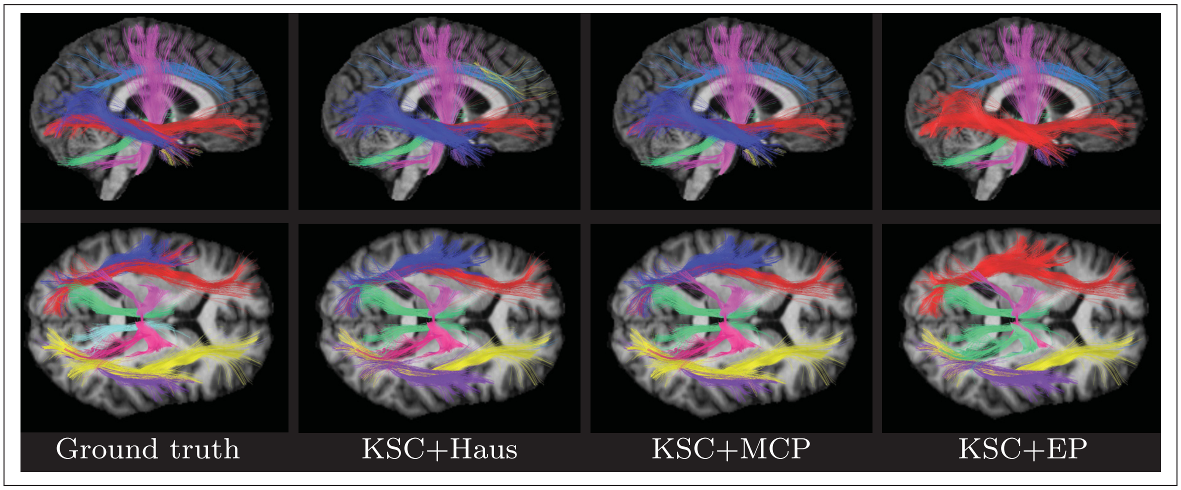


Figure 2.2 Right sagittal (**top**) and inferior axial (**bottom**) views of the ground truth, and bundles obtained by KSC ($S_{\max} = 3$) using the Haus, MCP and EP.

Figure 2.2 compares the ground truth clustering of the SCIL dataset with the outputs of KSC ($S_{\max}$=3) using the Haus, MCP and EP distances. Except for the superior cerebellar peduncle bundle (cyan and green colors in the ground truth), the bundles obtained by KSC+MCP and KSC+Haus are similar to those of the ground truth clustering. Also, we observe that the differences between KSC+MCP and KSC+Haus occur mostly in the right inferior fronto-occipital fasciculus and inferior longitudinal fasciculus bundles (yellow and purple colors in the ground truth). Possibly due to the large variance of endpoint distances in individual bundles, KSC+EP gives poor clustering results.

### 2.5.4 Impact of sparsity

Figure 2.3 reports the mean ARI (over $10$ runs) obtained on the SCIL dataset by our KSC approach, using $S_{\max}$=1,2,3, for an increasing number of clusters (i.e., dictionary size $m$). For comparison, the performance of KKM and Spect is also shown. When the Spectral Clustering initialization is near optimal (i.e., when $m$ is near the true number of clusters and using MCP), both methods find similar solutions. However, when the initial spectral clustering is poor (e.g., Haus and EP distance or small number of clusters) the improvement obtained by KSC is more significant than KKM. Hence, KSC ($S_{\max} \geq 2$) is more robust than hard clustering approaches (i.e., Spect, KKM or KSC with $S_{\max}$=1) to the number of clusters and distance measures.

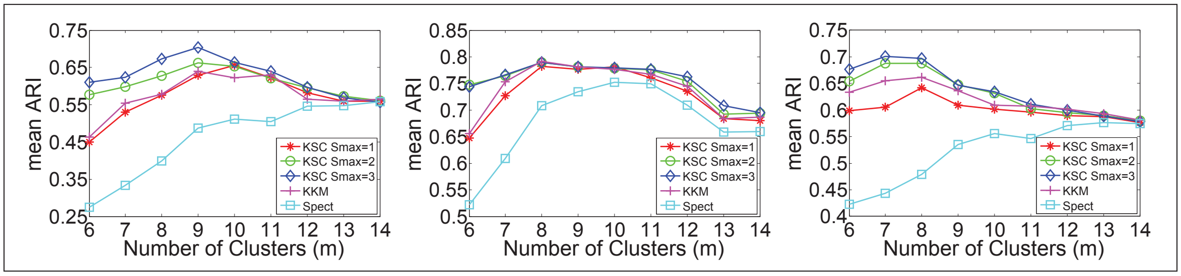


Figure 2.3 Mean ARI obtained on the SCIL dataset by KSC ($S_{\max} = 1, 2, 3$), KKM and Spect, using Haus (**left**), MCP (**center**), EP (**right**); for varying $m$.

To illustrate the soft clustering of KSC, Fig. 2.4 (**left**) and (**center**) show the membership level of streamlines to two different bundles. Streamline colors in each figure correspond to the values of a row in $\boldsymbol{W}$ normalized so that the minimum is 0 (blue) and the maximum is $1$ (red). We observe streamlines having a membership to both bundles (e.g., orange-colored streamlines in the left image), reflecting the uncertainty of this part of the clustering. In Fig. 2.4 (**right**), we show the importance of each streamline in defining the prototype of a bundle, using the

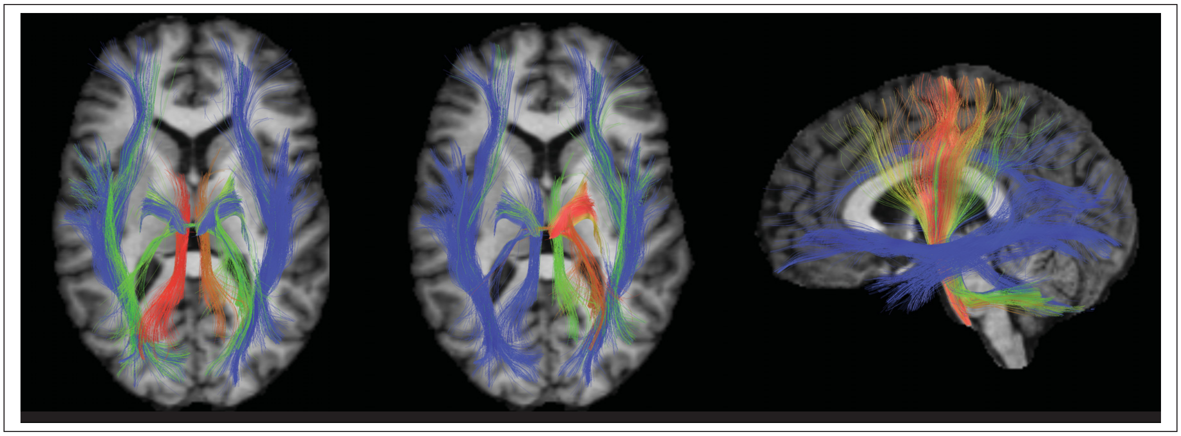
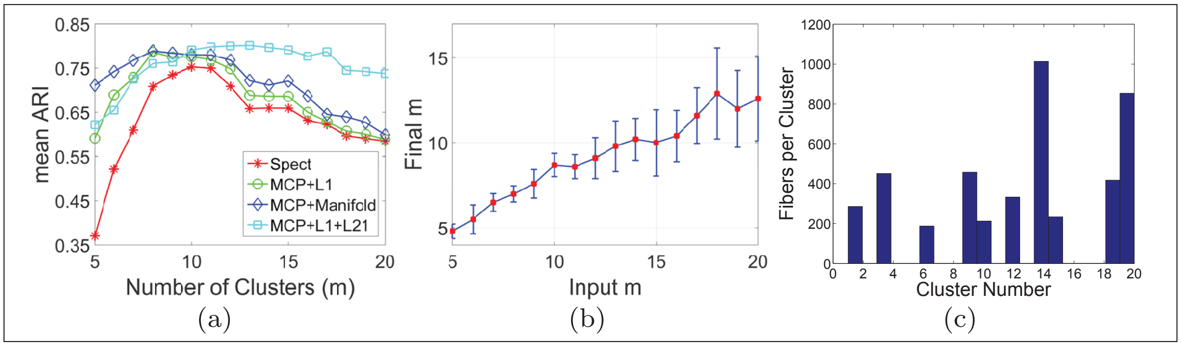


Figure 2.4 Membership level of streamlines to two different bundles (**left** and **center**), and importance of each streamline in defining the prototype of a bundle (**right**). Blue means a null membership/importance, while non-zero values are represented by a color ranging from green (lowest value) to red (highest value).

normalized value of a column in $\boldsymbol{A}$ as colors. It can be seen that only a few streamlines are used to define this bundle, confirming the sparsity of $\boldsymbol{A}$.

### 2.5.5 Group sparsity prior



(a) (b) (c)

Figure 2.5 **(a)** Mean ARI obtained on the SCIL dataset by GKSC, MCP+L1, MCP+Manifold and Spect, using MCP; for varying $m$; **(b)** mean and standard deviation of final $m$ for varying input $m$; **(c)** Distribution of bundle sizes for a sample run using $m = 20$.

Figure 2.5(a) plots the mean Adjusted Rand Index (ARI) obtained by our group sparse model (MCP+L1+L21) for various cluster numbers ($m$), over 10 runs with different spectral clustering

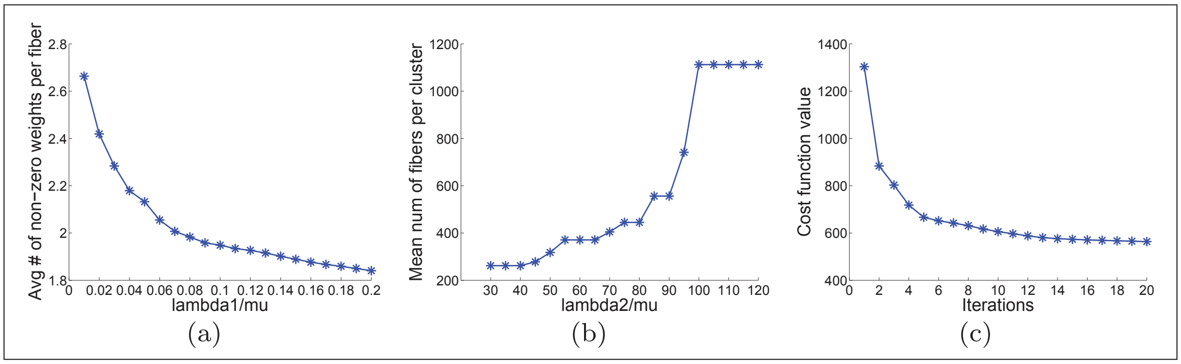
Table 2.2 Clustering accuracy of proposed methods using MCP distances and three types of priors: $L_1$ norm sparsity alone (L1), with group sparsity (L1+L21), and with manifold regularization (L1+Manifold). Reported values are the mean accuracy over $10$ initializations with (input) $m$=10 clusters. The best result for each accuracy metric is shown in boldface type.

| **Prior** | **RI**<br>mean (std) | **ARI**<br>mean (std) | **NARI**<br>mean (std) | **SI**<br>mean (std) |
|---|---|---|---|---|
| L1 | 0.947 (0.011) | 0.775 (0.049) | 0.714 (0.045) | 0.543 (0.029) |
| L1+Manifold | 0.948 (0.010) | 0.780 (0.044) | 0.717 (0.046) | 0.546 (0.033) |
| L1+L21 | **0.949 (0.006)** | **0.791 (0.025)** | **0.721 (0.035)** | **0.563 (0.039)** |

initializations. As baseline, we also report the ARI of spectral clustering and our method without group sparsity (MCP+L1), i.e. using $\lambda_2$=0. We see that employing group sparsity improves clustering quality and provides a greater robustness to the input value of $m$. The advantages of using a group sparse prior are further confirmed in Table 2.2, which gives the mean ARI, RI, NARI and average SI for $m$=10. Results show that MCP+L1+L21 outperforms MCP+L1 for all performance metrics. In a t-test, these improvements are statistically significant with $p$ <0.01.

As described in Section 2.4.4, group sparsity has the benefit of providing meaningful bundles, regardless of the number of clusters $m$ given as input. In Fig. 5(a), we see that the ARI of MCP+L1+L21 increases monotonically until reaching the ground-truth number of bundles $m^*$=10. While the clustering accuracy of other methods drops for $m$ >10, the performance of MCP+L1+L21 remains stable. This is explained in Fig. (b) which plots the number of non-empty clusters found by MCP+L1+L21 as a function of $m$: the number of output clusters stays near to $m^*$=10, even for large values of $m$. As additional confirmation, Fig. 2.5(c) shows the number of streamlines per cluster for a sample run of MCP+L1+L21 with $m = 20$. In this example, the output clustering contains $m^*$=10 non-empty clusters.

In Fig. 2.6, we measure the impact of sparse regularization parameters $\lambda_1$ and $\lambda_2$ for a fixed ADMM parameter of $\mu = 0.01$. As shown in (a), $\lambda_1/\mu$ controls the mean number of non-zero weights per streamline (i.e., how soft or hard is the clustering). Likewise, as illustrated in (b), $\lambda_2/\mu$ defines the size of bundles in the output. These results are consistent with the use $L_1$-norm and $L_{2,1}$-norm sparsity in (2.10). Finally, the optimization stability of the MCP+L1+L21 model is illustrated in Fig. 2.6(c), where convergence is reached around 20 iterations.



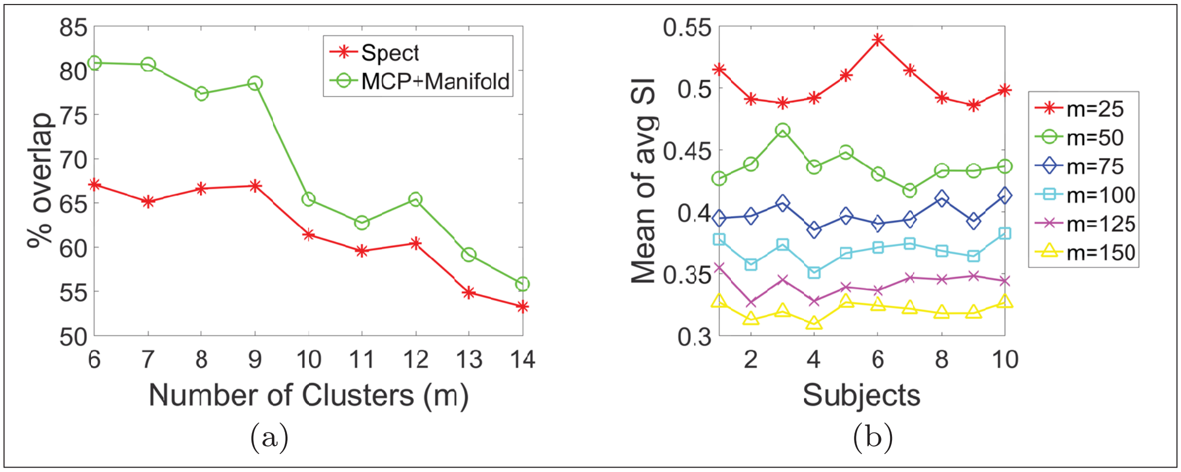
(a) (b) (c)

Figure 2.6 **(a)** Mean number of non-zero assignment weights per streamline, for $\lambda_2/\mu = 80$ and increasing $\lambda_1/\mu$. **(b)** Mean number of streamlines per bundle, for $\lambda_1/\mu = 0.1$ and increasing $\lambda_2/\mu$. **(c)** Cost function value at each of a sample run for MCP+L1+L21.

### 2.5.6 Manifold regularization prior

We apply the proposed manifold regularization prior to enforce the grouping of streamlines with similar end-points. The idea is to obtain bundles that correspond to localized regions of the cortex. To generate the Laplacian matrix in (2.12), we constructed a graph where the nodes are streamlines and two nodes are connected if the distance between their nearest endpoints is below some threshold. Following (Gori *et al.*, 2016), we used a distance threshold of 7mm, giving a Laplacian matrix with overall sparsity near 15%.

In Fig. 2.5(a), we see that the manifold regularization prior (MCP+Manifold) improves performance compared to spectral clustering baseline and $L_1$ norm sparsity (MCP+L1). This improvement is particularly important when the input number of clusters is below that of the ground truth (i.e., $m$ <10). Conversely, for $m > 10$, MCP+Manifold is outperformed by group sparsity (MCP+L1+L21) due to the over-segmentation of streamlines. Fig. 2.7(a) measures the



(a) (b)

Figure 2.7 **(a)** Percentage overlap with EP based Laplacian prior matrix, compared with baseline initialization of spect, for varying $m$. **(b)** Mean of avg SI for KSC+MCP clustering of $10$ unrelated HCP subjects for varying $m$.

the percentage of streamlines with nearby endpoints (i.e., edges in the graph) that are assigned to the same cluster, denoted as overlap in the figure. As expected, the prior helps preserve anatomical information defined by streamline endpoints in the clustering.

### 2.5.7 Validation on HCP data

We evaluated the performance of our kernel sparse clustering (KSC) method on a population of subjects from the Human Connectome Project (HCP). For this experiment, we used two datasets: $10$ unrelated HCP subjects, and subjects from the freely available MIDAS dataset (Bullitt *et al.*, 2005) (results in Supplement material). The objective here is to show applicability of our method across population-subjects, and analyse the impact of inter-subject variability.

Figure 2.7(b) shows the mean of average SI obtained for the $10$ unrelated subjects, using a varying number $m$ of clusters and $3$ runs for each $m$ value. This plot was generated by sampling $5000$ streamlines uniformly over the full tractography ((O'Donnell and Westin, 2007a; Kumar *et al.*, 2017c)) and computing their pairwise MCP distance. We observe that clustering quality decreases with higher values of $m$, and that this quality varies across subjects. A similar trend is observed for MIDAS dataset (Supplement material, Fig. 2). Comparing HCP and MIDAS datasets, a greater average SI is obtained for HCP possibly due to the higher resolution of images in this dataset. Full clustering visualization for $10$ subjects ($m = 50$) and subject 1 for $m = 25, 50, 75, 100, 125, 150$ are shown in Supplement material, Figure 3,4). Note the optimal number of streamline clusters is still an open challenge (O'Donnell and Westin, 2007a), we used $m = 50$ in this study for ease of visualization and interpretation.

Figure 2.8 shows sparse code memberships of streamlines in six different bundles: Corpus Callosum - anterior body (row 1) and central body (row 2), left Inferior Occipitofrontal Fasciculus (IOF) (row 3), left Cortico-Spinal-Tract (CST) (row 4), right IOF (row 5), and right CST (row 6). Results are reported for subject 1 ($m$=25 and $m$=50), subjects $2$ ($m$=50) and subject $3$ ($m$=50). Sparse code values are represented by a color ranging from green (lowest value) to red (highest value). While variations are observed across values of $m$ and subjects, the general shape of bundles recovered by our method is similar.

### 2.5.8 Application to automated tractography segmentation

In this section, we apply the proposed KSC method for the automated segmentation of new subject streamlines. Again, the focus of our analysis is on inter-subject variability and its effect on results. To label streamlines, we used as bundle atlas the dictionaries obtained from 40 unrelated HCP subjects ($4$ dictionaries, each one learned from $10$ subjects. Dictionaries were generated by sampling $5000$ streamlines in each subject and employing MCP as distance measure. Note that expert-labeled streamlines could also be used as dictionary.

The bundles encoded by these dictionaries are depicted in Figure 2.9. Moreover, segmentation results obtained for 4 different subjects using dictionary D1 are shown in Fig. 2.10. For each subject, we give the full segmentation as well as membership values for CC, left/right IOF, and left/right CST bundles. Additionally, to analyze the impact of sampling streamlines from a subject, segmentation results for $5$ instances of subject 1 using D1 are provided in Supplement material. Once more, while we observe variability across segmented streamlines from different subjects, the results obtained by our method are globally consistent across subjects. Similar

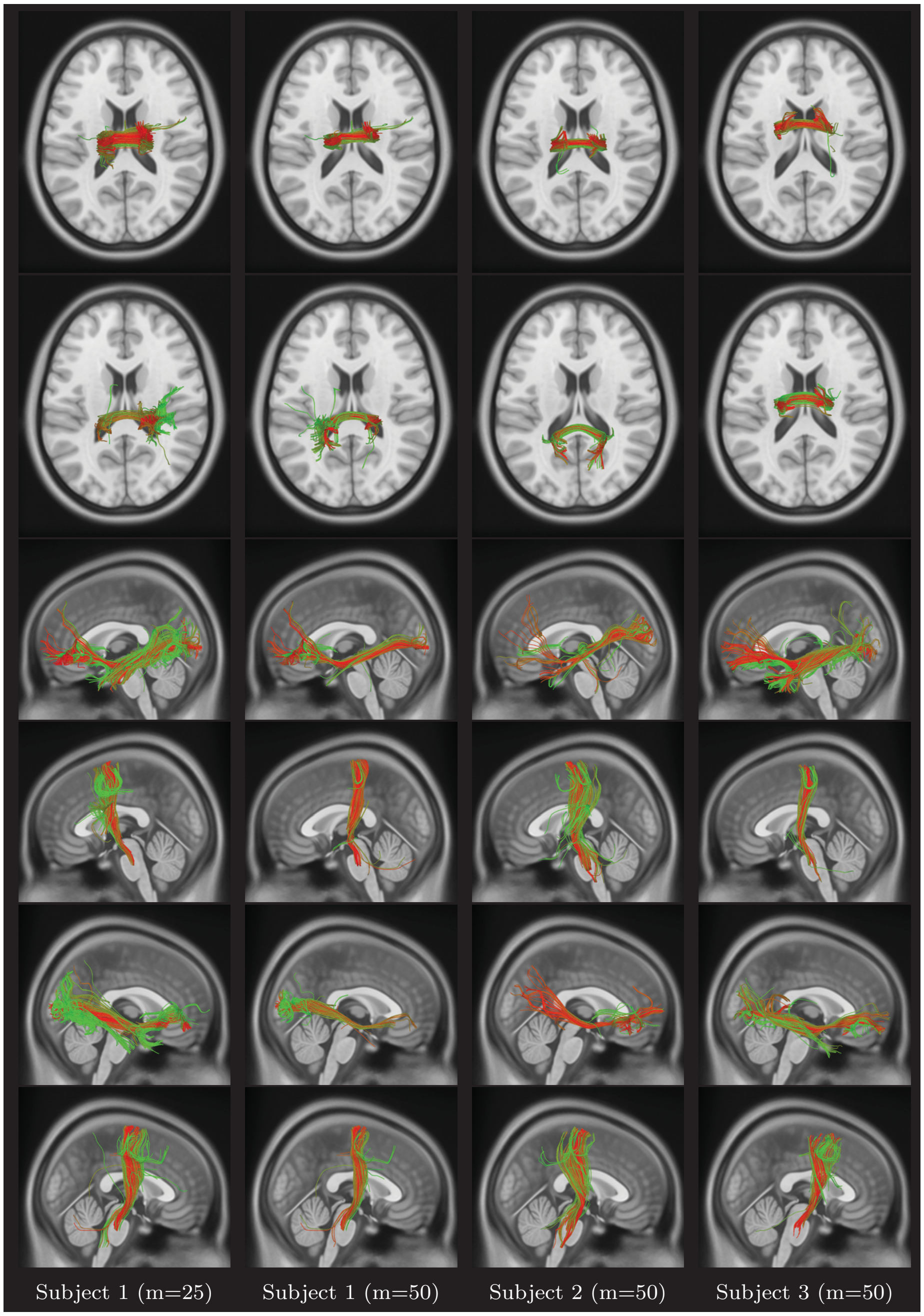


Figure 2.8 Color coded visualization of sparse code memberships of streamlines in Corpus Callosum (CC) (row-1,2), left Inferior Occipitofrontal Fasciculus (IOF) and Cortico-Spinal-Tract (CST) (row-3,4); and right IOF and CST (row-5,6).

consistency is found across multiple instances of the subject 1 (see Supplement material, Fig. 5).

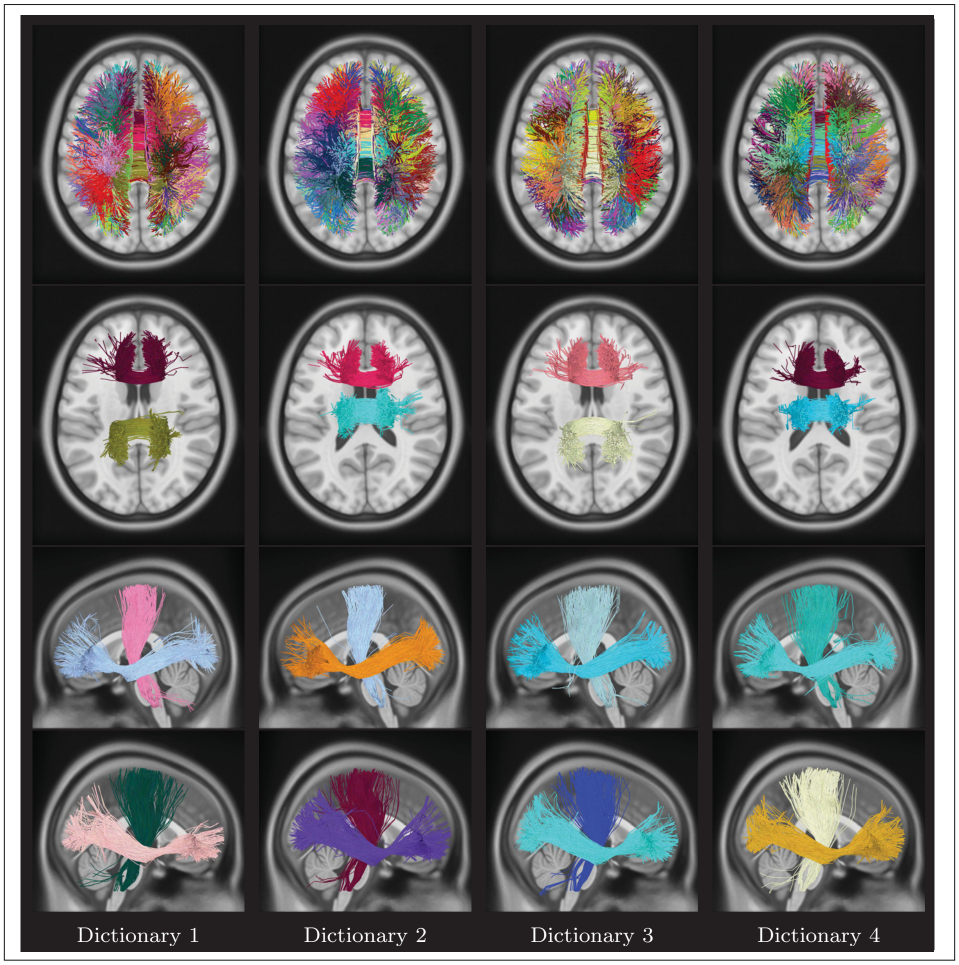


Figure 2.9 Unsupervised multi-subject dictionary visualization. Four different dictionaries and corresponding bundles. Top row: Axial view of full dictionary with a unique color assigned to each bundle; Second row: Anterior Body, and Central Body bundles in Corpus Callosum; Third row: Left CST, and Left IOF bundles; Last row: Right CST, and Right IOF bundles. Each dictionary has a different color code, while the bundles respect that dictionary color-code. (m=50 bundles).

## 2.6 Discussion

We now summarize and discuss the findings related to proposed approaches, impact of various priors, and their applications. We then highlight limitations and additional considerations of this study.

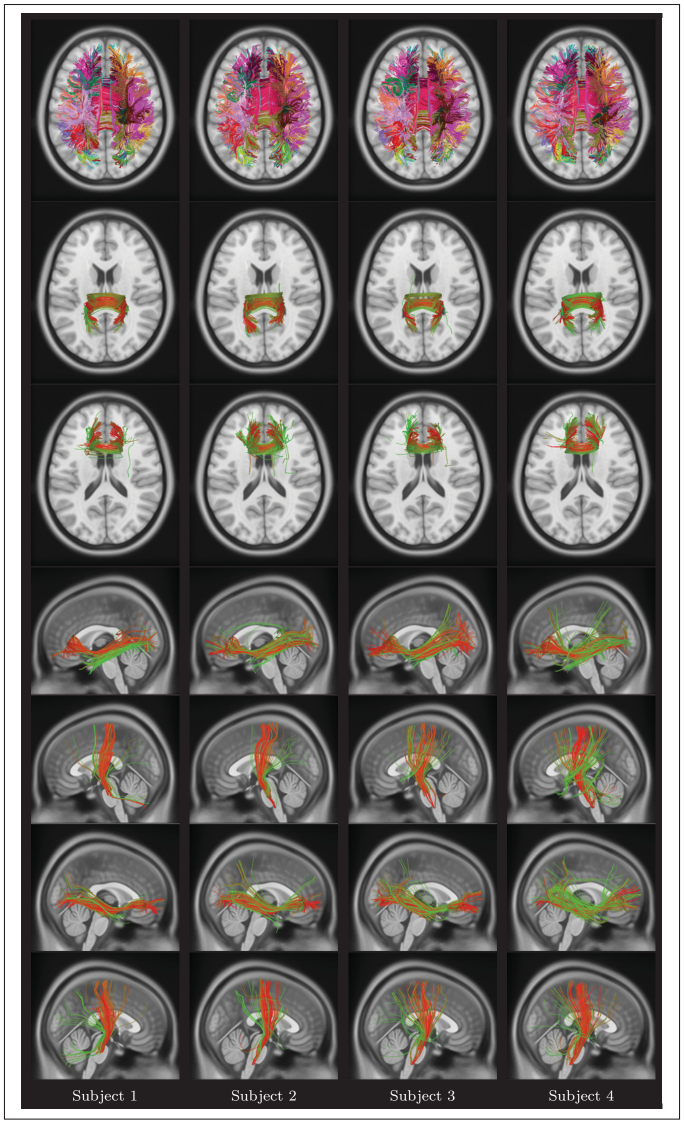


Figure 2.10 Automated segmentation visualization. Top row: full segmentation of 4 HCP subjects using dictionary D1, with a unique color assigned to each cluster, and same color code as D1. Rows 2-7: sparse code (bundle membership) visualization for the posterior body CC, anterior body CC, left IOF, left CST, right IOF, and right CST bundles. Membership values are represented by a color ranging from green (no membership) to red (highest membership).

### 2.6.1 Main findings

Our experiments have demonstrated the usefulness of our kernel sparse clustering (KSC) and various sparsity priors. The soft assignment provided by KSC ($S_{\max} \geq 2$) improved performance for all measures of clustering quality compared to a hard clustering approaches like kernel k-means. This improvement was most significant when the input number of clusters (parameter $m$) is not set close to the ground truth value. In such cases, soft assignment offers a greater robustness to the ambiguous membership of streamlines to bundles.

Comparing the different streamline distances, we found that mean of closest points (MCP) performed the best. Hausdorff distance measures the maximum distance between any point on a streamline and its closest point on another streamline, and thus fails to capture bundles with branching or diverging streamlines. Likewise, end points distances may be more affected by outlier streamlines or issues in diffusion tractography output. These observations are in line with previous analyses on streamline distances (Moberts *et al.*, 2005; Siless *et al.*, 2013).

Results revealed the input number of clusters to have a high impact on results. The true value of this parameter is largely unknown (O'Donnell and Westin, 2007a), and even in expert labeled set could be off the mark due to labeling errors (Moberts *et al.*, 2005). Our analysis showed that group sparsity provides robustness to this confound, and recovers meaningful bundles when it is set far from the ground-truth value. Likewise, the proposed manifold regularization prior helped the clustering by enforcing related pairs of streamlines to be grouped together. This could be useful in a wide range of applications where anatomical information (e.g., cortical parcellation atlas) is available.

Unsupervised clustering of subjects from HCP and MIDAS datasets showed that our KSC method can be employed for data driven analyses, our method finding plausible clusters corresponding to well known bundles. Moreover, the visualization of clusters and membership values demonstrates that KSC can effectively capture inter-subject variability. Experiments on automated streamline segmentation also revealed that KSC can accurately recover major bundles in new subjects, and that this segmentation is robust to the number of clusters, inter-individual variations, and the sampling of streamlines from the same subject.

### 2.6.2 Limitations and additional considerations

Due to the lack of gold standard clustering, as well as the various challenges in diffusion tractography (Maier-Hein *et al.*, 2017) and its interpretation (Jones *et al.*, 2013), validating streamline clustering approaches is difficult. A large scale and data-driven analysis, for example using data from over $1000$ HCP subjects, could lead to interesting observations on number of bundles and their population-wise variability.

An important aspect of our dictionary learning method is its initialization. While we employed spectral clustering for this task, considering other techniques could possibly lead to better results. For the automated segmentation of streamlines in new subjects, we learned the dictionary in an unsupervised setting, however expert-labeled streamlines set or atlas/clustering from other approaches can also be utilized.

One the main advantages of the proposed kernel-based framework is that it alleviates the need for an explicit streamline representation. Previous attempts in utilizing dictionary learning and sparse coding for streamline clustering might have been hindered by this. Employing kernels also provides flexibility and enables the extension to other streamline similarity measures, which can incorporate a richer set of characteristics such as along-tract diffusivity (Kumar *et al.*, 2017d; Charon and Trouvé, 2013; Charlier *et al.*, 2014).

Another key element of our study is the anatomical interpretation of clustering results. The streamlines generated from diffusion tractography provide a macro-scale inference of the underlying fibers(Jones *et al.*, 2013; Maier-Hein *et al.*, 2017). As such, the clustering for a given distance/similarity measures focuses primarily on the geometric aspect of streamlines. Although we considered end points proximity in our manifold regularization prior, additional information such as structural parcellation could be incorporated to improve the anatomic plausibility of the final clustering (O'Donnell *et al.*, 2013; Siless *et al.*, 2018).

The sparse code representation of streamlines conveys a wealth of information on inter-individual variability in terms of streamline geometry. Extension of this study could leverage this information for additional tasks, such as identifying noisy/spurious streamlines, discovering tract-based biomarkers to discriminate between healthy and diseased subjects (O'Donnell *et al.*, 2017), or establishing bundle-to-bundle correspondences across subjects.

## 2.7 Conclusion

We presented a novel framework using kernel dictionary learning with various sparsity priors for the unsupervised segmentation of white matter streamlines. The proposed framework does not require explicit streamline representation and enables using any streamline similarity measure. Dictionary bundles are encoded as a non-negative combination of training streamlines, and the kernel trick is used to model non-linear relationships between streamlines and bundles.

We compared our method against state-of-the-art streamline clustering approaches using expert-labeled data, as well as subjects from the HCP and MIDAS dataset. Results demonstrate the usefulness of having a soft assignment, and that our method is suitable for scenarios where streamlines are not clearly separated, bundles overlap, or when there is important inter-individual variability. Experiments using group sparsity ($L_{2,1}$ norm) and manifold regularization show that these priors can improve clustering quality by adding robustness to the input number of clustering or incorporating anatomical constraints in the clustering.

The benefits of the proposed approach in cases of inter-individual variability was showcased for the automated segmentation of streamlines from new subjects. In future work, we will investigate the usefulness of our approach for identifying and comparing major bundles in healthy vs diseased subjects, and for incorporating along-tract measures in the clustering process.

## 2.8 Supplement results

### 2.8.1 Non-negative kernel sparse clustering: Algorithm summary

**Algorithm 2.1:** Kernelized sparse clustering method

```
Input: Pairwise streamline distance matrix $S_{dist} \in \boldsymbol{R}^{n \times n}$;
Input: The desired number of streamline bundles $m$;
Input: The RBF kernel parameter $\gamma$;
Input: The sparsity level $S_{\max}$ and maximum number of iterations $T_{\max}$;
Output: The sparse assignment matrix $\boldsymbol{W} \in \boldsymbol{R}^{n \times m}$;
Output: The hard assignment vector $\boldsymbol{c} \in \{1, \dots, m\}^n$;
Initialize the kernel matrix: $k_{ij} = \exp(-\gamma \cdot dist_{ij}^2)$ ;
Initialize $\boldsymbol{A}$ as a random selection matrix;
for $t = 1, \dots, T_{\max}$ do
    Update each column $\boldsymbol{w}_i$ of $\boldsymbol{W}$ using NNKOMP (Algorithm 2.2);
    Update dictionary until convergence:
    $\boldsymbol{A}_{ij} \leftarrow \boldsymbol{A}_{ij} \frac{\left(\boldsymbol{K}\boldsymbol{W}^{\top}\right)_{ij}}{\left(\boldsymbol{K}\boldsymbol{A}\boldsymbol{W}\boldsymbol{W}^{\top}\right)_{ij}}, \quad i = 1, \dots, n, \quad j = 1, \dots, m.$;
    $t_{\text{out}} \leftarrow t_{\text{out}} + 1$;
end
Compute hard assignment: $c_i = \arg\max_{k'} w_{im'}, \; i = 1, \dots, n$ ;
return $\{\boldsymbol{W}, c\}$ ;
```

#### Algorithm complexity

In Algorithm 2.1, the user provides a matrix $S_{dist}$ of pairwise streamline distances, as well as the desired number of bundles (clusters), and obtains in return the soft (matrix $\boldsymbol{W}$) and hard (vector $\boldsymbol{c}$) streamline clusterings. Various distance measures, suitable for streamlines, are described in experiments. The distances are converted into similarities by using a Gaussian (RBF) kernel of parameter $\gamma$. Note that the obtained kernel is semi-definite positive only if the distance is a metric. However, non-metric distances, such as the Hausdorff distance (see experiments), have been shown to be quite useful in practice (Laub and Müller, 2004). In the main loop, the dictionary matrix $\boldsymbol{A}$ and sparse streamline-to-bundle assignment matrix $\boldsymbol{W}$ are optimized alternatively, until convergence or $T_{\max}$ iterations have been reached. The soft clustering of $\boldsymbol{W}$ is converted to a hard clustering by assigning each streamline $i$ to the bundle $m$ for which $w_{im}$ is maximum.

**Algorithm 2.2:** Non-negative kernelized orthogonal matching pursuit

```
Input: The dictionary matrix $\boldsymbol{A} \in \boldsymbol{R}_{+}^{n \times m}$ and kernel matrix $\boldsymbol{K} \in \boldsymbol{R}^{n \times n}$;
Input: The streamline index $i$ and sparsity level $S_{\max}$;
Output: The set of non-zero weights $I_s$ and corresponding weight values $\boldsymbol{w}_s$;
Initialize set of selected atoms and weights: $I_0 = \emptyset, \ \boldsymbol{w}_0 = \emptyset$;
for $s = 1, \ldots, S_{\max}$ do
    $\tau_j = \big[\boldsymbol{A}^\top\big(\boldsymbol{k}_i - \boldsymbol{K}\,\boldsymbol{A}_{[I_s]}\boldsymbol{w}_s\big)\big]_j \,/\, \big[\boldsymbol{A}^\top \boldsymbol{K} \boldsymbol{A}\big]_{jj}, \ \ j = 1, \ldots, m$;
    $j_{\max} = \arg\max_{j \notin I_{s-1}} \tau_j, \ \ I_s = I_{s-1} \cup j_{\max}$;
    $\boldsymbol{w}_s = \arg\min_{\boldsymbol{w} \in \boldsymbol{R}_{+}^{s}} \ \boldsymbol{w}^\top \boldsymbol{A}_{[I_s]}^\top \boldsymbol{K} \boldsymbol{A}_{[I_s]} \boldsymbol{w} \ - \ 2\,\boldsymbol{k}_i{}^\top \boldsymbol{A}_{[I_s]} \boldsymbol{w}$;
end
return $\{I_s, \, \boldsymbol{w}_s\}$ ;
Note: $\boldsymbol{A}_{[I_s]}$ contains the columns of $\boldsymbol{A}$ whose index is in $I_s$ ;
```

### 2.8.2 Group sparse kernel dictionary learning: Algorithm summary

The clustering process of our proposed method is summarized in Algorithm 2.3. In this algorithm, the user provides a matrix $S_{dist}$ of pairwise streamline distances (see experiments for more details), the maximum number of clusters $m$, as well as the trade-off parameters $\lambda_1, \lambda_2$, and obtains as output the dictionary matrix $\boldsymbol{A}$ and the cluster assignment weights $\boldsymbol{W}$. At each iteration, $\boldsymbol{W}$, $\boldsymbol{Z}$ and $\boldsymbol{U}$ are updated by running at most $T_{\text{in}}$ ADMM loops, and are then used to update $\boldsymbol{A}$. This process is repeated until $T_{\text{out}}$ iterations have been completed or the cost function $f(\boldsymbol{D}, \boldsymbol{W})$ converged. The soft assignment of $\boldsymbol{W}$ can be converted to a hard clustering by assigning each streamline $i$ to the bundle $m$ for which $w_{im}$ is maximum.

**Algorithm 2.3:** ADMM method for group sparse kernelized clustering

```
Input: Pairwise streamline distance matrix $S_{dist} \in \boldsymbol{R}^{n \times n}$;
Input: The maximum number of streamline bundles $m$;
Input: The RBF kernel parameter $\gamma$;
Input: The cost trade-off parameters $\lambda_1, \lambda_2$ and Lagrangian parameter $\mu$;
Input: The maximum number of inner and outer loop iterations $T_{\text{in}}, T_{\text{out}}$;
Output: The dictionary $\boldsymbol{A} \in \boldsymbol{R}^{n \times m}$ and assignment weights $\boldsymbol{W} \in \boldsymbol{R}_{+}^{n \times m}$;
Initialize the kernel matrix: $k_{ij} = \exp(-\gamma \cdot dist_{ij}^2)$;
Initialize $\boldsymbol{A}$ as a random selection matrix and $t_{\text{out}}$ to 0;
while $f(\boldsymbol{D}, \boldsymbol{W})$ not converged and $t_{\text{out}} \leq T_{\text{out}}$ do
    Initialize $\boldsymbol{U}$ and $\boldsymbol{Z}$ to all zeros and $t_{\text{in}}$ to zero;
    while $||\boldsymbol{W} - \boldsymbol{Z}||_F^2$ not converged and $t_{\text{in}} \leq T_{\text{in}}$ do
        Update $\boldsymbol{W}$, $\boldsymbol{Z}$ and $\boldsymbol{U}$:
            $\boldsymbol{W} \leftarrow \big(\boldsymbol{A}^\top \boldsymbol{K} \boldsymbol{A} + \mu \boldsymbol{I}\big)^{-1} \big(\boldsymbol{A}^\top \boldsymbol{K} + \mu(\boldsymbol{Z} - \boldsymbol{U})\big)$;
            $\hat{z}_{ij} \leftarrow \max\Big\{w_{ij} + u_{ij} - \frac{\lambda_1}{\mu}, 0\Big\}, \quad i \leq m, \ j \leq n$;
            $\boldsymbol{z}_{i\cdot} \leftarrow \max\Big\{||\hat{\boldsymbol{z}}_{i\cdot}||_2 - \frac{\lambda_2}{\mu}, 0\Big\} \cdot \frac{\hat{\boldsymbol{z}}_{i\cdot}}{||\hat{\boldsymbol{z}}_{i\cdot}||_2}, \quad i \leq m$;
            $\boldsymbol{U} \leftarrow \boldsymbol{U} + \big(\boldsymbol{W} - \boldsymbol{Z}\big)$;
        $t_{\text{in}} \leftarrow t_{\text{in}} + 1$;
    end
    Update dictionary until convergence:
    $\boldsymbol{A}_{ij} \leftarrow \boldsymbol{A}_{ij} \frac{\big(\boldsymbol{K}\boldsymbol{W}^\top\big)_{ij}}{\big(\boldsymbol{K}\boldsymbol{A}\boldsymbol{W}\boldsymbol{W}^\top\big)_{ij}}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m.$;
    $t_{\text{out}} \leftarrow t_{\text{out}} + 1$;
end
return $\{\boldsymbol{A}, \boldsymbol{W}\}$ ;
```

The complexity of this algorithm is mainly determined by the initial kernel computation, which takes $O(n^2)$ operations, and updating the assignment weights in each ADMM loop, which has a total complexity in $O(T_{\text{out}} \cdot T_{\text{in}} \cdot m^2 \cdot n)$. Since $T_{\text{out}}$, $T_{\text{in}}$ and $m$ are typically much smaller than $n$, the main bottleneck of the method lies in computing the pairwise distances $S_{dist}$ used as input. For datasets having a large number of streamlines (e.g., more than $n = 100,000$ streamlines), this matrix could be computed using an approximation strategy such as the the Nyström method (Fowlkes *et al.*, 2004), described later in the paper.

### 2.8.3 Kernel dictionary learning with manifold prior: Algorithm summary

#### Update W: Bartels-Stewart Algorithm summary

Algorithm 2.4 provides the summary of $\boldsymbol{W}$ update using Bartels-Stewart Algorithm. The Schur forms can be precomputed to speed up the computation.

**Algorithm 2.4:** Bartels-Stewart Algorithm summary

```
Input: $P$, $Q$, and $R$;
Output: $W$

Step 1: Transfrom $P$ and $Q$ into Schur form
$C_c = R$;
$[Q_a, T_a] = schur(P)$;    $C_c = Q_a^\top C_c$
$[Q_b, T_b] = schur(Q)$;    $C_c = C_c Q_b$

Step 2: Solve following Simplified Sylvester equation using back substitution
  $T_a W + W T_b = C_c$

Step 3: Recover $W$:
$W = Q_a W Q_b^\top$
return $\{W\}$ ;
```

Algorithm 2.5 provides summary of the methods, while complexity can be computed similar to previous section, with only difference being update of W.

**Algorithm 2.5:** ADMM method for kernelized dictionary learning with Laplacian prior

```
Input: Pairwise streamline distance matrix $S_{dist} \in \boldsymbol{R}^{n \times n}$;
Input: The maximum number of streamline bundles $m$;
Input: The RBF kernel parameter $\gamma$;
Input: The cost trade-off parameters $\lambda_1, \lambda_L$ ;
Input: Lagrangian parameter $\mu_1$;
Input: The maximum number of inner and outer loop iterations $T_{\text{in}}, T_{\text{out}}$;
Output: The dictionary $\boldsymbol{A} \in \boldsymbol{R}^{n \times m}$ and assignment weights $\boldsymbol{W} \in \boldsymbol{R}_{+}^{n \times m}$;
Initialize the kernel matrix: $k_{ij} = \exp(-\gamma \cdot dist_{ij}^2)$;
Initialize $\boldsymbol{A}$ as a random selection matrix and $t_{\text{out}}$ to 0;
Precompute $schur(\lambda_L \boldsymbol{L})$;
while $f(\boldsymbol{D}, \boldsymbol{W})$ not converged and $t_{\text{out}} \leq T_{\text{out}}$ do
    Initialize $\boldsymbol{U}$ and $\boldsymbol{Z}$ to all zeros and $t_{\text{in}}$ to zero;
    while $||\boldsymbol{W} - \boldsymbol{Z}||_F^2$ not converged and $t_{\text{in}} \leq T_{\text{in}}$ do
        Update $\boldsymbol{W}$, $\boldsymbol{Z}$, and $\boldsymbol{U}$:
            $\boldsymbol{W} \leftarrow Sylvester\big((\boldsymbol{A}^\top \boldsymbol{K} \boldsymbol{A} + \mu_1 \boldsymbol{I}), \lambda_L \boldsymbol{L}, (\boldsymbol{A}^\top \boldsymbol{K} + \mu_1(\boldsymbol{Z} - \boldsymbol{U}))\big)$
            $\hat{z}_{ij} \leftarrow \max\Big\{w_{ij} + u_{ij} - \frac{\lambda_1}{\mu_1}, 0\Big\}, \quad i \leq m, \ j \leq n;$
            $\boldsymbol{U} \leftarrow \boldsymbol{U} + \big(\boldsymbol{W} - \boldsymbol{Z}\big);$
        $t_{\text{in}} \leftarrow t_{\text{in}} + 1$;
    end
    Update dictionary until convergence:
    $\boldsymbol{A}_{ij} \leftarrow \boldsymbol{A}_{ij} \frac{\big(\boldsymbol{K}\boldsymbol{W}^\top\big)_{ij}}{\big(\boldsymbol{K}\boldsymbol{A}\boldsymbol{W}\boldsymbol{W}^\top\big)_{ij}}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m.$;
    $t_{\text{out}} \leftarrow t_{\text{out}} + 1$;
end
return $\{\boldsymbol{A}, \boldsymbol{W}\}$ ;
```

### 2.8.4 Group sparsity and manifold prior visualization

The bundles obtained by group sparsity (MCP+L1+L21) for the input number of clusters $m = 20$ are presented in Figure 2.11 (middle). We observe that the clustering is similar to the ground truth clustering, except for small differences in left/right inferior longitudinal fasciculus bundles (purple and blue colors in the ground truth). Also, we observe that superior cerebellar peduncle bundles (cyan and green colors in the ground truth) are well separated.

Figure 2.11(right) shows clustering output using this method for $m = 10$ and MCP for a sample run. We observe that the clustering is similar to the ground truth clustering, except for small differences in left/right inferior longitudinal fasciculus bundles (purple and blue colors in the ground truth).

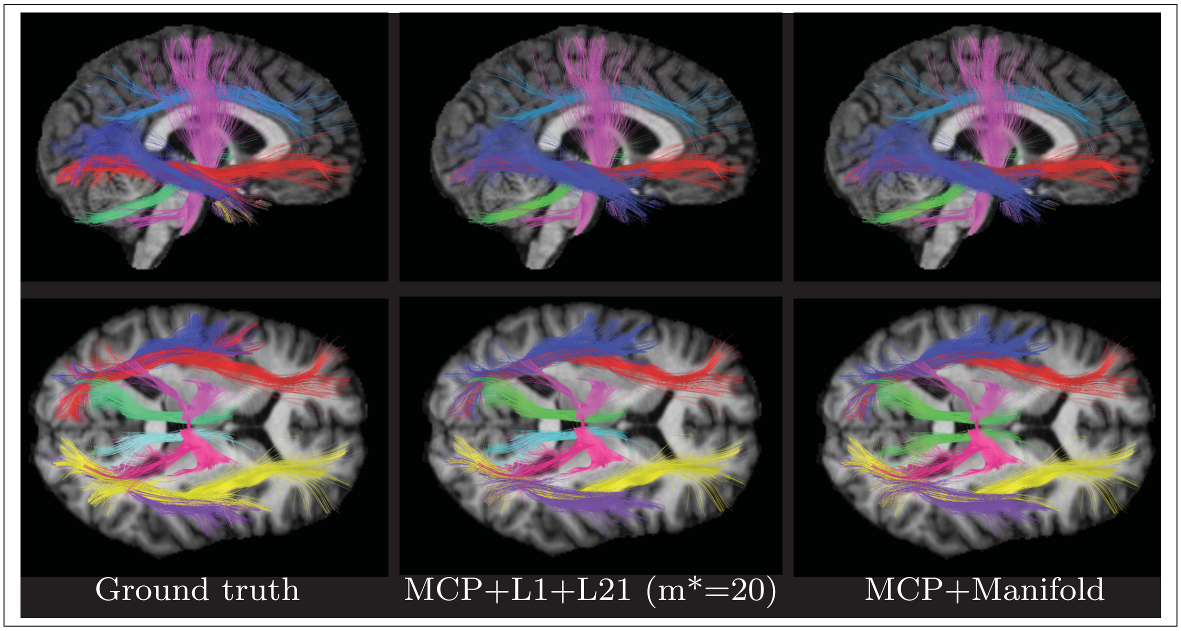


Figure 2.11 Right sagittal (**top**) and inferior axial (**bottom**) views of the ground truth (left), and bundles obtained by MCP+L1+L21 (middle, m=20, final m=10), and MCP+L1+Lap (right, m=10).

### 2.8.5 Results on multi-subject MIDAS dataset (KSC+MCP)

**Data:** To evaluate the performance of our method on multiple subjects, we also used the data of 12 healthy volunteers (6 males and 6 females, 19 to 35 years of age) from the freely available MIDAS dataset (Bullitt *et al.*, 2005). For fiber tracking, we used the tensor deflection method (Lazar *et al.*, 2003) with the following parameters: minimum fractional anisotropy of 0.1, minimum streamline length of 100 mm, threshold for streamline deviation angle of 70 degrees. A mean number of 9124 streamlines was generated for the 12 subjects.

**Results:** Figure 2.12 (left) shows the mean SI (averaged over all clusters) obtained by KSC ($S_{\max} = 3$), KKM and Spect with MCP, on 12 subjects of the MIDAS dataset. We see that our soft clustering method outperforms the hard clustering approaches, especially for a small number of clusters. In Figure 2.12 (right), the results obtained for $m = 35$ are detailed for each subject. Error bars in the plot show the mean and variance of SI values obtained over 10 different initializations. As can be seen, our method shows a greater accuracy and less variance across subjects.

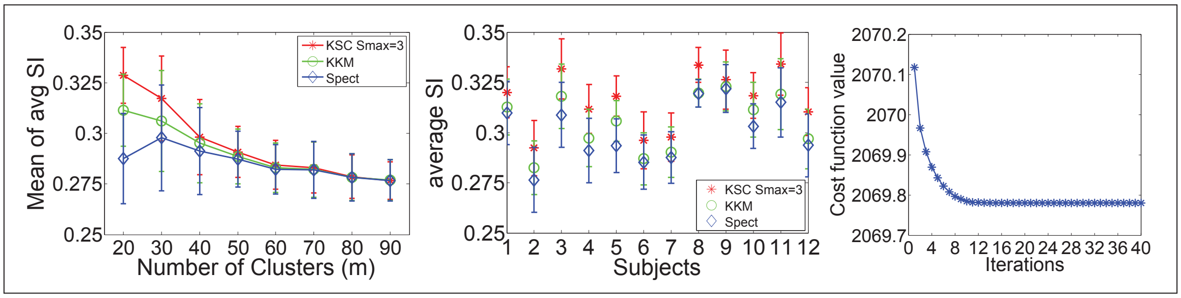


Figure 2.12 MIDAS: Mean of average SI using MCP: varying $m$ mean over 12 subjects (**left**); for $m = 35$ for each subject (**middle**); Convergence plot (**right** for KSC+MCP )

### 2.8.6 Additional results on Human Connectome Project subjects

Figure 2.13 shows clustering output for 10 HCP subjects for $m = 50$, with an unique color assigned to each cluster. For this simplified visualization each streamline is assigned to a single cluster by taking the maximum for each column of the matrix $\boldsymbol{W}$. Note, we have used a unique color code for each subject, as establishing a cluster correspondence across subjects is itself a challenging problem. We observe that the overall pattern of clustering across subjects looks similar. However, there are subtle variations for clusters across subjects.

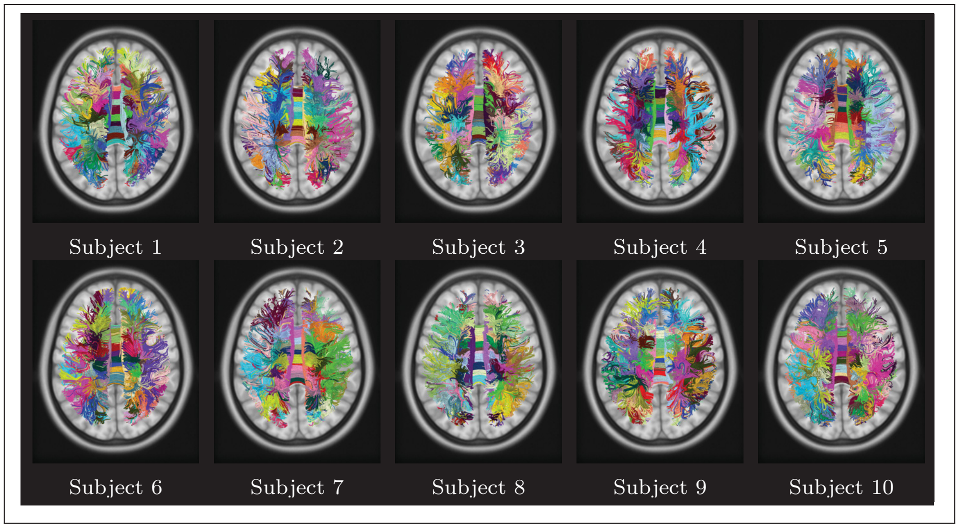


Figure 2.13 Visualization of clustering output for 10 unrelated HCP subjects using KSC, for m = 50.

Similarly, Figure 2.14 shows simplified visualization of clustering output for subject 1, for varying $m$. As expected, going from $m = 25$ to $m = 150$ the clusters split into smaller ones, for example, observe the clusters in corpus callosum.

### 2.8.7 Multi-subject clustering as dictionary

#### Computing the kernel matrix using the Nyström method

When there can be multiple subjects, each subject having several thousands of streamlines, computing the similarity between all pairs of training streamlines in $\boldsymbol{K}$ is impossible. To alleviate this problem, we approximate the kernel matrix using the Nyström method (Fowlkes *et al.*, 2004; O'Donnell and Westin, 2007b). In this method, a set of $p$ representative streamlines are sampled from while set of training streamlines, where $p \ll |\boldsymbol{X}|$. The pairwise similarities between all selected streamlines are then computed in a reduced kernel matrix $\boldsymbol{K}_a \in \boldsymbol{R}^{p\times p}$. Likewise, the similarity between the selected and non-selected ones are obtained in a matrix $\boldsymbol{K}_b \in \boldsymbol{R}^{p\times(|\boldsymbol{X}|-p)}$. The whole kernel matrix is then reconstructed using a low-rank approx-

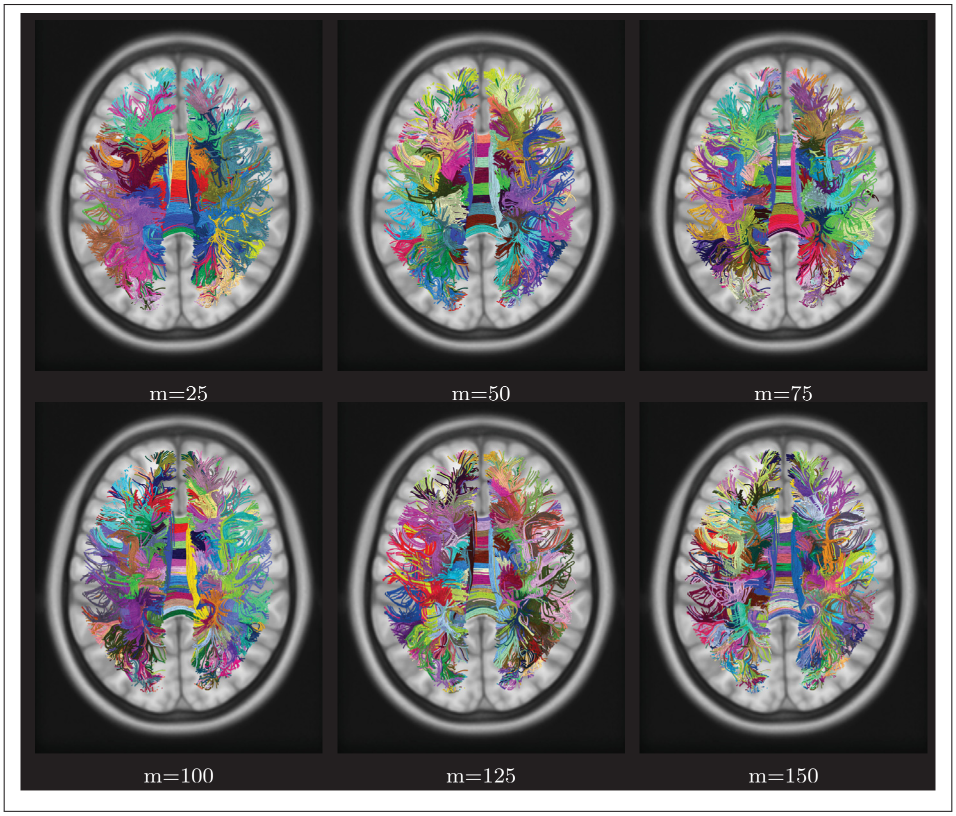


Figure 2.14 Visualization of clustering output for subject 1 using KSC, for varying $m$.

imation $\boldsymbol{K} = \boldsymbol{G}\boldsymbol{G}^\top$, where $\boldsymbol{G}^\top = \boldsymbol{K}_a^{-\frac{1}{2}}\left[\boldsymbol{K}_a^\top \boldsymbol{K}_b^\top\right]$. In practice, the most computationally expensive step of this method is the SVD decomposition of $\boldsymbol{K}_a$.

### HCP multi-subject clustering

When there are multiple subjects, each subject having several thousands of streamlines, computing the similarity between all pairs of training streamlines in $\boldsymbol{K}$ is impossible. To alleviate this problem, we approximate the kernel matrix using the Nyström method (Fowlkes *et al.*, 2004; O'Donnell and Westin, 2007b). Figure 9 (manuscript) shows simplified visualization of $\boldsymbol{A}$ matrix of 4 sets of $10$ unrelated HCP subjects. (These subjects are utilized as dictionary in next section). We utilized $50,000$ streamlines for each set, sampling $5,000$ streamlines from each subject.

For simplification, we show full clusterings and select bundles including Anterior Body, and Central Body bundles in Corpus Callosum; Third row: Left Cortico-Spinal-Tract, and Left Arcuate Fasciculus bundles; Last row: Right Cortico-Spinal Tract, and Right Inferior Occipitofrontal Fasciculus bundles. The objective here is to show that the multi-subject clustering output provides plausible clusters, corresponding to well-known anatomical bundles. Also, comparing multi-subject clustering with single subject clustering, we observe overall similarity in terms of clusters, while also reflecting variation. We also observe subtle variations across multi-subject clustering sets, for example, within IOF or CST bundles.

### 2.8.8 Application: automated segmentation of new subject streamlines

To analyze the impact of sampling streamlines from a subject, Figure 2.15 shows segmentation output for 5 instances of subject 1 using dictionary D1.

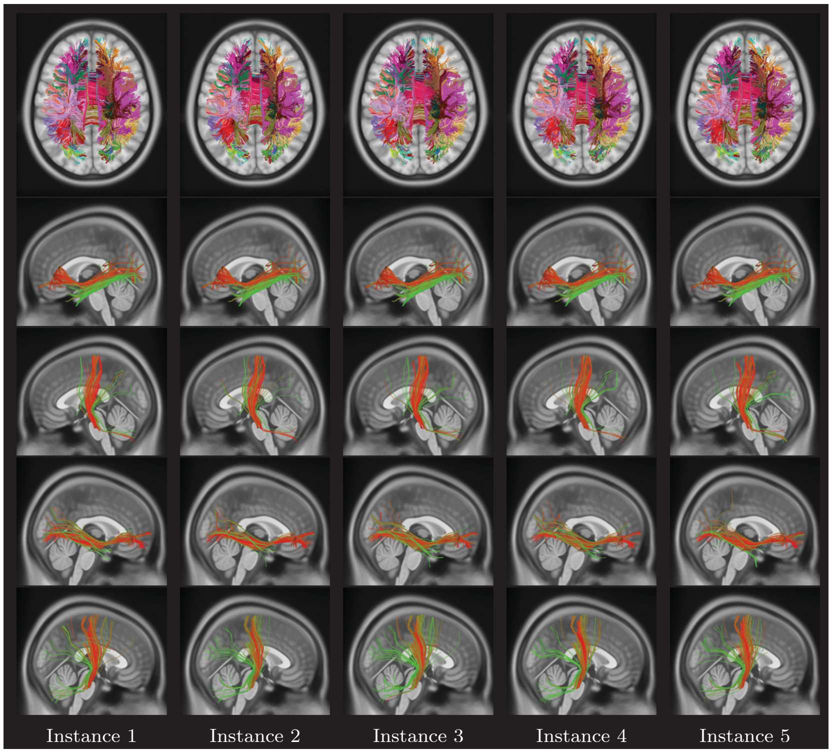


Figure 2.15 Automated segmentation of 5 instances of subject 1 using dictionary D1.

## 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. We thank the Sherbrooke Connectivity Imaging Laboratory for generously providing the labeled dataset used in this work.

# CHAPTER 3

## FIBERPRINT: A SUBJECT FINGERPRINT BASED ON SPARSE CODE POOLING FOR WHITE MATTER FIBER ANALYSIS

Kuldeep Kumar[a,*], Christian Desrosiers[a], Kaleem Siddiqi[b],
Olivier Colliot[c,d,e], Matthew Toews[a]
[a]LIVIA, École de technologie supérieure, Montreal, Canada
[b]CIM, McGill University, Montreal, Canada
[c]Sorbonne Universités, UPMC, Inserm, CNRS, ICM, Paris, France
[d]Inria Paris, Aramis project-team, Paris, France
[e]AP-HP, Dept of Neurology and Neuroradiology, Hôpital Pitié-Salpêtrière, Paris, France
Email: kkumar@livia.etsmtl.ca, christian.desrosiers@etsmtl.ca

This article was published in *NeuroImage*, Elsevier, on June 30, 2017
https://doi.org/10.1016/j.neuroimage.2017.06.083

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

## 3.2 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 (Basser *et al.*, 1994). Recent advances in dMRI acquisition protocols have lead to significant improvements in signal reconstruction (Assemlal *et al.*, 2011; Tuch, 2004; Wedeen *et al.*, 2005), 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 (Chiang *et al.*, 2011; Jahanshad *et al.*, 2012; Thompson *et al.*, 2013) is an important step in the study of the human brain, essential to understanding the mechanisms of neurological function and disease (Chung *et al.*, 2011; Griffa *et al.*, 2013; Hagmann *et al.*, 2006; Thomason and Thompson, 2011).

Over the years, several approaches have been proposed to provide a simplified quantitative description of white matter connections, to allow for cross-population inferences (Guevara *et al.*, 2012; Jin *et al.*, 2014; O'Donnell and Westin, 2007a; Prasad *et al.*, 2014). 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 (Amunts *et al.*, 2000; Mangin *et al.*, 2004; Rademacher *et al.*, 2001), function (Barch *et al.*, 2013; Grabner *et al.*, 2007; Mueller *et al.*, 2013; Ruiz-Blondet *et al.*, 2016; Rypma and D'Esposito, 1999; Zilles and Amunts, 2013), and white matter geometry (Bürgel *et al.*, 2006; de Schotten *et al.*, 2011). Motivated by this, the concept of connectome fingerprinting, which characterizes individuals using unique connectivity profiles, has recently drawn significant interest from the neuroscience community (Armstrong *et al.*, 2015; Finn *et al.*, 2015; Miranda-Dominguez *et al.*, 2014; Mišić and Sporns, 2016; Wachinger *et al.*, 2015a; Yeh *et al.*, 2016b,c).

So far, most studies on subject fingerprinting have centered around functional (Armstrong *et al.*, 2015; Finn *et al.*, 2015; Miranda-Dominguez *et al.*, 2014) and structural data (Wachinger *et al.*, 2015a; Toews and Wells, 2016). Recently, a novel approach was proposed for building individual connectome profiles based on dMRI data (Yeh *et al.*, 2016c,a). 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 (Colby *et al.*, 2012), 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 (Kumar *et al.*, 2015). 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 (Kumar and Desrosiers, 2016). 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 (Yang *et al.*, 2009) 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 (Bürgel *et al.*, 2006; de Schotten *et al.*, 2011), 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 3.4 then presents the proposed Fiberprint approach, based on non-negative kernel sparse coding. In Section 3.5, 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 3.6, we discuss our main observations and experimental findings. We conclude with a summary of our contributions and a discussion of possible extensions.

## 3.3 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.

### 3.3.1 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 (Basser *et al.*, 2000; Conturo *et al.*, 1999; Mori *et al.*, 1999). 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 (O'Donnell *et al.*, 2013). Over the years, a wide range of approaches have been proposed to cluster fiber trajectories, including methods based on hierarchical clustering (Corouge *et al.*, 2004; Gerig *et al.*, 2004) and spectral clustering (Brun *et al.*, 2004; Jonasson *et al.*, 2005; O'Donnell and Westin, 2005). Most of these methods group fiber trajectories using problem-specific measures of similarity, such as the Hausdorff distance (Corouge *et al.*, 2004; Gerig *et al.*, 2004; Moberts *et al.*, 2005) or a mean of closest points (MCP) distance (Corouge *et al.*, 2004; Ding *et al.*, 2003; Gerig *et al.*, 2004; Moberts *et al.*, 2005).

Various studies have also focused on the segmentation of white matter tracts, toward the goal of drawing cross-population inferences (Guevara *et al.*, 2012; Jin *et al.*, 2014; O'Donnell and Westin, 2007a; Prasad *et al.*, 2014). These studies either follow an atlas based approach (Guevara *et al.*, 2012; Jin *et al.*, 2014; O'Donnell and Westin, 2007a) or align specific tracts directly across subjects (Garyfallidis *et al.*, 2015; O'Donnell *et al.*, 2012). Multi-step or multi-level approaches have also been proposed to segment fiber trajectories, for example, by combining both voxel and fiber trajectory groupings (Guevara *et al.*, 2012), fusing labels from multiple hand-labeled atlases (Jin *et al.*, 2014), using a white matter voxel-space atlas and a bundle representation based on maximum density paths (Prasad *et al.*, 2014), or using Gaussian processes (Wassermann *et al.*, 2010). A few studies have also investigated the representation of specific fiber trajectory bundles using different techniques such as gamma mixture models (Maddah *et al.*, 2008), hierarchical Dirichlet processes (Wang *et al.*, 2011c), and the computational model of rectifiable currents (Durrleman *et al.*, 2009; Gori *et al.*, 2016).

### 3.3.2 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 (Elad *et al.*, 2010; Wright *et al.*, 2009, 2010; Yang *et al.*, 2009). This technique has also shown promise for various neuroimaging applications, such as the reconstruction (Lustig *et al.*, 2008) or segmentation (Tong *et al.*, 2013) of MRI data, and for functional connectivity analysis (Lee *et al.*, 2016a,b). For diffusion data, sparse coding has been used successfully for clustering white matter voxels from Orientation Density Function (ODF) data (Çetingül *et al.*, 2014), and for finding a population-level dictionary of key white matter tracts (Zhu *et al.*, 2016).

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 (Kumar *et al.*, 2015). 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 (Kumar and Desrosiers, 2016).

### 3.3.3 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 (Barch *et al.*, 2013; Bürgel *et al.*, 2006; Grabner *et al.*, 2007; Mangin *et al.*, 2004; Mueller *et al.*, 2013; Rademacher *et al.*, 2001; Ruiz-Blondet *et al.*, 2016). A recent study has shown that functional connectivity profiles act as robust and reliable fingerprints that can identify individual subjects from a large group (Finn *et al.*, 2015). 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 (Wachinger *et al.*, 2015a), 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 (Yeh *et al.*, 2016b,a). 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 (Colby *et al.*, 2012). 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.

## 3.4 Materials and methods

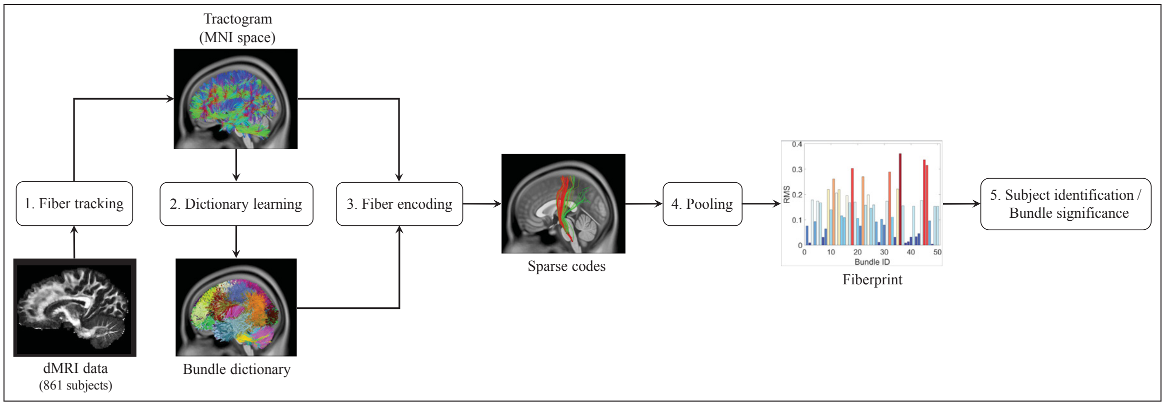


Figure 3.1 Pipeline of the proposed *Fiberprint* approach based on sparse code pooling.

Figure 3.1 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 preprocessed dMRI data of 861 subjects from the Human Connectome Project (Van Essen *et al.*, 2012, 2013). 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.

### 3.4.1 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 (Glasser *et al.*, 2013; Van Essen *et al.*, 2012, 2013), 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 $^{s}/_{mm^2}$, 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[1].

For signal reconstruction and tractography, we used the freely available DSI Studio toolbox. All subjects were reconstructed in MNI space using the Q-space diffeomorphic reconstruction (QSDR) (Yeh and Tseng, 2011) option in DSI Studio. QSDR is an extension of generalized q-sampling imaging (GQI, (Yeh *et al.*, 2010)), 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 (Ashburner, 2000). 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

---

[1] http://www.humanconnectome.org/documentation/Q3/

provided with pre-processed 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 (Basser *et al.*, 2000; Yeh *et al.*, 2013), 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 (Jones *et al.*, 2013; Thomas *et al.*, 2014). 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.

### 3.4.2 Learning the fiber trajectory dictionary

Out of the 861 available subjects, 10 unrelated ones (O'Donnell *et al.*, 2017) 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 (Kumar *et al.*, 2015; Kumar and Desrosiers, 2016), which we now summarize.

Let $\boldsymbol{X}$ 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 $\boldsymbol{x}_i \in \boldsymbol{R}^d$, and that $\boldsymbol{X}$ 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 $\boldsymbol{D}$. Formally, we write this as $\boldsymbol{x}_i \sim \boldsymbol{D}\boldsymbol{w}_i$, where $\boldsymbol{w}_i$ 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 $\boldsymbol{\phi} : \boldsymbol{R}^d \rightarrow \boldsymbol{R}^q$, such that $\boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\phi}(\boldsymbol{x}') = k(\boldsymbol{x}, \boldsymbol{x}')$ 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 (Corouge *et al.*, 2004; Gerig *et al.*, 2004; Moberts *et al.*, 2005), the mean of closest points (MCP) distance (Corouge *et al.*, 2004; Ding *et al.*, 2003; Gerig *et al.*, 2004; Moberts *et al.*, 2005) or the Minimum average Direct Flip (MDF) distance (Garyfallidis *et al.*, 2012). 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. $k(\boldsymbol{x}, \boldsymbol{x}') = \exp\big(-\gamma \cdot \mathrm{dist}_{\mathrm{MDF}}(\boldsymbol{x}, \boldsymbol{x}')\big)$. The fiber trajectories were sampled to $15$ equidistant points for distance computation (Garyfallidis *et al.*, 2012) and the kernel bandwidth parameter was set empirically to $\gamma = 0.0001$.

Using $\boldsymbol{\Phi} \in \boldsymbol{R}^{q \times n}$ to denote the matrix of mapped training fiber trajectories, the kernel matrix of pairwise similarities then corresponds to $\boldsymbol{K} = \boldsymbol{\Phi}^\top \boldsymbol{\Phi}$. Using the idea proposed in (Nguyen *et al.*, 2012), we express the dictionary as a non-negative linear combination of training examples, i.e., $\boldsymbol{D} \sim \boldsymbol{\Phi}\boldsymbol{A}$, and formulate the dictionary learning task as the following optimization problem:

$$\operatorname*{arg\,min}_{\boldsymbol{A}, \boldsymbol{W} \geq 0} \frac{1}{2}\|\boldsymbol{\Phi} - \boldsymbol{\Phi}\boldsymbol{A}\boldsymbol{W}\|_F^2 \quad \text{s.t.} \ \ \|\boldsymbol{w}_i\|_0 \leq S_{\max}, \ \ i = 1, \ldots, n, \tag{3.1}$$

where $\|\boldsymbol{w}_i\|_0$ is the $L_0$ norm (i.e., number of non-zero elements) of $\boldsymbol{w}_i$, constraining each fiber trajectory to be encoded using at most $S_{\max}$ prototypes, $\boldsymbol{A} \in \boldsymbol{R}^{n\times m}$ is the dictionary coefficient matrix, and $\boldsymbol{W} \in \boldsymbol{R}^{m\times n}$ is the sparse code matrix for all fiber trajectories. When $S_{\max} = 1$, this formulation corresponds to the kernel K-means problem (Dhillon *et al.*, 2004). As shown in Section 3.5.1.4, 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 (Kumar *et al.*, 2015), which updates the sparse codes $\boldsymbol{W}$ and dictionary matrix $\boldsymbol{A}$ iteratively, until convergence. In the sparse coding step, each column of $\boldsymbol{W}$ is updated independently by optimizing the following sub-problem:

$$\operatorname*{arg\,min}_{\boldsymbol{w}_i \geq 0} \ \frac{1}{2}\boldsymbol{w}_i{}^{\top}\boldsymbol{A}^{\top}\boldsymbol{K}\boldsymbol{A}\boldsymbol{w}_i \ - \ k_i^{\top}\boldsymbol{A}\boldsymbol{w}_i \quad \text{s.t.} \ \ \|\boldsymbol{w}_i\|_0 \leq S_{\max}, \tag{3.2}$$

where $k_i \in \boldsymbol{R}^n$ 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 (Kumar *et al.*, 2015). The dictionary matrix $\boldsymbol{A}$ is then obtained using a kernel version of the non-negative matrix tri-factorization approach proposed in (Ding *et al.*, 2006), which applies the following update scheme until convergence:

$$\boldsymbol{A}_{ij} \ \leftarrow \ \boldsymbol{A}_{ij} \cdot \frac{\left(\boldsymbol{K}\boldsymbol{W}^{\top}\right)_{ij}}{\left(\boldsymbol{K}\boldsymbol{A}\boldsymbol{W}\boldsymbol{W}^{\top}\right)_{ij}}, \quad \forall\, i,j. \tag{3.3}$$

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 $\boldsymbol{A}$.

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 (Fowlkes *et al.*, 2004; O'Donnell and Westin, 2007a). 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.

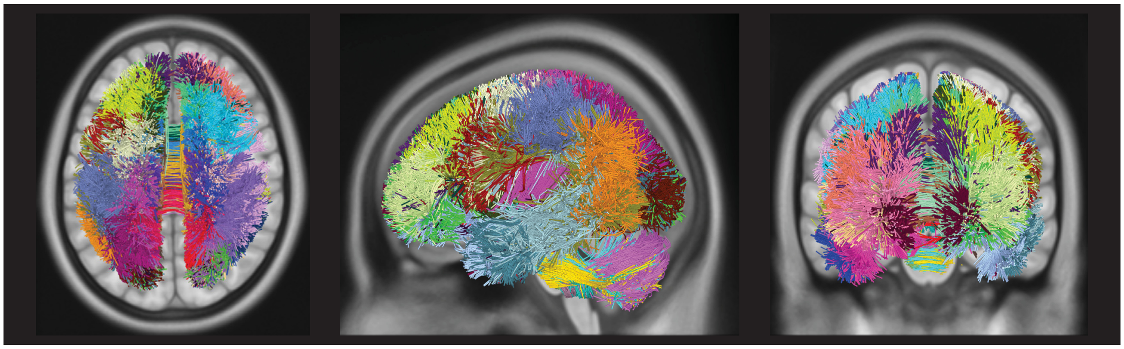


Figure 3.2 Dictionary visualization. Visualization of $m = 50$ fiber trajectory prototypes learned from $10$ subjects, with an unique color assigned to each dictionary prototype. For this simplified visualization each fiber trajectory is assigned to a single prototype by taking the maximum for each row of the matrix $\boldsymbol{A}$. (superior axial, left sagittal, and anterior coronal views respectively)

Figure 3.2 gives a qualitative visualization of $m = 50$ fiber trajectory prototypes learned in the dictionary (the impact of parameter $m$ is analyzed in Section 3.5.1.2), 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[2]. 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 presenta-

---

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

tion, we use the term bundle dictionary when referring to the output of the dictionary learning step.

### 3.4.3 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. (3.2), 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 3.2, are illustrated in Figure 3.3. We represent bundles using the same colors as in Figure 3.2, 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 (Kumar and Desrosiers, 2016). 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 3.4, 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.

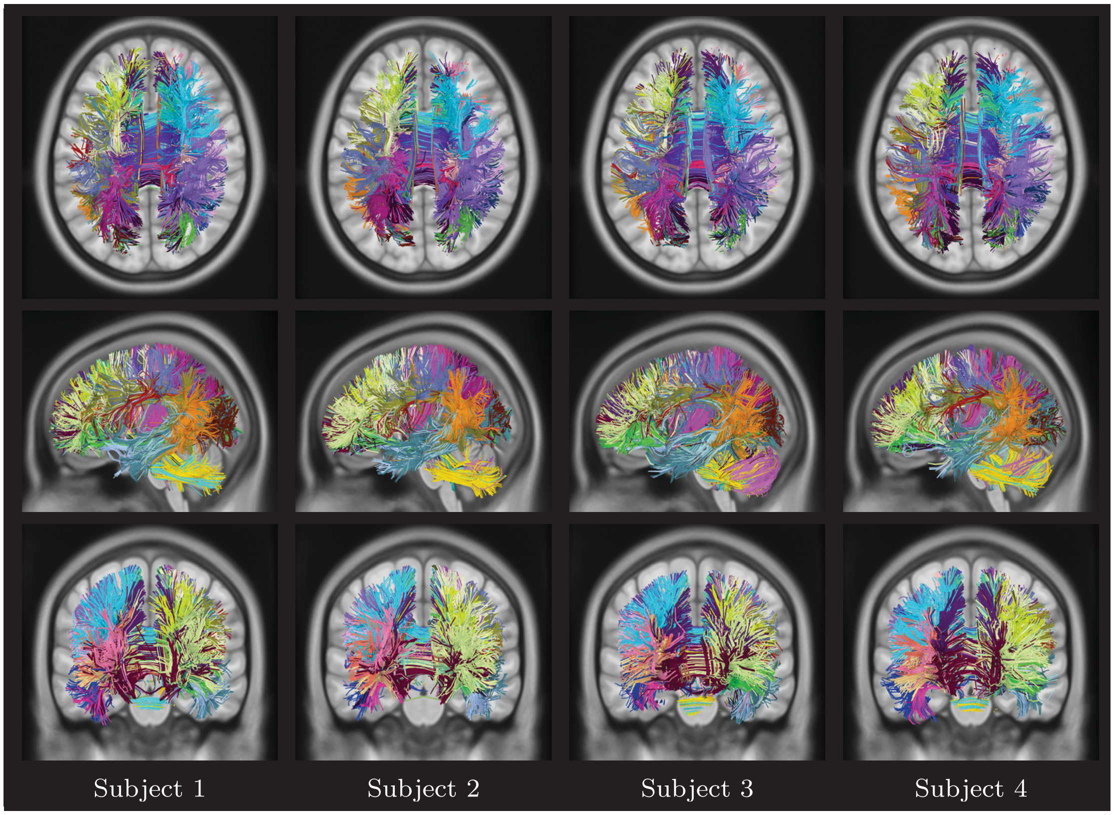


Figure 3.3 Visualization of sparse code representation of fiber trajectories from four subjects. Each fiber trajectory is assigned to a single bundle by taking the maximum of the sparse code vector. Bundles are represented using the same colors as in Figure 3.2. (superior axial (top), left sagittal (middle), and anterior coronal (bottom) views respectively)

### 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 achieved using a sparse code pooling function (Yang *et al.*, 2009) that combines, for each dictionary bundle, the relationship between this bundle and all fiber trajectories of a subject into a single value. Let $\boldsymbol{W} \in \boldsymbol{R}^{m \times n}$ be the sparse code matrix obtained in the previous step, each column corresponding to a different fiber trajectory of the

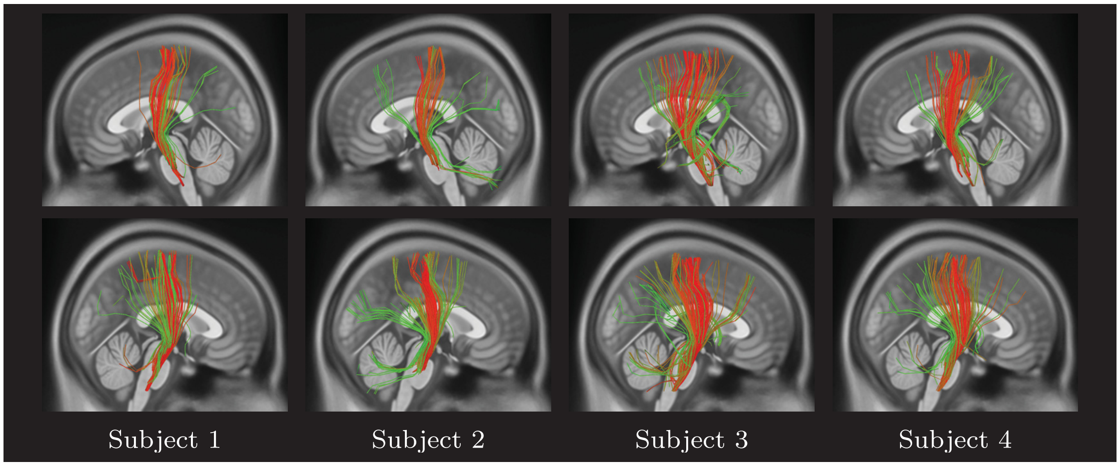


Figure 3.4 Color coded visualization of sparse code memberships of fiber trajectories for the left (top row) and right (bottom row) corticospinal bundles from four subjects. Green and red represent, a low and a high membership of a fiber trajectory to a bundle, respectively. Fiber trajectories with a zero membership to the bundle are removed for a simplified visualization.

subject to encode. We consider three pooling functions frequently used in the literature, based on the root mean square (RMS), mean and maximum:

$$\big[f_{\text{RMS}}(W)\big]_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n} w_{ji}^2} \tag{3.4}$$

$$\big[f_{\text{Mean}}(W)\big]_j = \frac{1}{n}\sum_{i=1}^{n} |w_{ji}| \tag{3.5}$$

$$\big[f_{\text{Max}}(W)\big]_j = \max\big\{|w_{j1}|, |w_{j2}|, \ldots, |w_{jn}|\big\}. \tag{3.6}$$

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 $f_{\text{mean}}$ computes the average sparse code value of fiber trajectories belonging to a bundle, thus giving an estimate of the bundle's density. $f_{\text{RMS}}$ 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 3.5 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 3.6, 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.

## 3.5 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. 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).

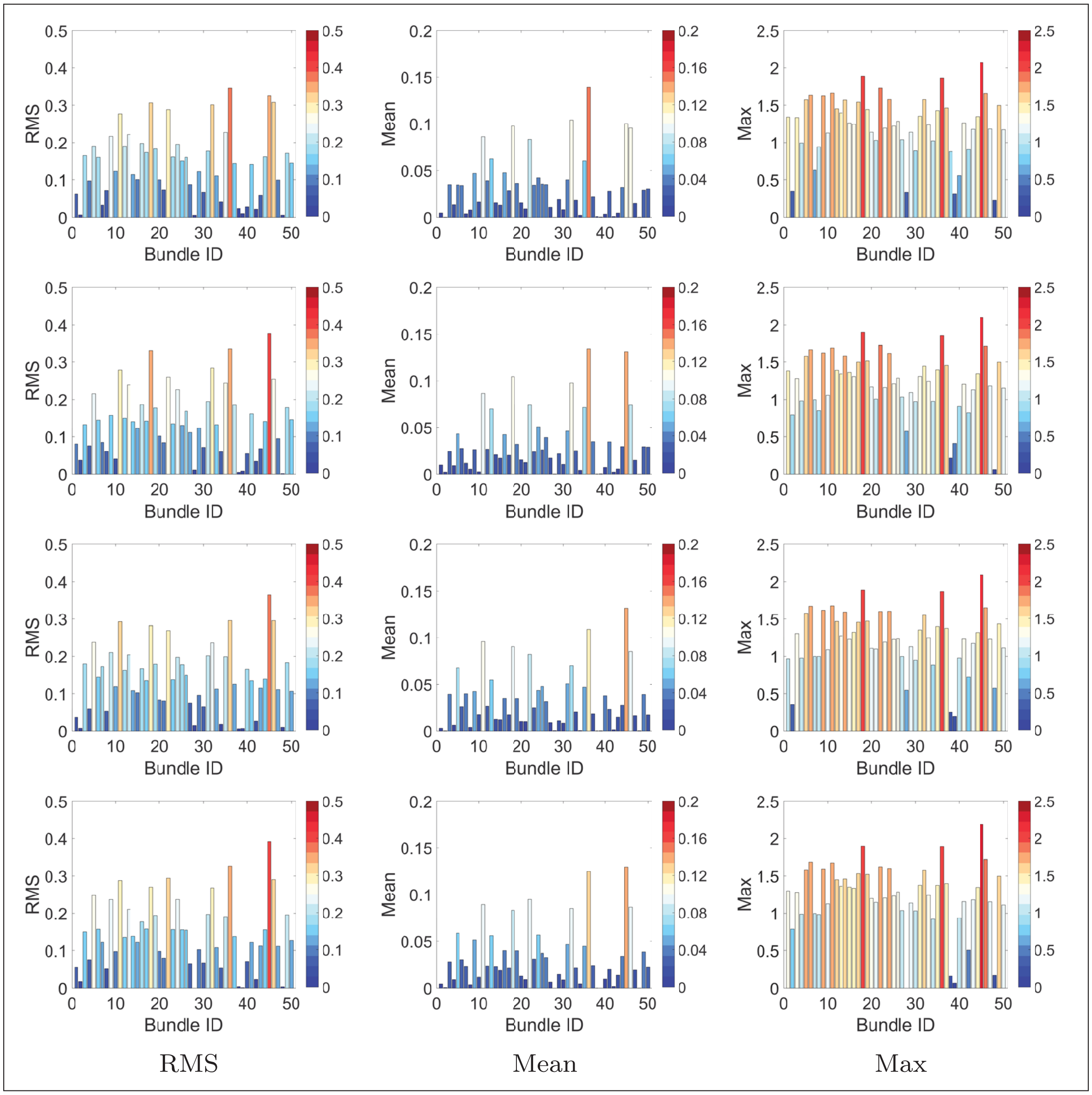


Figure 3.5 Subject fingerprint visualization. Color coded bar plot representation for four subjects (rows) and three pooling functions (RMS, Mean, and Max; columns), plotted as a value per bundle ID.

### 3.5.1 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

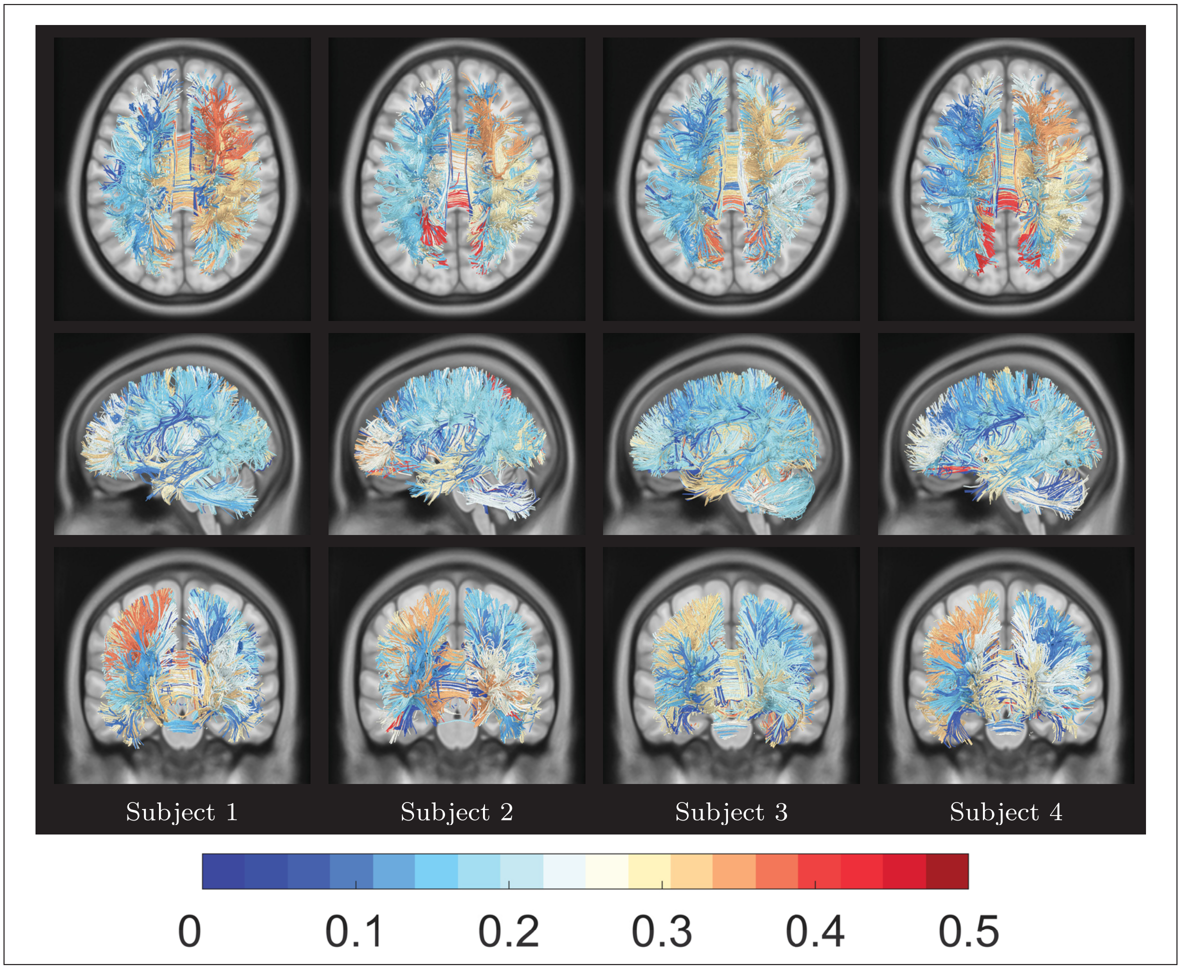


Figure 3.6 Subject fingerprint visualization. Color coded bundles from four subjects representing the magnitude of their corresponding RMS pooling function values. We use the same color code scheme as in Figure 3.5. (superior axial (top), left sagittal (middle), and anterior coronal (bottom) views respectively)

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 3.4.3, 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 (Gale and Perkel, 2010) is used to obtain a quantitative measure of separability:

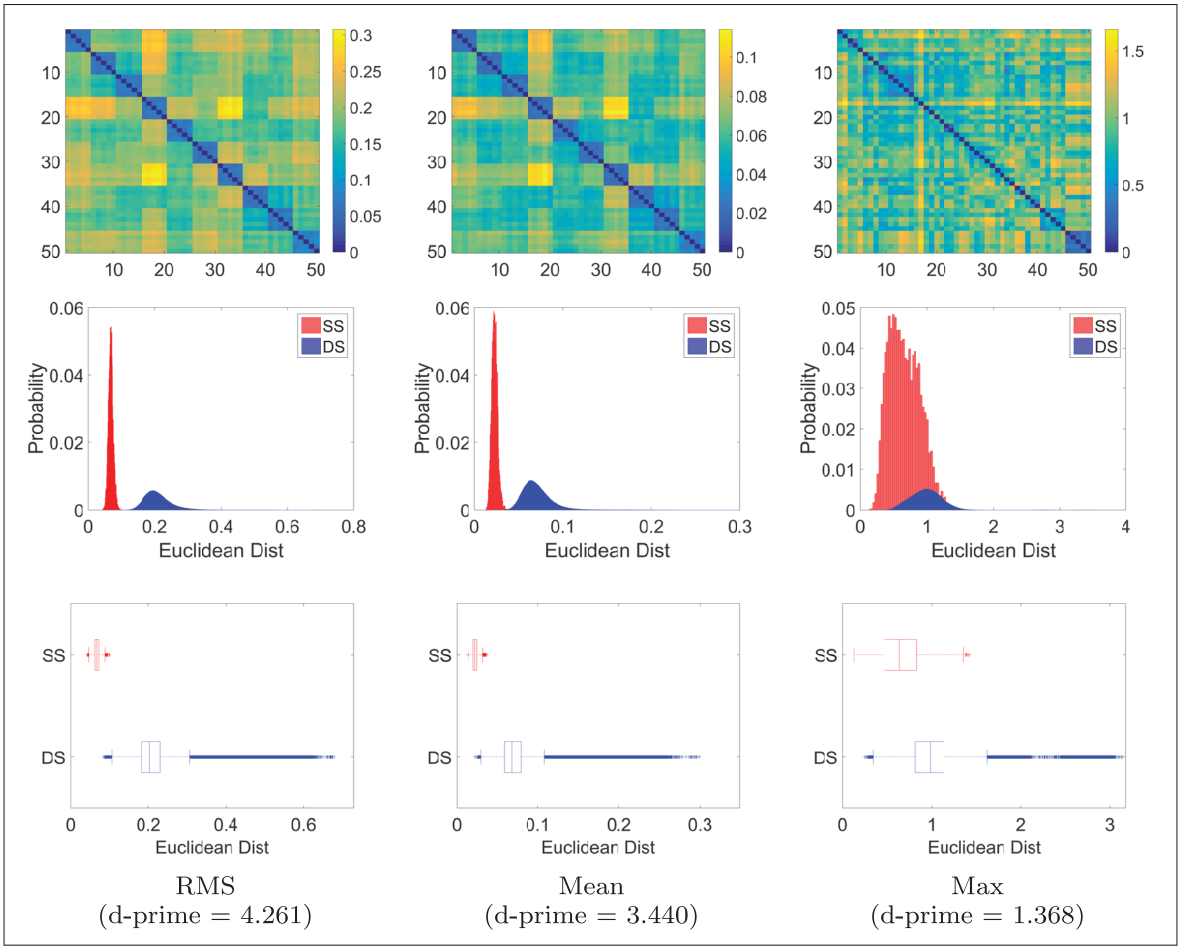
$$\text{d-prime} \;=\; \frac{\mu_1 - \mu_2}{\sqrt{\frac{1}{2}\left(\sigma_1^2 \;+\; \sigma_2^2\right)}}, \tag{3.7}$$

where, $\mu_1, \mu_2$ are the means and $\sigma_1, \sigma_2$ the standard deviations of the compared distributions. Higher d-prime values indicate better separability. In this work we report absolute value of d-prime.

#### 3.5.1.1 Pooling function

The impact of the pooling function on the fingerprint's ability to distinguish subjects is analyzed in Figure 3.7. 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., $5 \times 5$ 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



RMS
(d-prime = 4.261)

Mean
(d-prime = 3.440)

Max
(d-prime = 1.368)

Figure 3.7 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)

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.

#### 3.5.1.2 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.

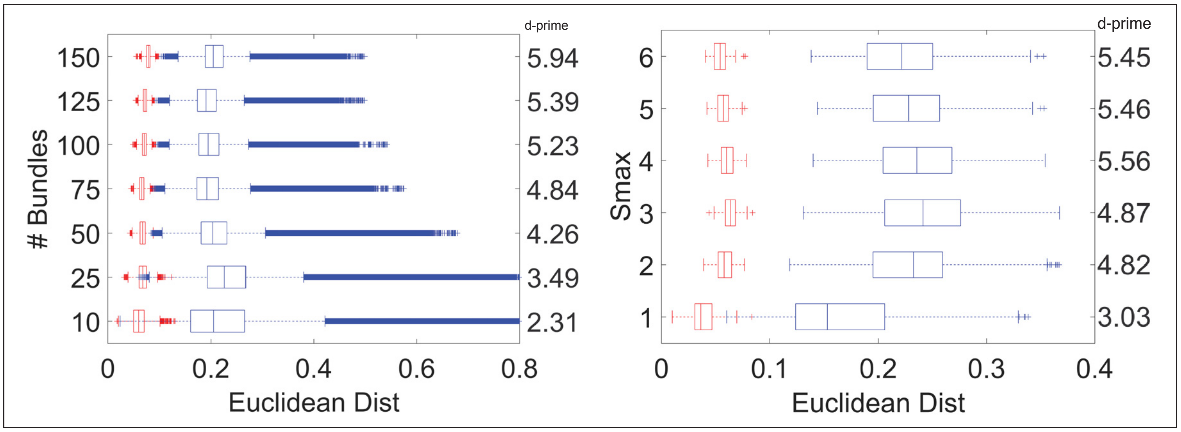


Figure 3.8 Impact of the size of the dictionary and the level of sparsity $S_{\max}$ on subject fingerprint. Box plot of Euclidean distances between same-subject (red) and different-subject (blue) instances for seven different dictionary sizes using all $851$ subjects (left); and for varying level of the sparsity parameter $S_{\max}$ using $10$ subjects (right).

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 3.8 (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 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.

#### 3.5.1.3 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 3.9 (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.

#### 3.5.1.4 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 3.8 (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.

#### 3.5.1.5 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 (Basser *et al.*, 2000; Yeh *et al.*, 2013)), 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 3.9 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 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 (Garyfallidis *et al.*, 2012; O'Donnell *et al.*, 2017), 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 (Garyfallidis *et al.*, 2012). 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 dis-

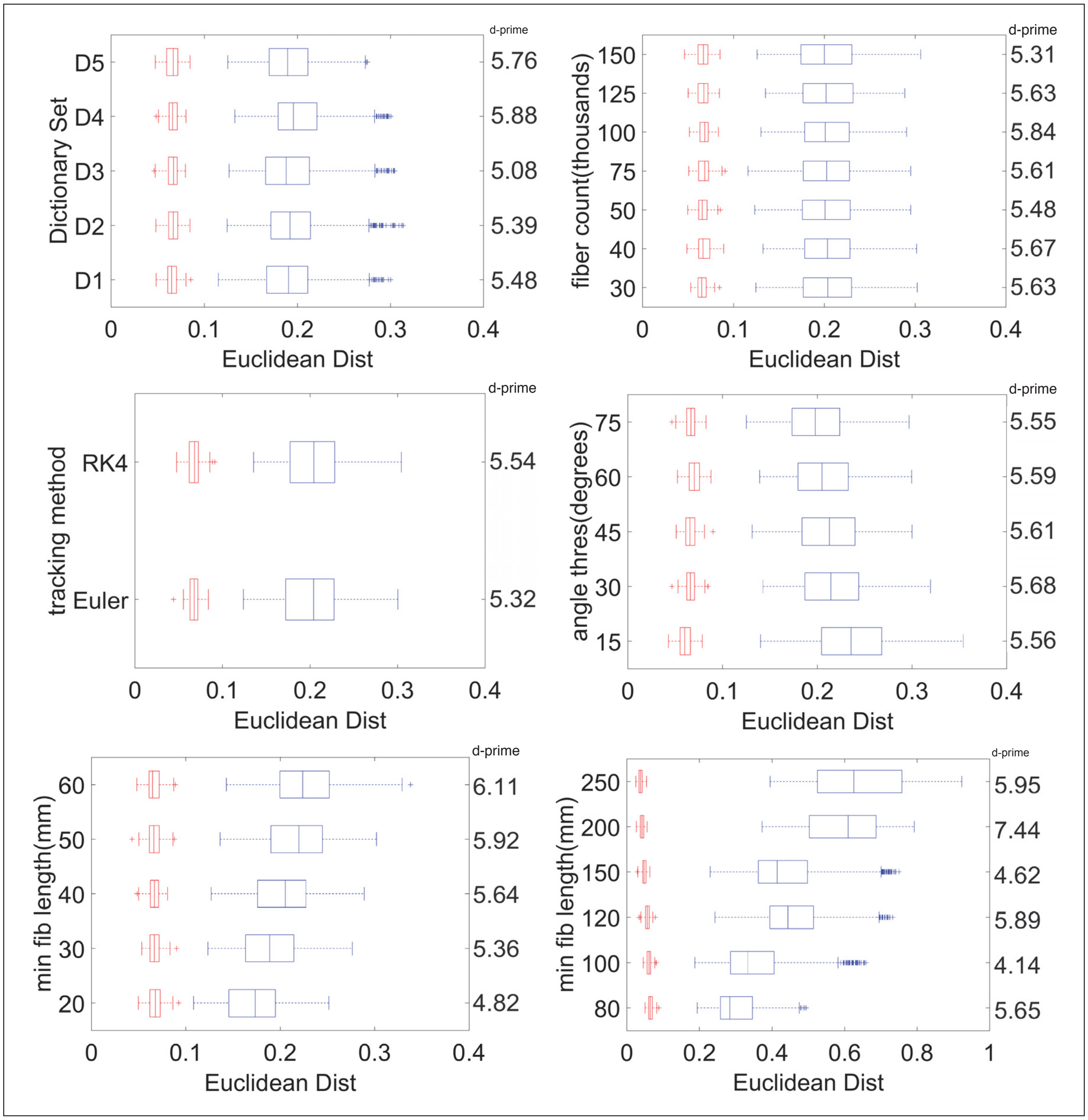


Figure 3.9 Impact of independent dictionary sets and fiber tracking parameters on subject fingerprints. Box plots of Euclidean distances between same-subject (red) and different-subject (blue) instances using 10 subjects for: independent sets of dictionaries; the number of output fiber trajectories; 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)

tances. 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.

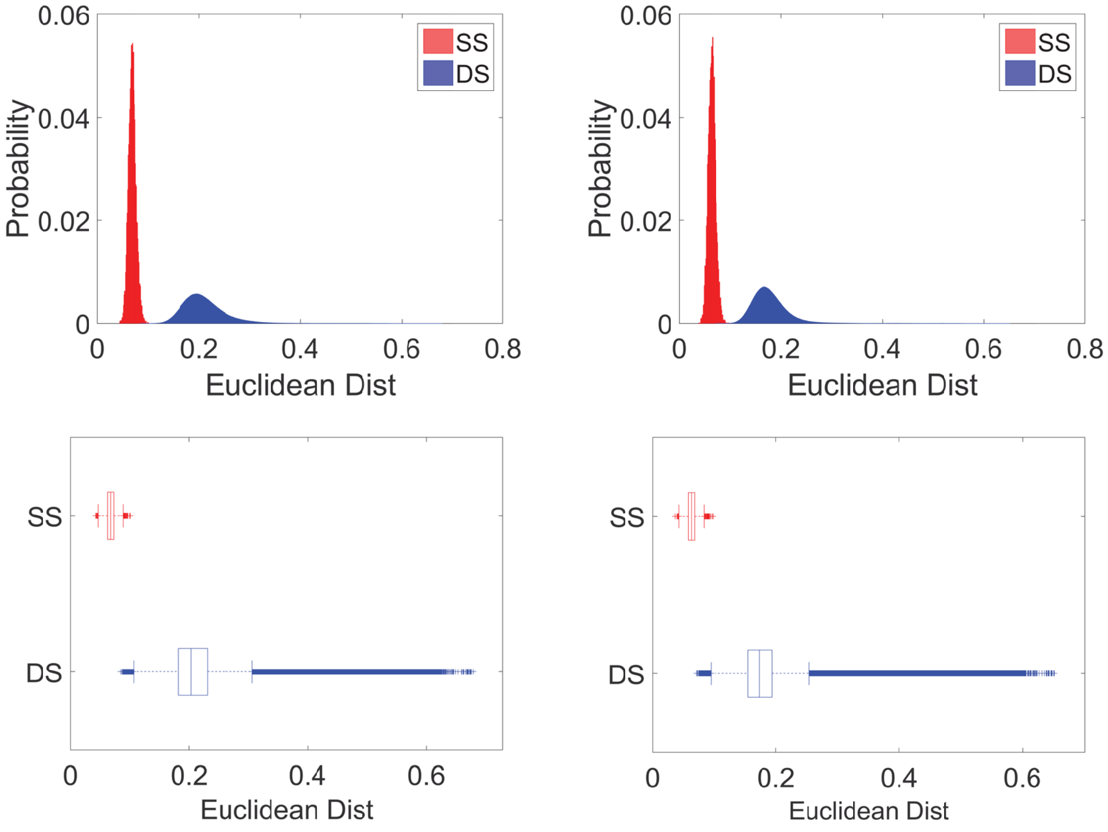
#### 3.5.1.6 Inclusion of cerebellum

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 3.10 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.

#### 3.5.1.7 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 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 3.11 (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



With cerebellum
(d-prime $= 4.347$)

Without cerebellum
(d-prime $= 3.995$)

Figure 3.10 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.

a discriminative fingerprint (for a dictionary size of $m = 50$). To understand how the number of fiber trajectories affects the fingerprint, Figure 3.11 (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.

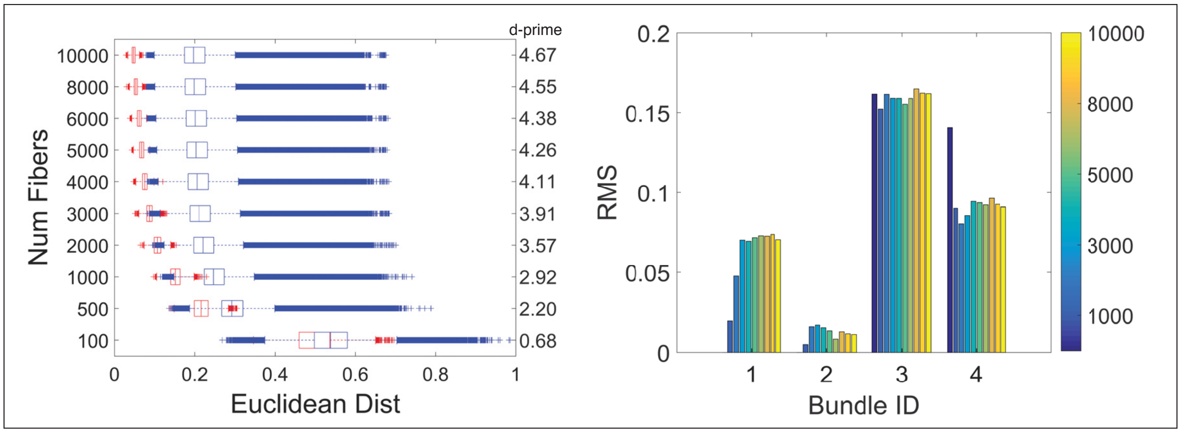


Figure 3.11 Impact of the number of fiber trajectories used to generate a subject fingerprint. Box plot for Euclidean distances between same-subject (red) and different-subject (blue) instances for all 851 subjects (left). Bar plot of RMS pooled features corresponding to four different bundles of a subject, obtained with varying numbers of fiber trajectories (right).

### 3.5.2 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:

$$\text{precision@}k \;=\; \frac{|T \cap S_k|}{k}, \quad \text{recall@}k \;=\; \frac{|T \cap S_k|}{|T|}. \tag{3.8}$$

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

#### 3.5.2.1 Same subject identification

Table 3.1 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 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

Table 3.1 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 |

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 [3] (MNI space, 1mm resolution, similar to the QSDR reconstruction output) using FSL (Jenkinson *et al.*, 2012) 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 3.4.3 for details). Note that the same dictionary as in previous experiments was employed for obtaining these fingerprints.

Figure 3.12 (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 3.13, 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 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 3.12 (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 $k = 10$ near-

---

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

est neighbors, further demonstrating the robustness of our fiberprint to alignment and signal reconstruction methods.

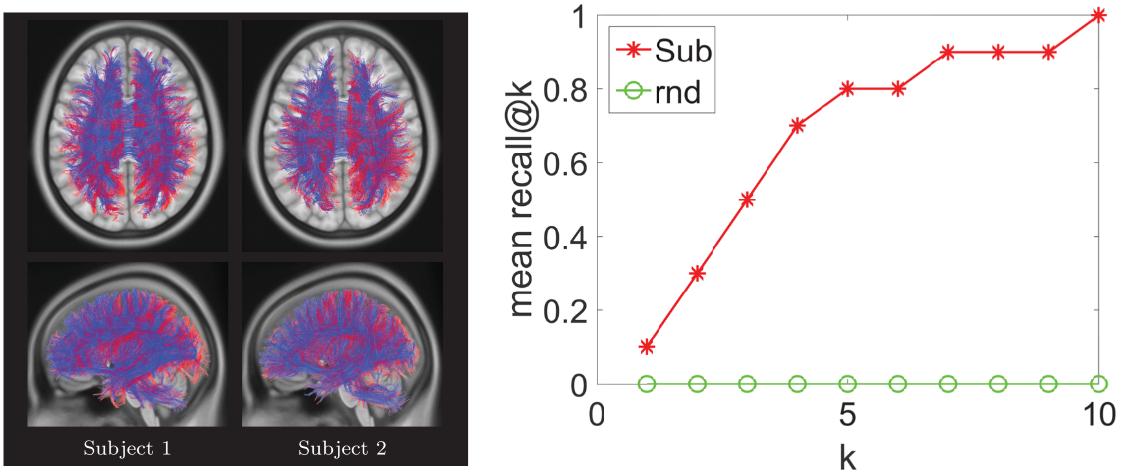


Figure 3.12 Comparison of QSDR+rigid alignment (blue) and DTI+affine alignment (red) based tractographies for subject 1 and subject 2 (left). Mean recall@k for DTI+affine alignment based fiberprint identification using 851 QSDR+rigid alignment fiberprints (right)

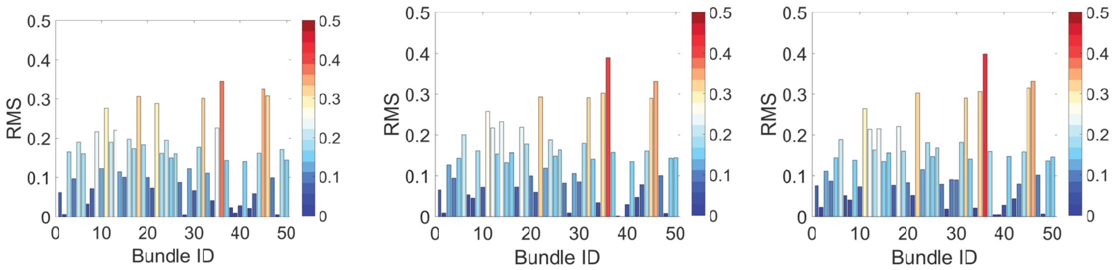


Figure 3.13 Color-coded bar plot representation of a subject's fiberprint, compared across the different alignment and signal reconstruction methods. Column 1 is a fiberprint based on QSDR and rigid alignment (Figure 3.5); columns 2 and 3 show fiberprint instances obtained with DTI and affine alignment.

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

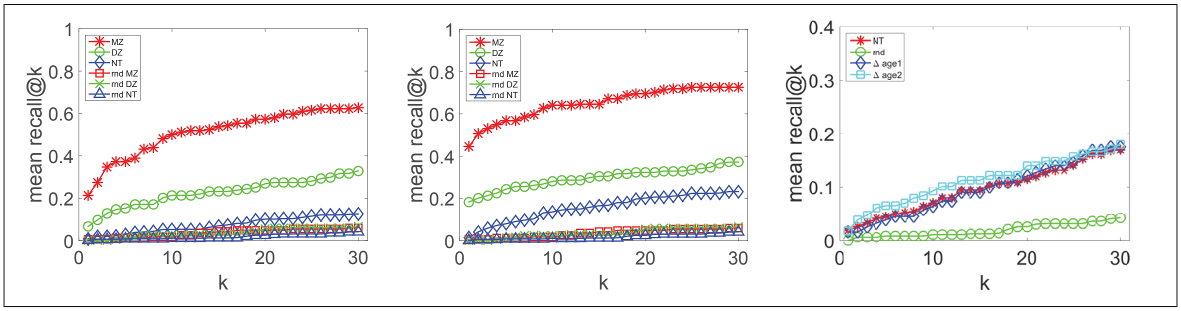


Figure 3.14 Genetically-related subject identification. The mean recall@k for MZ-twin ($82$-pairs), DZ-twin ($82$-pairs), Non-Twin siblings ($166$ pairs) using Fiberprint (left) and full T1w images rigidly aligned to MNI space as fingerprint (middle). The age difference impact on Non-Twin sibling identification, 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 3.14 (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. 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 3.14 (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 3.15 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.

#### 3.5.2.3 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 3.14 (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 T1-weighted images, is related to genetic proximity and

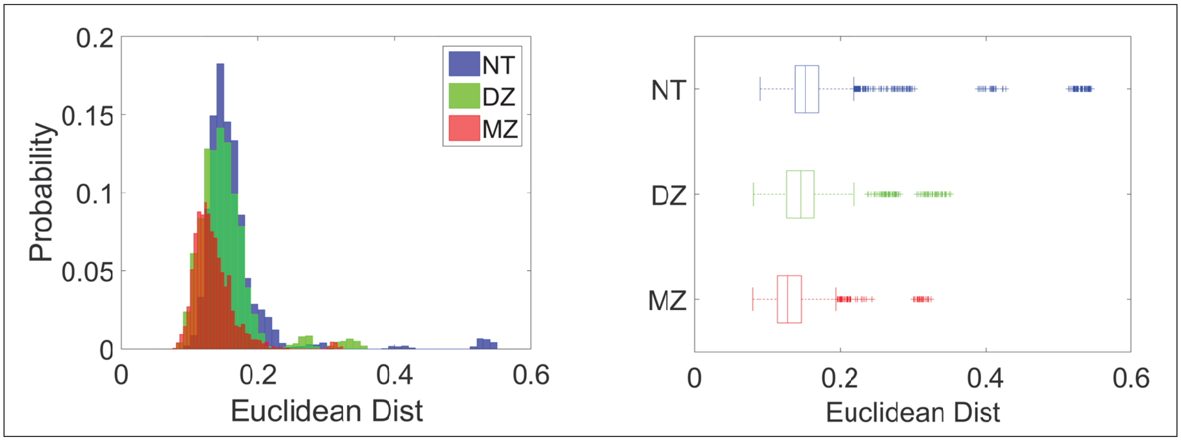


Figure 3.15 Differences between fingerprints of genetically-related subjects. Probability normalized histogram and box plot of Euclidean distances between instances belonging to MZ, DZ, and Non-Twin siblings

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 3.14 and considered the identification of a sibling as successful if this sibling's fingerprint is found within the $k = 30$ nearest neighbors.

Table 3.2 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 3.2 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** | $11.35\%$ | $19.81\%$ | $14.51\%$ | $54.33\%$ |

### 3.5.3 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[4], 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 (Holm, 1979) and consider as significant the bundles with corrected $p < 0.05$.

---

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

#### 3.5.3.1 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 3.16 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 3.17.

#### 3.5.3.2 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: $28.05 \pm 3.65$) and $436$ females (age: $29.33 \pm 3.55$), 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 3.18 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.

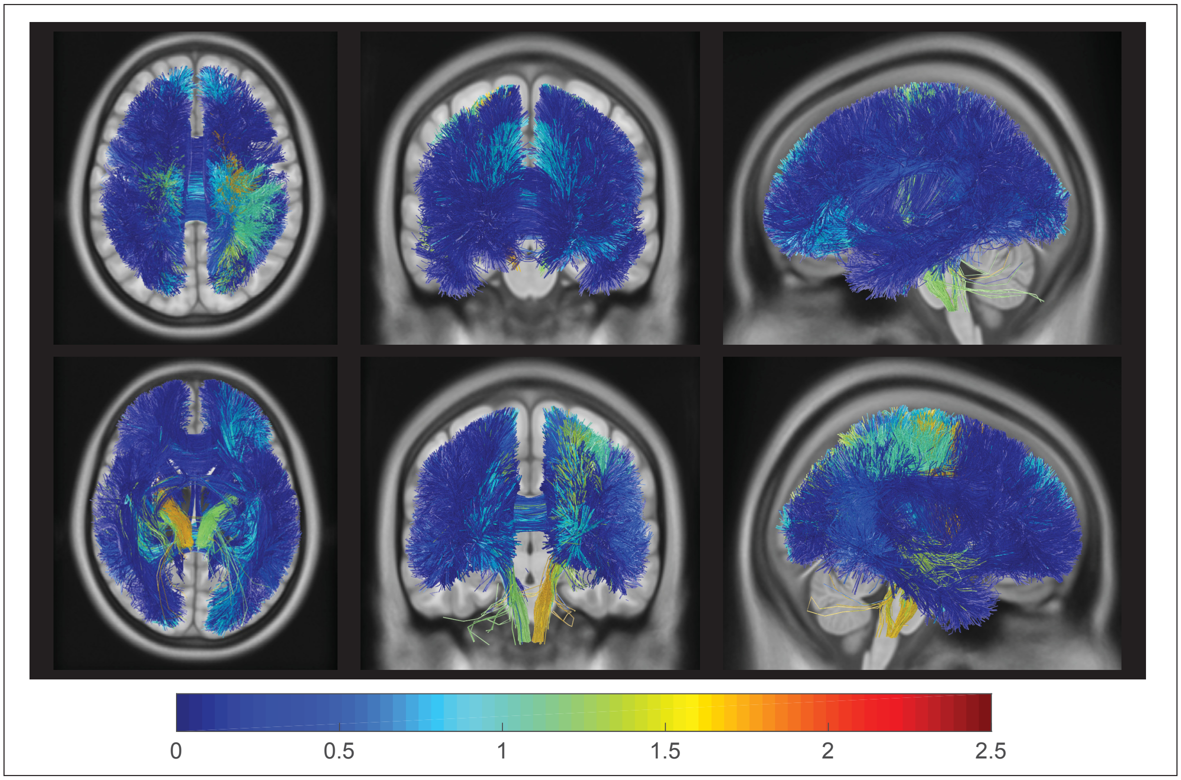


Figure 3.16 **MZ vs NT**. Differences between MZ-twin and Non-Twin siblings. Color coded bundle visualization for Holm-Bonferroni corrected p-values (in $-\log_{10}$ scale) obtained using a Wilcoxon rank-sum test. (superior axial, anterior coronal, and left sagittal views (top row); inferior axial, posterior coronal, and right sagittal views (bottom row);)

## 3.6 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.

### 3.6.1 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

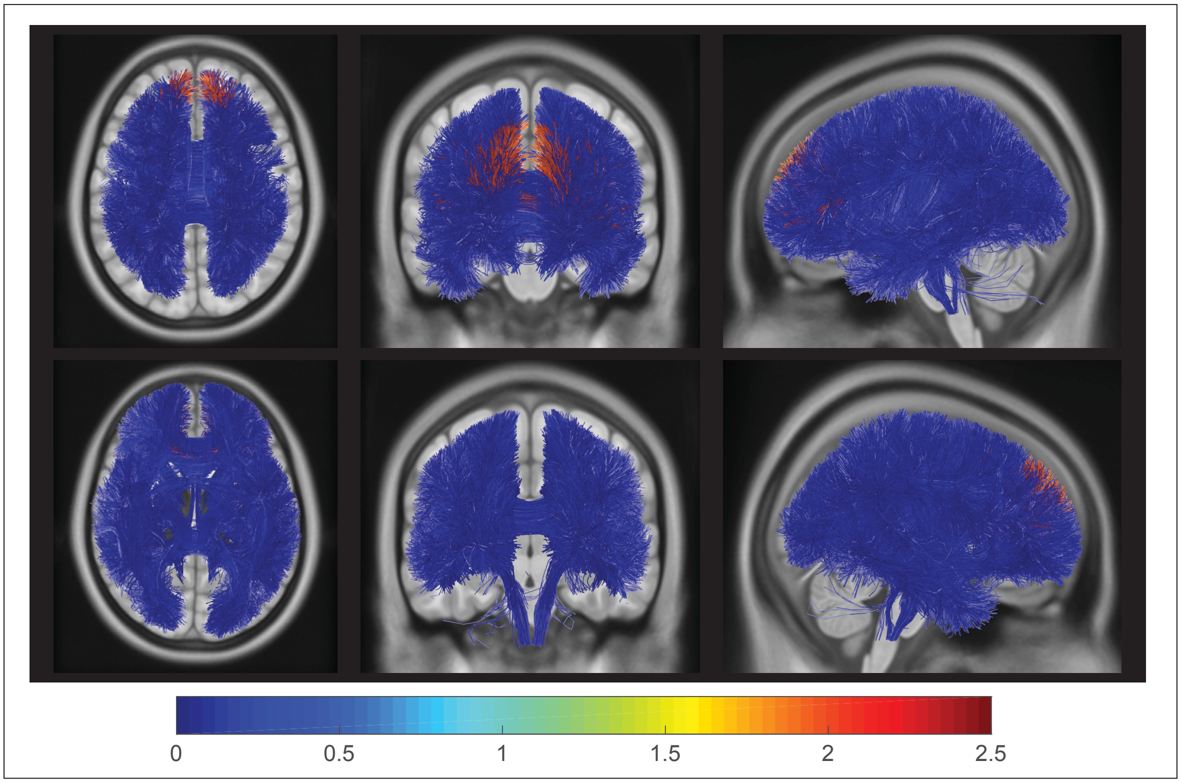


Figure 3.17 **DZ vs NT**. Differences between DZ-twin and Non-Twin siblings. Color coded bundle visualization for Holm-Bonferroni corrected p-values (in -$\log_{10}$ scale) obtained using a Wilcoxon rank-sum test. (superior axial, anterior coronal, and left sagittal views (top row); inferior axial, posterior coronal, and right sagittal views (bottom row);)

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 $50$ bundles or more, a number consistent with previous studies on the topic of fiber trajectory clustering and segmentation (Guevara *et al.*, 2012; O'Donnell and Westin, 2007a). 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

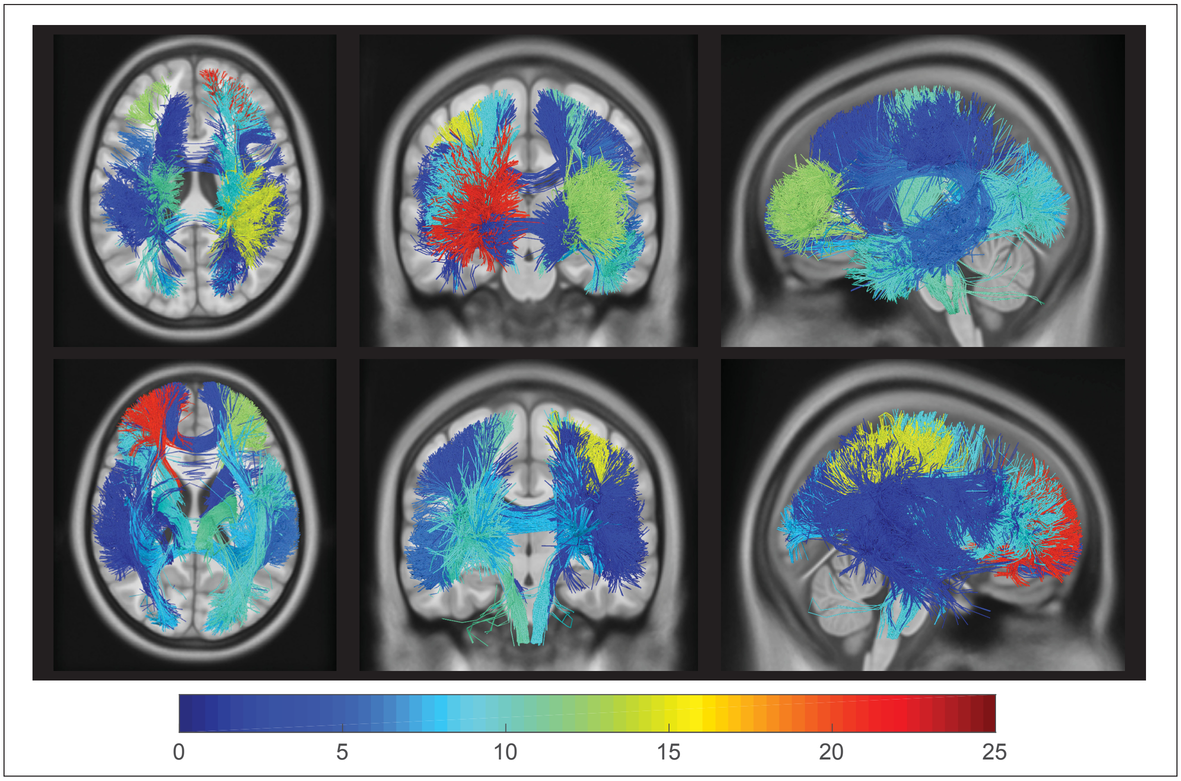


Figure 3.18 **Male vs Female**. Differences related to gender. Color coded bundle visualization for Holm-Bonferroni corrected p-values (in $-\log_{10}$ scale) obtained using a Wilcoxon rank-sum test. (superior axial, anterior coronal, and left sagittal views (top row); inferior axial, posterior coronal, and right sagittal views (bottom row);) **Note**: for visualization purposes, fibers in non-signifcant bundles are not shown.

was also found to be optimal in our previous work on fiber trajectory segmentation (Kumar and Desrosiers, 2016).

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 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 (Mangin *et al.*, 2004), function (Barch *et al.*, 2013; Mueller *et al.*, 2013), and white matter micro-structure (Bürgel *et al.*, 2006; de Schotten *et al.*, 2011).

### 3.6.2 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 (Kochunov *et al.*, 2015). 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 (Chiang *et al.*, 2011), 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 (Kochunov *et al.*, 2011, 2015; Van Essen *et al.*, 2012). 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.

### 3.6.3 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) (Kochunov *et al.*, 2015).

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) (Chou *et al.*, 2011; Gong *et al.*, 2011) or structural network analysis (Yan *et al.*, 2011).

### 3.6.4 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 (Colby *et al.*, 2012). 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.

### 3.6.5 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 (Jones *et al.*, 2013), 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 (Yeh *et al.*, 2016b) 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.

### 3.7 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.

## 3.8 Supplement results

### 3.8.1 Dictionary bundle visualization

We observed bundles from the dictionary, and assigned the hemisphere label to a bundle based on where majority of fiber trajectories lie. We had 6 bundles in the cerebellum (Figure 3.22), 12 inter-hemispheric bundles (Figure 3.19), 17 bundles in the left hemisphere (Figure 3.20), and 15 bundles in the right hemisphere (Figure 3.21).

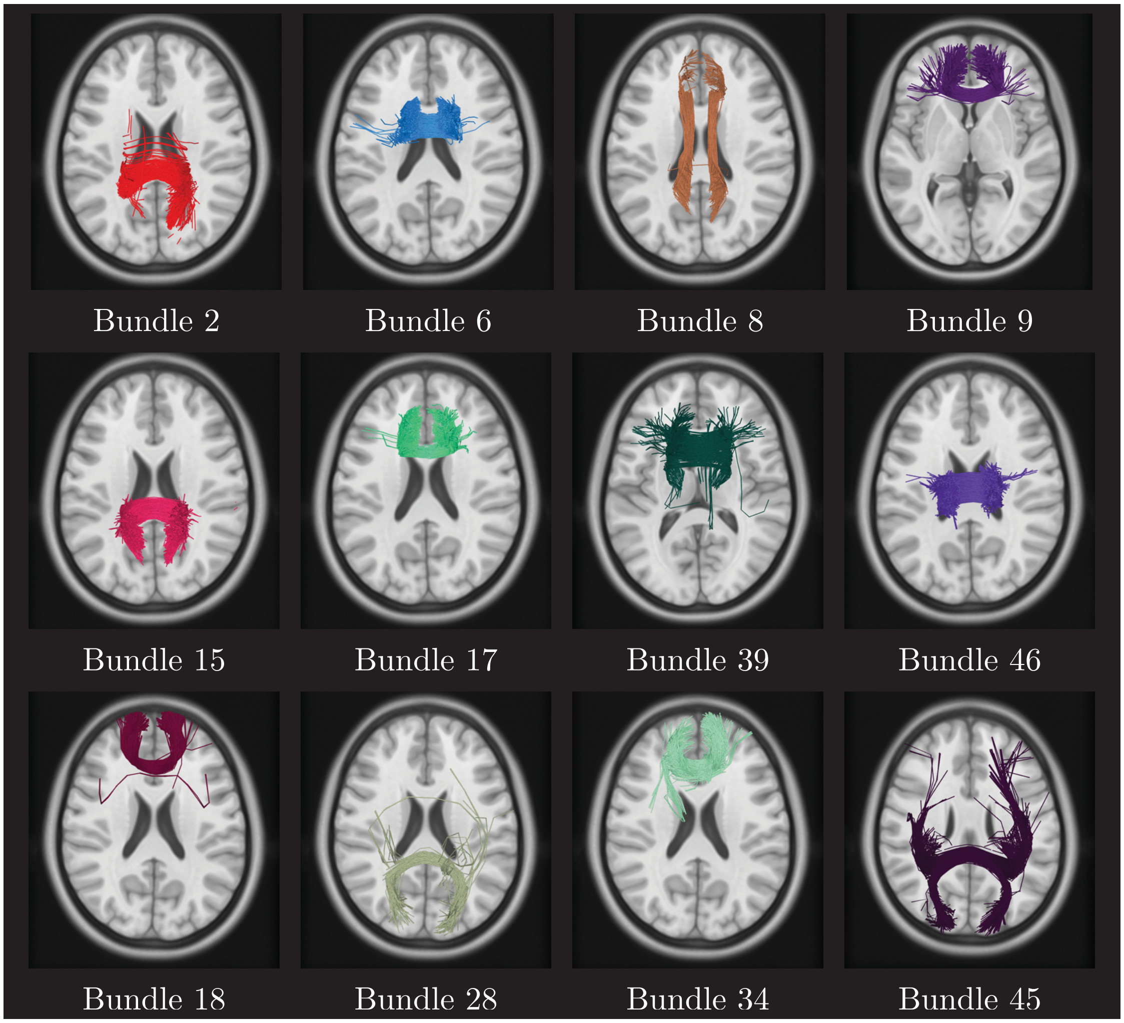


Figure 3.19 Inter-hemispheric dictionary bundles with respective bundle IDs. (Top two rows show superior axial view, bottom row shows inferior axial view)

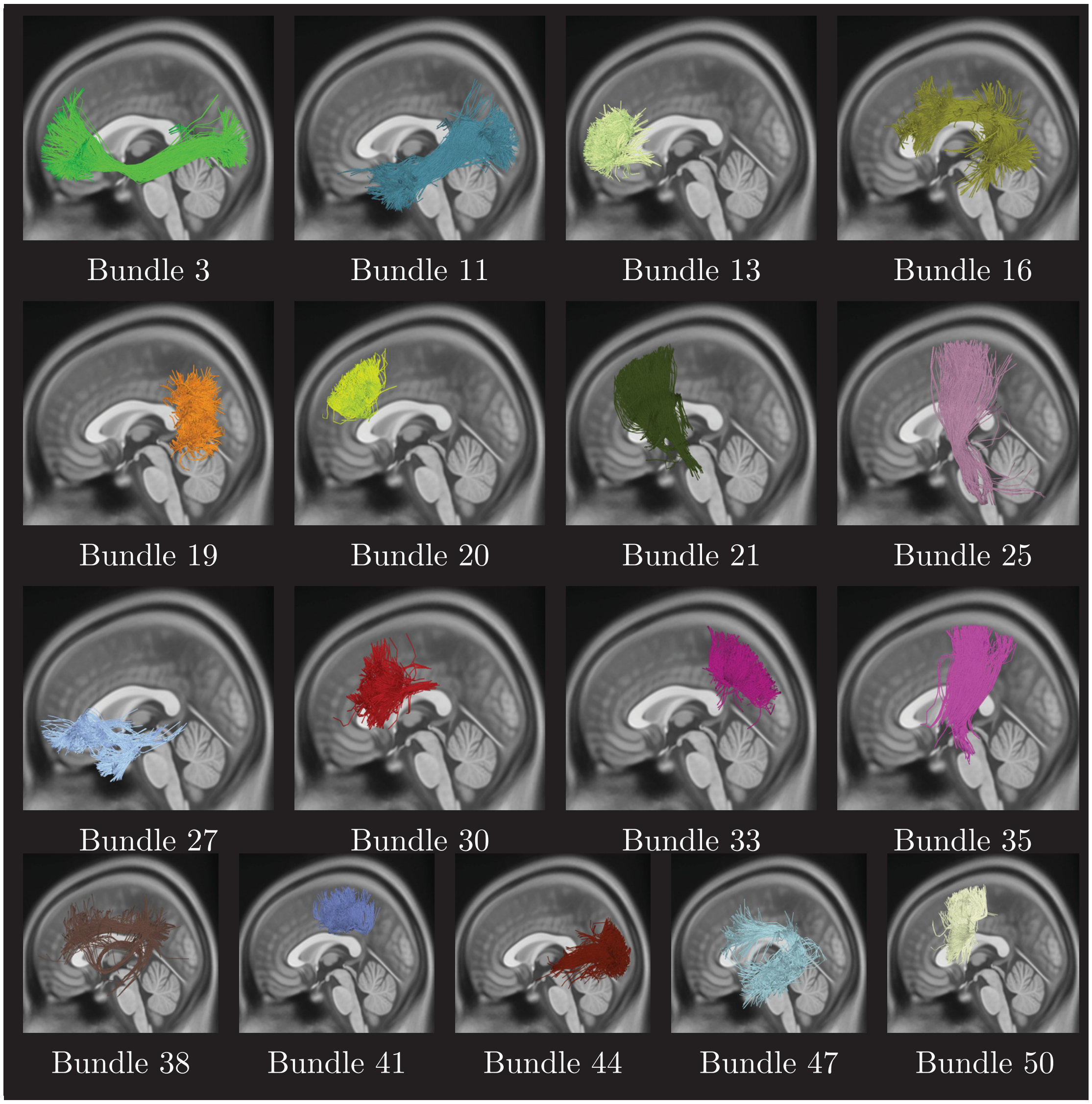


Figure 3.20 Left hemisphere dictionary bundles with respective bundle IDs. (Left sagittal view)

### 3.8.2 Comparison of subject fingerprint across instances and Encoding sparsity

Figure 3.23 compares the fingerprints for two independent instances of a given subject. We observe that the fingerprints are similar across instances for all pooling functions.

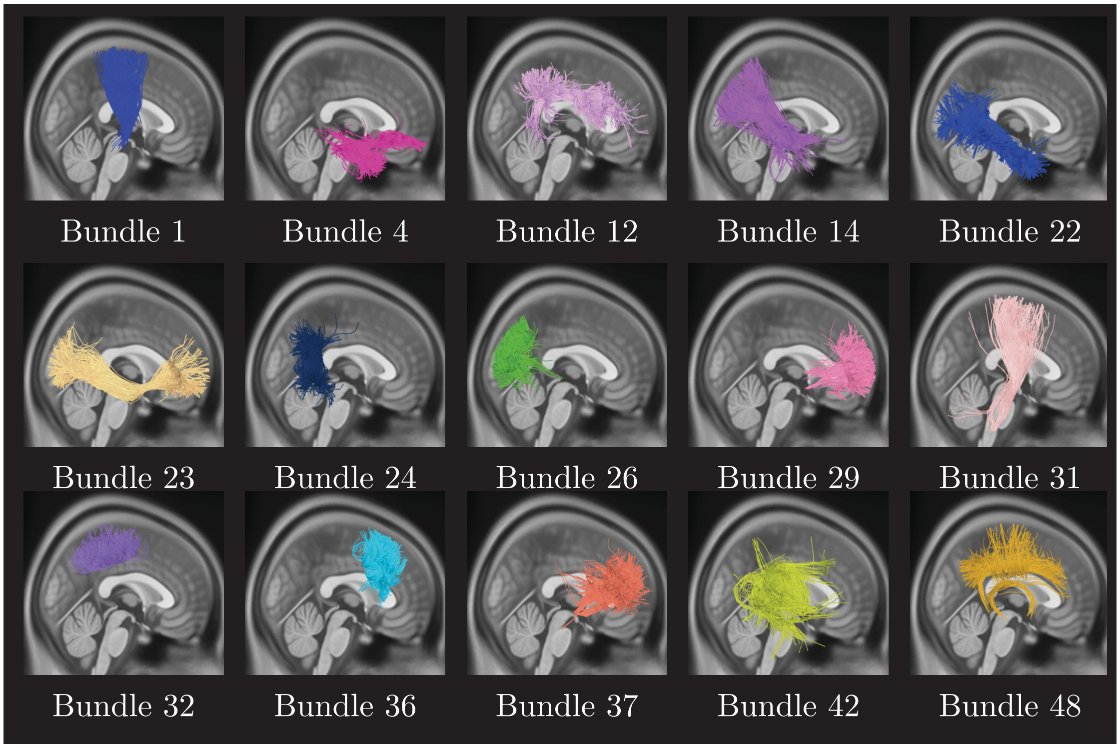


Figure 3.21 Left hemisphere dictionary bundles with respective bundle IDs. (Right sagittal view)

Figure 3.24 (Right) shows box plot for impact of the level of sparsity $S_{\max}$ on subject fingerprint for $m = 100$. We observe trend to be similar to box plots for $m = 50$, thus, justifying our choice of $S_{\max}$ parameter for our experiments.

### 3.8.3 Impact of Handedness on subject identification

Table 3.3 shows mean precision@1 (in $\%$) for a varying number of fiber trajectories using the RMS pooling function and $80$ Left handed and $80$ Right handed subjects, in a nearest neighbor analysis.

### 3.8.4 Impact of age on twin identification

Figure 3.24 (left) measures the impact of age on MZ-twin identification, where twin pairs were divided based on the median age. We observe that age does not impact MZ-twin identification,

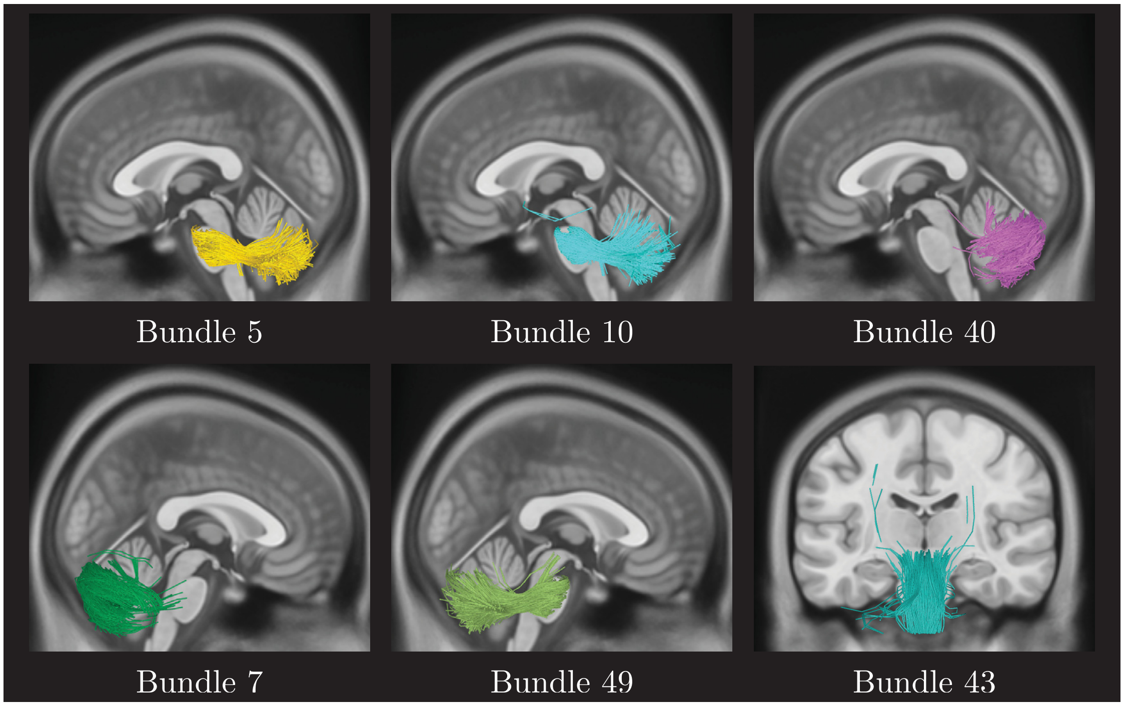


Figure 3.22 Cerebellum dictionary bundles with respective bundle IDs. (Left sagittal view in top row; right sagittal and posterior coronal view in bottom row)

pairs in different age groups having similar mean recall@k values. This is consistent with HCP study design, which aims to minimize the impact of age by selecting subjects in the plateau of white matter development. Figure 3.24 (right) gives the results of a similar analysis of DZ twins. While group mean recall@k plots are significantly higher than random mean recall@k, differences between the mean recall@k values obtained in the two age groups are not statistically significant.

### 3.8.5 Twin fingerprint analysis

Figure 3.25 gives the distributions of Euclidean distances between fingerprint instances of subject pairs corresponding to MZ, DZ and non-twin siblings. We observe that distances between instances of MZ twins (BMZ) are smaller than those between instances of DZ twins (BDZ) and distances between instance of Non-Twin siblings (BNT). This figure also highlights that finger-

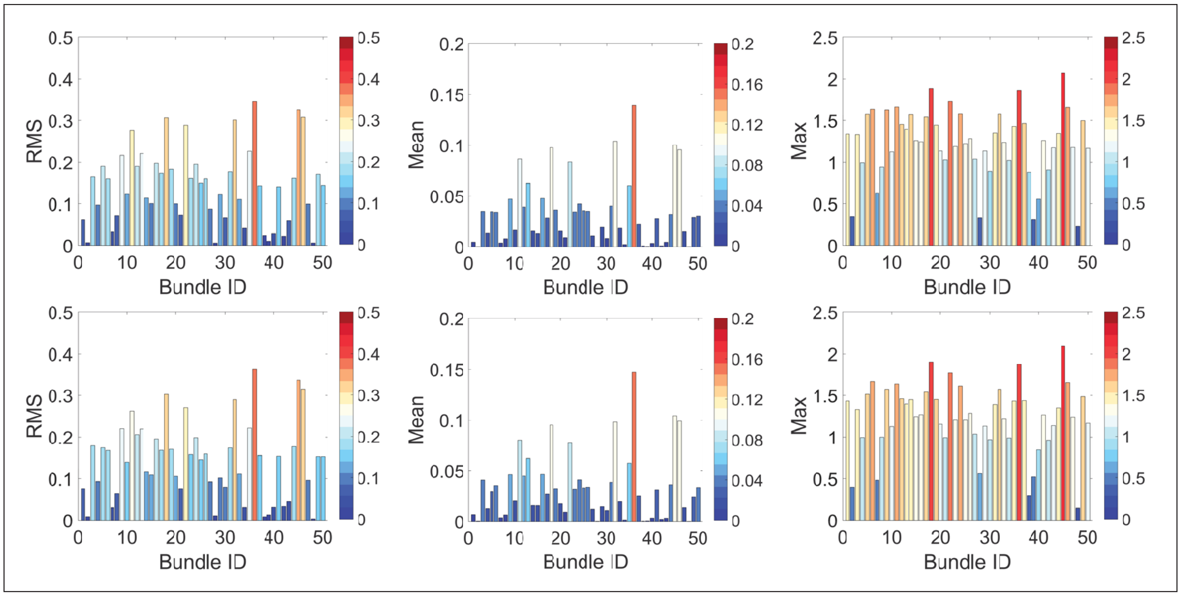


Figure 3.23 Fingerprint comparison across two instances of a subject. Color coded bar representation of subject fingerprint (subject $1$, Instance $1$ and $2$); with Columns representing: RMS, Mean, and Max pooling functions respectively

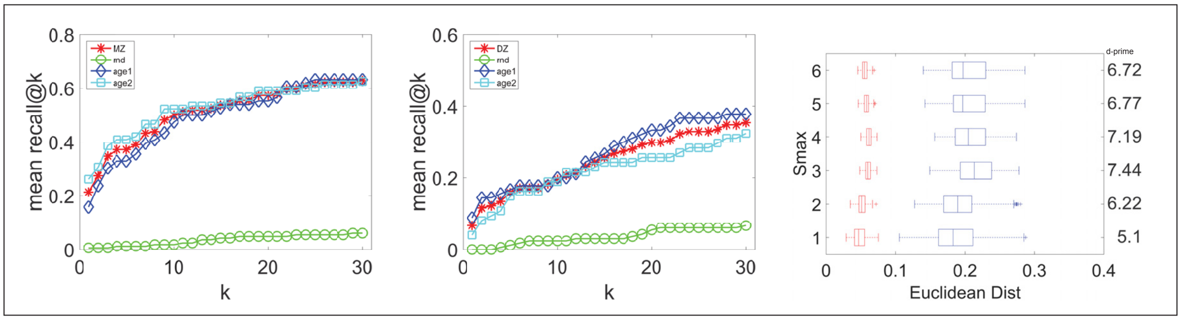


Figure 3.24 Impact of age on MZ/DZ-twin identification; and the level of sparsity $S_{\max}$ on subject fingerprint for $m = 100$. (Left) The mean recall@k for MZ-Twin identification (82 pairs); where, $22 \leq age1 \leq 29$, and $30 \leq age2 \leq 35$, 30 is the median age for MZ-twin pairs. (Middle) The mean recall@k for DZ-Twin identification (82 pairs); where, $22 \leq age1 \leq 29$, and $30 \leq age2 \leq 35$, 29 is the median age for DZ-Twin pairs. (Right) Impact of the level of sparsity $S_{\max}$ on subject fingerprint for $m = 100$. Note: mean recall@k for random lists of nearest-neighbors is identified by rnd.

prints of MZ twins are more similar than those of DZ and non-twin siblings, but same-subject instances are still more similar to one another. Thus, subject signatures preserve individual differences, and show an expected trend for MZ, DZ and non-twins siblings.

Table 3.3 Same-subject instance identification. Mean precision@1 (in %) for a varying number of fiber trajectories using the RMS pooling function and 80 Left handed and 80 Right handed 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 0.6% 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 | 6.0 | 3.4 | 2.3 | 1.8 | 1.9 |
| 500 | 58.3 | 40.3 | 12.3 | 13.8 | 9.3 |
| 1 000 | 94.3 | 85.0 | 33.8 | 30.0 | 23.8 |
| 2 000 | 99.6 | 98.9 | 74.4 | 65.4 | 47.0 |
| **3 000** | **100.0** | **99.9** | 92.0 | 86.9 | 69.0 |
| 4 000 | 100.0 | 100.0 | 96.9 | 93.8 | 78.8 |
| 5 000 | 100.0 | 100.0 | 98.6 | 96.0 | 87.6 |
| 6 000 | 100.0 | 100.0 | 99.4 | 97.9 | 92.4 |
| 8 000 | 100.0 | 100.0 | 99.5 | 99.5 | 96.9 |
| 10 000 | 100.0 | 100.0 | 99.9 | 99.9 | 98.6 |

### 3.8.6 Additional Plots on bundle significance tests

Figures 3.26, 3.27 and 3.28 show the results of the bundle-wise significance analysis, using an unpaired t-test. Plots give the color coded Holm-Bonferroni corrected p-values ($-\log_{10}$ scale), corresponding to each dictionary bundle. These results validate those obtained using the Wilcoxon rank sum tests.

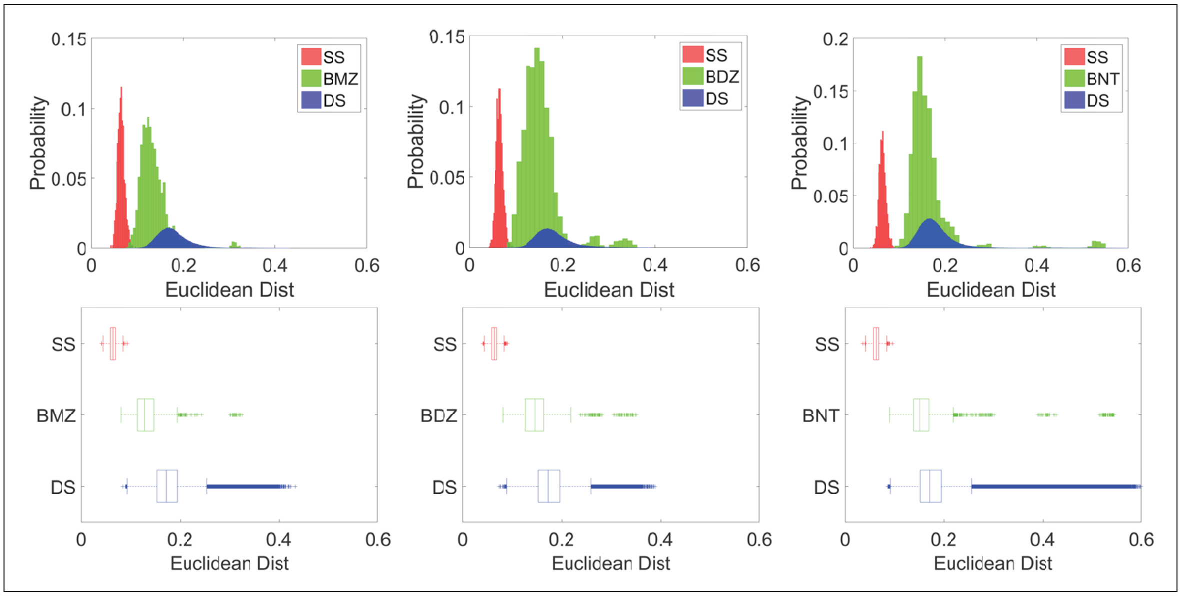


Figure 3.25 Euclidean distance based differences between fingerprints of genetically-related subjects w.r.t. same-subject (SS) and different-subject (DS) instances. Probability normalized histogram and box plots of Euclidean distances for MZ twins (164 subjects), DZ twins (164 subjects), and Non-Twin siblings (215 subjects) using RMS pooling function

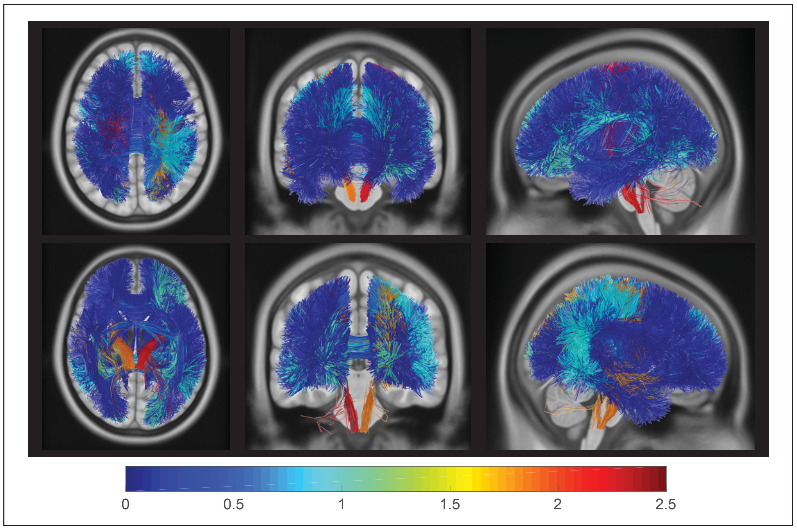


Figure 3.26 **MZ vs NT**. Differences between MZ-twin and Non-Twin siblings. Color coded bundle visualization for Holm-Bonferroni corrected p-values (in -$\log_{10}$ scale) obtained using an unpaired t-test. (superior axial, anterior coronal, and left sagittal views (top row); inferior axial, posterior coronal, and right sagittal views (bottom row);)

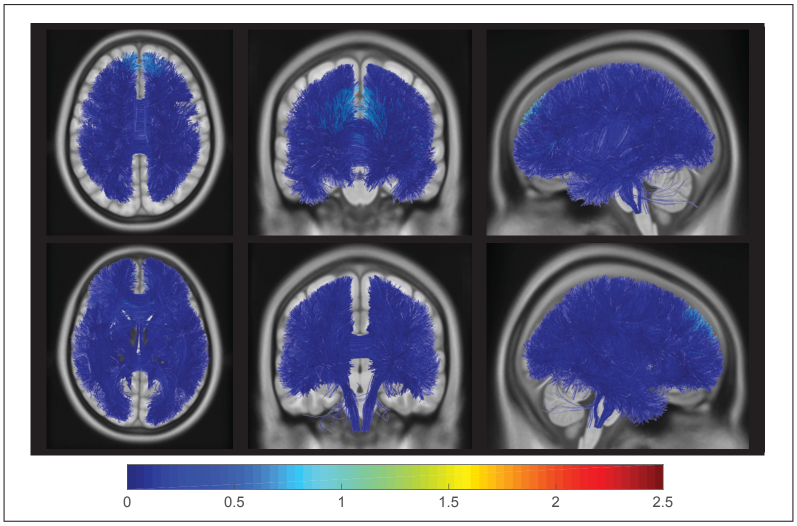


Figure 3.27 **DZ vs NT**. Differences between DZ-twin and Non-Twin siblings. Color coded bundle visualization for Holm-Bonferroni corrected p-values (in -$\log_{10}$ scale) obtained using an unpaired t-test. (superior axial, anterior coronal, and left sagittal views (top row); inferior axial, posterior coronal, and right sagittal views (bottom row);)

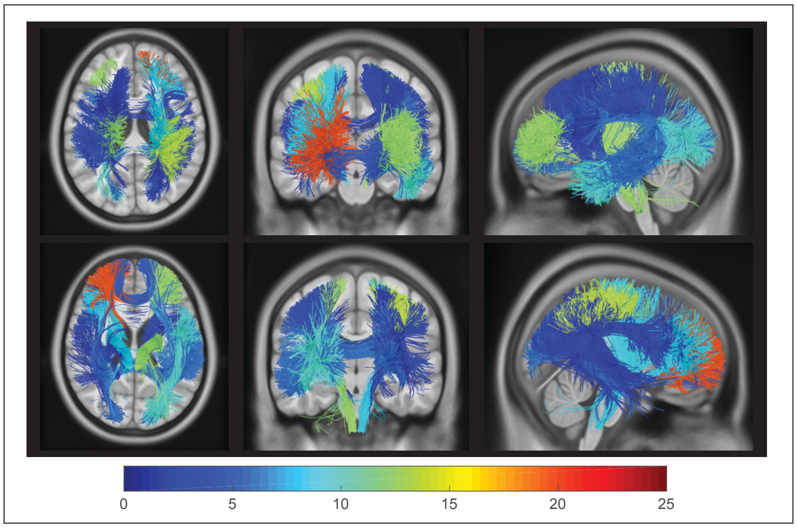


Figure 3.28 **Male vs Female**. Differences related to gender. Color coded bundle visualization for Holm-Bonferroni corrected p-values (in -$\log_{10}$ scale) obtained using an unpaired t-test. (superior axial, anterior coronal, and left sagittal views (top row); inferior axial, posterior coronal, and right sagittal views (bottom row); **Note**: for visualization purposes, fibers in non-significant bundles are not shown.

## 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.

# CHAPTER 4

# MULTI-MODAL BRAIN FINGERPRINTING: A MANIFOLD APPROXIMATION BASED FRAMEWORK

Kuldeep Kumar[a,c,*], Matthew Toews[a], Laurent Chauvin[a],
Olivier Colliot[b,c,d], Christian Desrosiers[a]
[a]LIVIA, École de technologie supérieure, Montreal, Canada
[b]Sorbonne Universités, UPMC, Inserm, CNRS, ICM, Paris, France
[c]Inria Paris, Aramis project-team, Paris, France
[d]AP-HP, Dept of Neurology and Neuroradiology, Hôpital Pitié-Salpêtrière, Paris, France
Email: kkumar@livia.etsmtl.ca, christian.desrosiers@etsmtl.ca

This article was submitted to NeuroImage, Elsevier, on March 18, 2018

## 4.1 Abstract

This work presents an efficient framework, based on manifold approximation, for generating brain fingerprints from multi-modal data. The proposed framework represents images as bags of local features which are used to build a subject proximity graph. Compact fingerprints are obtained by projecting this graph in a low-dimensional manifold using spectral embedding. Experiments using the T1/T2-weighted MRI, diffusion MRI, and resting state fMRI data of $945$ Human Connectome Project subjects demonstrate the benefit of combining multiple modalities, with multi-modal fingerprints more discriminative than those generated from individual modalities. Results also highlight the link between fingerprint similarity and genetic proximity, monozygotic twins having more similar fingerprints than dizygotic or non-twin siblings. This link is also reflected in the differences of feature correspondences between twin/sibling pairs, occurring in major brain structures and across hemispheres. The robustness of the proposed framework to factors like image alignment and scan resolution, as well as the reproducibility

of results on retest scans, suggest the potential of multi-modal brain fingerprinting for characterizing individuals in a large cohort analysis.

## 4.2 Introduction

Despite sharing gross similarities, individual brains show a significant amount of variability (Gordon *et al.*, 2017b) in terms of structure (Mangin *et al.*, 2004), function (Barch *et al.*, 2013; Gordon *et al.*, 2017a; Mueller *et al.*, 2013), and white matter architecture (Bürgel *et al.*, 2006; de Schotten *et al.*, 2011). Recently, various studies have focused on characterizing this variability using brain *fingerprints*, for instance, based on shape (Wachinger *et al.*, 2015a), functional connectivity (Finn *et al.*, 2015; Liu *et al.*, 2018), white matter fiber geometry (Kumar *et al.*, 2017b), or voxel-wise diffusion density (Yeh *et al.*, 2016a). These studies are motivated by the fact that brain characteristics are largely determined by genetic factors that are often unique to individuals (Thompson *et al.*, 2013). Moreover, various neurological disorders like Parkinson (Geevarghese *et al.*, 2015) and autism (Goldman *et al.*, 2013) have been linked to specific brain abnormalities that are difficult to describe at the population level. With the rapid improvements in MRI acquisition hardware and analysis tools, and thanks to large brain-related initiatives like the Human Connectome Project (HCP) (Van Essen *et al.*, 2013) and UK Biobank (Sudlow *et al.*, 2015), researchers are better poised to study individual subjects in addition to groups (Dubois and Adolphs, 2016; Gordon *et al.*, 2017c), thus taking a step towards precision medicine (Hampel *et al.*, 2017) and precision psychiatry (Finn and Constable, 2016).

The importance of brain fingerprinting is evident from the recent surge in studies on this topic. For example, Yeh et al. (Yeh *et al.*, 2016a) built a local connectome fingerprint using dMRI data, and applied this fingerprint to the analysis of genetically-related subjects. Kumar et al. (Kumar *et al.*, 2017b) proposed another dMRI-based fingerprint called *Fiberprint*, which characterizes white matter fiber geometry. Finn et al. (Finn *et al.*, 2015) considered the correlation between time courses of atlas-defined nodes to generate a functional connectivity profile, and used this profile to identify individuals across scan sessions, both for task and rest conditions.

Liu et al. (Liu *et al.*, 2018) use dynamic brain connectivity patterns for identifying individuals and predicting higher cognitive functions. Moreover, Miranda et al. (Miranda-Dominguez *et al.*, 2014) proposed a linear model to describe the activity of brain regions in resting-state fMRI as a weighted sum of its functional neighboring regions. Their functional fingerprint, derived from the model's coefficients, has the ability to predict individuals using a limited number of non-sequential frames.

Various morphometry-based fingerprints have also been proposed for structural MRI modalities like T1- or T2-weighted images. For example, Wachinger et al. (Wachinger *et al.*, 2015a) quantify the shape of cortical and subcortical structures via the spectrum of the Laplace-Beltrami operator. The resulting representation, called *Brainprint*, is used for subject identification and analyzing potential genetic influences on brain morphology. Toews et al. (Toews *et al.*, 2010) represent images as a collection of localized image descriptors, and apply scale-space theory to analyze their distribution at the characteristic scale of underlying anatomical structures. This representation is employed to identify distinctive anatomical patterns of genetically-related individuals or subjects with a known brain disease.

So far, fingerprinting studies in the literature have focused on a single modality. However, each modality captures unique properties of the brain and combining multiple modalities can provide a richer, more discriminative information (Calhoun and Sui, 2016; Groves *et al.*, 2012). Hence, the fusion of multiple modalities has been shown superior than single-modality data to identify diseases like schizophrenia, bipolar disorder, major depressive disorder and obsessive-compulsive disorder (Calhoun and Sui, 2016). Multi-modal neuroimaging biomarkers have also been proposed to predict cognitive deficits in schizophrenia (Sui *et al.*, 2015). Similarly, the combination of multiple MRI modalities has led to the improved segmentation of isointense infant brain images (Zhang *et al.*, 2015). Multi-modal imaging data can also be used to predict the brain-age of subjects and detect cognitive impairments (Liem *et al.*, 2017). Detailed reviews on multi-modal methods and investigations for psychopathology can be found in (Calhoun and Sui, 2016; Liu *et al.*, 2015a,b).

Due to the challenges of combining multiple modalities in a single framework (Calhoun and Sui, 2016; Liu *et al.*, 2015b), defining a multi-modal brain fingerprinting remains to this day an elusive task. Morphometry-based approaches, such as Brainprint (Wachinger *et al.*, 2015a), could potentially be extended to other modalities like dMRI. However, this requires solving non-trivial problems such as the cross-modality alignment of images with different resolutions, the segmentation and correspondence of neuroanatomical structures, etc. Computational efficiency is another important issue when dealing with large-scale, multi-subject and multi-modal datasets like the Human Connectome Project (HCP) (Van Essen *et al.*, 2013) and UK Biobank (Sudlow *et al.*, 2015). In this work, we propose a multi-modal brain fingerprinting that overcomes these challenges using manifold approximation. The key idea is to represent each image as a bag of local features, and derive a subject-level proximity graph using feature correspondences over the entire set of images (Toews *et al.*, 2010). This subject proximity graph provides an approximation of the image appearance subspace (i.e., the manifold), which can be used to obtain a compact fingerprint representation.

Manifold learning has been extensively studied in machine learning (Bengio *et al.*, 2013) with many approaches like Isomap (Tenenbaum *et al.*, 2000), Locally Linear Embedding (LLE) (Roweis and Saul, 2000) and Spectral Embedding (Belkin and Niyogi, 2003)proposed over the years. As detailed in (Aljabar *et al.*, 2012), such techniques have also been used for various problems of medical imaging like registration, segmentation, and classification. For example, in (Gerber *et al.*, 2010), Gerber et al. use manifold learning to perform a population analysis of brain images. Similarly, a deep learning based approach is explored in (Brosch *et al.*, 2013) to learn the manifold of brain MRIs. A key factor in such methods is image representation. For instance, the manifold could be approximated using the Euclidean distance between image pairs, however this would not be robust to translation, rotation or scaling, and would suffer from high computational costs.

Representations based on local features, often referred to as bag of features (BoF), offer an efficient alternative for encoding and matching image structures without the stringent requirement of one-to-one correspondence (Lowe, 2004; Tsai, 2012). In brain imaging, BoFs have been shown to automatically identify known structural differences between healthy controls and Alzheimer's subjects in a fully data-driven fashion (Toews *et al.*, 2010). They have also been used successfully to model the development of infant brains (Toews *et al.*, 2012) and align images of different modalities (Toews and Wells, 2013). Despite their numerous advantages, BoFs have thus far not been explored for brain fingerprinting. This is mainly due to their large and variable size, which makes comparing two fingerprints non-trivial.

The key contributions of this work are as follows:

- **Novel framework:** We propose a data-driven approach based on BoFs and manifold approximation that combines the information from multiple imaging modalities into a common fingerprint. By embedding BoFs in a low-dimensional manifold, the proposed approach circumvents the problem of variable representation size, and provides a compact description of brain structure that enables efficient comparisons across subjects. Furthermore, we show how this manifold-based approach can be used to encode non-structural brain characteristics, for instance, modeling functional connectivity profiles from fMRI. To our knowledge, this is the first work to combine structural, diffusion, and functional modalities in a single fingerprint.
- **Large-scale analysis:** We present a comprehensive analysis of the proposed fingerprint using a large-scale dataset from the Human Connectome Project (HCP), where numerous properties/factors are investigated: fingerprint parameters (e.g., manifold dimensionality and proximity graph connectivity), contribution of individual modalities and/or their combination to the fingerprint's discriminativeness, robustness to image alignment and scan resolution, and reproducibility of results with re-test or corrupted scans. Using genetically verified zygosity labels from the HCP twin dataset, we also analyze the proposed fingerprint's ability to identify genetically-related subjects (i.e., monozygotic twins, dizygotic twins and non-twin siblings) from a large cohort, and show our

multi-modal fingerprint to outperform single-modality approaches or fingerprints based on raw images. In an analysis of local feature correspondences, we identify for individual modalities the neuroanatomical regions having the most significant differences across groups of genetically-related subjects, between males and females, and across brain hemispheres.

This study extends our preliminary work in (Toews and Wells, 2016; Kumar *et al.*, 2017a), where BoF representations were used to identify and compare subjects in a population. Here, we show how these variable-length representations can be converted to fixed-sized fingerprints via manifold embedding, and present an out-of-sample strategy to generate fingerprints for new subjects. While our previous work only considered structural and diffusion MRI data, the current study also investigates the benefit of including fMRI-based information, as well as different combinations of sMRI, dMRI, and fMRI data. Additionally, it offers a much deeper analysis where the impact of multiple factors like the inclusion of skull tissue, image alignment and scan resolution are evaluated. The present study also complements the recent work of Colclough et al. (Colclough *et al.*, 2017), which analyzes the heritability of functional connectivity profiles from multi-modal data (i.e., fMRI and MEG) using the HCP twin dataset. Unlike this recent work, our study analyzes the relationship between genetic proximity and fingerprint similarity based on a rank retrieval analysis, and shows that a higher retrieval accuracy can be obtained when combining structural, diffusion, and functional imaging data.

The rest of this paper is organized as follows. We first present the proposed multi-modal brain fingerprinting framework, detailing the data pre-processing steps, the BoF representation and proximity graph computation, as well as the manifold embedding of this graph. In Section 4.4, we then conduct an extensive experimental validation using the T1-weighted, T2-weighted, diffusion-weighted MRI, and resting state fMRI data of $945$ subjects from the HCP dataset. Finally, we conclude with a summary of our contributions and a discussion of possible extensions.

## 4.3 Materials and methods

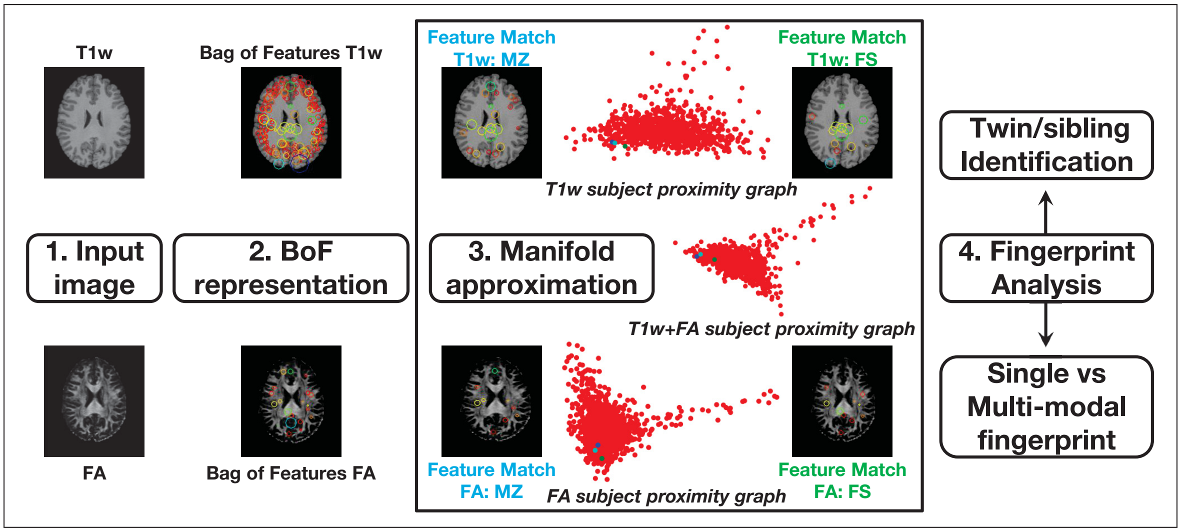


Figure 4.1 Pipeline of the proposed framework and analysis. For a given input image, a BoF representation is first obtained by extracting local features. This representation is then converted to a fingerprint by matching features across the entire set of images, and embedding the resulting proximity graph into the manifold. The manifold approximation block shows the 2D embedding coordinates (i.e., fingerprint) of HCP subjects (red dots) obtained with T1w (top), FA (bottom) and combined T1w+FA (middle) images. The fingerprints of a specific subject (blue dot), his/her monozygotic twin (MZ, cyan dot) and full sibling (FS, green dot) are highlighted in each plot. The pairwise feature matches of these two sibling pairs, for T1w and FA images, are shown in the corner images of the block.

Figure 4.1 summarizes the pipeline of the proposed multi-modal brain fingerprint framework, which is comprised of four steps. In the first step, we start with pre-processed structural MRI (sMRI) and diffusion MRI (dMRI) data of 945 subjects from the Human Connectome Project (Van Essen *et al.*, 2012, 2013). Diffusion Tensor Imaging (DTI) and Generalized Q-Ball Imaging (GQI) based Diffusivity measures are obtained from dMRI scans, including: fractional anisotropy (FA), axial diffusivity (AD), mean diffusivity (MD), radial diffusivity (RD) and generalized fractional anisotropy (GFA). The second step then extracts local features from the images of each subject, and encodes subjects as a bag of features (BoF). In the third step, the multi-modal fingerprints of subjects are computed using manifold approximation. Towards this goal, a subject-level proximity graph is first constructed by matching the features of each modality across images, and identifying pairs of subjects with a high number of correspondences. Fingerprints are then obtained by embedding this graph in a low-dimensional subspace. In the last step, we perform various analyses on the subject fingerprints. The informativeness of individual modalities and their link to genetic proximity are first measured in a twin/sibling identification analysis. This analysis is then extended to multi-modal fingerprints, showing the combined effect and complementarity of multiple modalities. Resting state fMRI network matrices and FreeSurfer derived measures of volume, thickness, and area provided by HCP are also used for fingerprint analysis. Finally, the distribution of feature correspondences between pairs of subjects are used to identify regions showing significant differences across different sibling types. The following subsections describe each of these steps in greater detail.

### 4.3.1 Data and pre-processing

We used the pre-processed structural and diffusion MRI data, and the resting state fMRI network matrices of $945$ subjects from the HCP1200 release of the Human Connectome Project (Van Essen *et al.*, 2013). The retest data of $42$ subjects from the same dataset were also considered in our study to evaluate reproducibility. The HCP1200 release provides genetically-verified labels for twins and siblings, and is a rich resource to analyze the importance of environmental and genetic influences for traits, phenotypes, and disorders (Kochunov *et al.*, 2015; Van Essen *et al.*, 2012). Table 4.1 provides the demographic details of the subjects used in this study.

Data were acquired on a Siemens Skyra $3$T scanner (Sotiropoulos *et al.*, 2013) and pre-processed as described in (Glasser *et al.*, 2013). The structural acquisitions include high resolution T1-weighted (T1w) and T2-weighted (T2w) images ($0.7\,$mm isotropic, FOV = $224\,$mm, matrix = $320$, $256$ sagittal slices in a single slab), the diffusion acquisition used following parameters: sequence = Spin-echo EPI; repetition time (TR) = $5520\,$ms; echo time (TE) = $89.5\,$ms; resolution = $1.25 \times 1.25 \times 1.25\,\text{mm}^3$ voxels, and the resting-state fMRI acquisition involved four $15\,$min runs at $2\,$mm isotropic resolution and a repetition time of $0.72\,$s ($4800\,$vol per subject). Further details can be obtained from the HCP1200 data release manual[1]. We used the `hcp2blocks.m` script (described in the HCP1200 release) to generate a FamilyID based matrix, only considering subjects having dMRI, sMRI, and rfMRI netmats data, and for which the `HasGT` field is true. Using this selection criterion, we obtained a total of $238$ monozygotic (MZ) subjects, $126$ dizygotic (DZ) subjects, and $581$ non-twin subjects. The sibship size ranged between $1$ and $6$. In a next step, using the mother ID, father ID, family ID and family type information from the output of `hcp2blocks.m` script, we obtained $119$ monozygotic twin pairs, $63$ dizygotic twin pairs, $546$ full-sibling (FS) pairs, $39$ maternal half sibling (MHS) pairs, and $5$ paternal half sibling (PHS) pairs. These pairs are used for twin/sibling identification task in the following sections.

Table 4.1 Demographics. We considered the HCP1200 release subjects with structural MRI, diffusion MRI, and resting state fMRI netmats data, and for which the HasGT field is true (genetically verified data). The family structure and links are obtained from the output of hcp2blocks.m script listed in data release manual. The sibship sizes are between 1 and 6.

| **Type** | **Total** | **Gender** | | **Age** | **Handedness** | |
|---|---|---|---|---|---|---|
| | | F | M | Range (median) | L | R |
| All | 945 | 501 | 444 | 22-36 (29) | 84 | 861 |
| MZ | 238 | 138 | 100 | 22-36 (30) | 19 | 219 |
| DZ | 126 | 70 | 56 | 22-35 (29) | 13 | 113 |
| NotTwin | 581 | 293 | 288 | 22-36 (28) | 52 | 529 |

For structural MRI, we considered T1-weighted ($0.7\,$mm) and T2-weighted ($0.7\,$mm), with and without skull. The images are in native space and skull stripped, unless explicitly specified. In the case of dMRI data, signal reconstruction was performed with the freely available DSI Studio toolbox (Yeh *et al.*, 2010) using the Diffusion Tensor Imaging (DTI) and Generalized Q-Ball Imaging (GQI) reconstruction options. Four widely used DTI-based measures were extracted to characterize white matter micro-structure: fractional anisotropy (FA), axial dif-

---

[1] https://www.humanconnectome.org/documentation/S1200/

fusivity (AD), mean diffusivity (MD) and radial diffusivity (RD). The interpretation of these measures are discussed in (Alexander *et al.*, 2007). In addition, to utilize the multi-shell information and high angular resolution of the HCP data, Generalized Q-Ball Imaging (GQI) (Yeh *et al.*, 2010) based measures including generalized fractional anisotropy (GFA) and quantitative anisotropy (QA) were also obtained. For resting state fMRI, we used the connectivity matrices (netmats), provided by the HCP 1200 release, derived using the FSLNets toolbox, either via full correlation or the partial correlation (Smith *et al.*, 2015), the latter being calculated by inverting the covariance matrix. For analyzing the impact of alignment, we also used the MNI space aligned data for T1-weighted ($0.7$ mm) and T2-weighted ($0.7$ mm) provided by the HCP 1200 release. In addition, to combine structural modalities with dMRI, and to analyze impact of scan resolution, we re-sampled T1- and T2-weighted images to a $1.25$ mm resolution using the linear registration (FLIRT) tool of FSL (Jenkinson *et al.*, 2012). Finally, our analysis also considered FreeSurfer derived measures of sub-cortical volumes, cortical thickness and area, as well as T1w/T2w MRI ratio images ($0.7$ mm, myelin content information).

### 4.3.2 Multi-modal brain fingerprint

Generating brain fingerprints of subjects based on their multi-modal data involves multiple steps: extracting local descriptors in images to build a bag of features (BoF) representation of subjects, building a subject proximity graph by comparing their BoF representations, and embedding this graph in a low-dimensional manifold. Additionally, once the manifold has been constructed, an out-of-sample extension strategy is required to compute the fingerprint of new subjects.

#### 4.3.2.1 Bag of feature (BoF) representation of subjects

In the first step, a set of local descriptors (Lowe, 2004) is obtained from each available image (3D scan). Various local feature extraction and representation approaches (Tuytelaars *et al.*, 2008) can be used, for example, Scale Invariant Feature Transfrom (SIFT) (Lowe, 1999) or

Speeded UP Robust Features (SURF) (Bay *et al.*, 2006). In this work, we use 3D SIFT descriptors as they have been well studied for neuro-image analysis (Toews *et al.*, 2010; Toews and Wells, 2013; Toews *et al.*, 2015) and can be computed efficiently.

3D keypoints are located in the scans of each subject by finding the local extrema (i.e., maxima or minima) of the difference of Gaussians (DoG) occurring at multiple scales. Keypoints with a low contrast or corresponding to edge response are discarded, and remaining ones are encoded into a feature vector (i.e, the descriptor) using the histogram of oriented gradients (HOG) within a small neighborhood. Note that these descriptors are robust to changes in illumination, scale and rotation, and are thus efficient for comparing images acquired using different scanners or imaging parameters. Each subject is then represented as an orderless bag of features (BoF), containing all the descriptors found in this subject's scans. This representation provides a simple, robust and extensible way of incorporating data from multiple modalities.

#### 4.3.2.2 Subject proximity graph

Because the BoFs of two subjects may contain different numbers of descriptors, they are difficult compare directly. To circumvent this problem, we construct an intrinsic manifold of subject appearance using a nearest-neighbor (NN) graph in feature space. In this graph, each descriptor is represented by a node and is connected to its $K$ most similar appearance descriptors based on Euclidean distance. The $K$-nearest neighbors of each descriptor can be computed in sublinear time, for example, using randomized KD-search trees (Muja and Lowe, 2009). This feature graph is then used to induce a subject proximity graph by considering, for each pair of subjects, the number descriptors in their BoF that are linked in the feature graph.

Let $\boldsymbol{B}_i^m$ and $\boldsymbol{B}_j^m$ be the BoFs (i.e., set of descriptors) of subjects $i$ and $j$ for modality $m \in \boldsymbol{M}$, where $\boldsymbol{M}$ is the set of available modalities. The similarity between these subjects is evaluated as

$$s_{ij} \;=\; \frac{\sum_{m\in\boldsymbol{M}} |\boldsymbol{B}_i^m \cap \boldsymbol{B}_j^m|}{\sum_{m\in\boldsymbol{M}} |\boldsymbol{B}_i^m \cup \boldsymbol{B}_j^m|} \;=\; \frac{\sum_{m\in\boldsymbol{M}} |\boldsymbol{B}_i^m \cap \boldsymbol{B}_j^m|}{\sum_{m\in\boldsymbol{M}} \big(|\boldsymbol{B}_i^m| + |\boldsymbol{B}_j^m| - |\boldsymbol{B}_i^m \cap \boldsymbol{B}_j^m|\big)}, \tag{4.1}$$

where $|\boldsymbol{B}_i^m \cap \boldsymbol{B}_j^m|$ is the number of edges in the feature graph between nodes in $\boldsymbol{B}_i^m$ and those in $\boldsymbol{B}_j^m$. When using a single modality, this measure corresponds to the well-known Jaccard similarity. Here, we extend it to a multi-modal setting by comparing the descriptors of each modality separately. We note that the Jaccard distance, defined as one minus the Jaccard similarity, is a *metric* and thus well-suited for constructing the manifold.

When defining the feature graph, $K$ determines the number of nearest-neighbor connections for each descriptor. In manifold learning approaches, this parameters controls the locality of the manifold approximation at each point (Bengio *et al.*, 2013). Its value should be large enough to capture the manifold's local structure, but also restricted so that distances to nearest-neighbors are close to the geodesic. In our experiments, we tested $K \in \{10, 20, 30, 40, 50\}$ and found similar results for these values. In what follows, we report results obtained with $K = 20$.

#### 4.3.2.3 Manifold embedding

A manifold embedding technique is used to obtain compact brain fingerprints from the subject proximity graph. While various approaches could be employed for this task, for instance Isomap (Tenenbaum *et al.*, 2000) or locally linear embedding (LLE) (Roweis and Saul, 2000), we performed the embedding using Laplacian eigenmaps (Belkin and Niyogi, 2003). This technique, which is connected to the well-known Laplace-Beltrami operator, has the advantage of being efficient and allowing out-of-sample extensions.

In Laplacian eigenmaps, each subject $i$ is mapped to a coordinate vector $\boldsymbol{x}_i \in \boldsymbol{R}^k$ of the manifold, whose dimension $k$ is a user parameter. The embedding of subjects in the manifold is made such that two subjects $i$ and $j$ with a high similarity $s_{ij}$ will be close to one another. Let $\boldsymbol{S} \in \boldsymbol{R}^{n \times n}$ be the adjacency matrix of the subject proximity graph, as defined in Eq (4.1), and denote as $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{S}$ the Laplacian of $\boldsymbol{S}$, where $\boldsymbol{D}$ is a diagonal matrix containing the row sums of $\boldsymbol{S}$. The embedding is accomplished by solving the following problem:

$$\arg\min_{\boldsymbol{X}} \sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 = \mathrm{tr}(\boldsymbol{X}^\intercal \boldsymbol{L}\boldsymbol{X}), \quad \text{s.t. } \boldsymbol{X}^\intercal \boldsymbol{D}\boldsymbol{X} = \boldsymbol{I}. \tag{4.2}$$

The constraint on $\boldsymbol{X}$ removes the arbitrary scaling factor in the embedding. As described in (Belkin and Niyogi, 2003), the solution to this problem is given by the leading $k$ eigenvectors of the normalized adjacency matrix $\overline{\boldsymbol{S}} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{S}\boldsymbol{D}^{-\frac{1}{2}}$, starting from the second one[2]. Once computed, the rows of matrix $\boldsymbol{X}$ correspond to the $n$ subject fingerprints of size $k$.

#### 4.3.2.4 Out-of-sample extension

The manifold embedding technique described above computes the fingerprint of all subjects at once. If new subjects are added, this process must be repeated over again, which is inefficient and changes the fingerprint of previous subjects. To alleviate these problems, we use an out-of-sample extension of Laplacian eigenmaps, based on the Nystrom method (Bengio *et al.*, 2004; Fowlkes *et al.*, 2004).

Suppose we want to compute the manifold embedding of $m$ new subjects. The first step is to update the nearest-neighbor feature graph with the local descriptors of these new subjects, leaving unchanged the nearest-neighbors of the $n$ base subjects. We then evaluate the pairwise similarities between new subjects and the base ones. Let $\boldsymbol{P} \in \boldsymbol{R}^{n\times m}$ be the matrix containing these similarities, the adjacency matrix of the extended subject proximity graph $\boldsymbol{S}_{\text{ext}} \in \boldsymbol{R}^{(n+m)\times(n+m)}$ is given by

$$\boldsymbol{S}_{\text{ext}} = \begin{bmatrix} \boldsymbol{S} & \boldsymbol{P} \\ \boldsymbol{P}^\intercal & \boldsymbol{Q} \end{bmatrix}. \tag{4.3}$$

Using the formula in (Belkin and Niyogi, 2003), the matrix $\boldsymbol{Q}$ of similarities between new subjects can be approximated as $\boldsymbol{P}^\intercal \boldsymbol{S}^{-1}\boldsymbol{P}$.

---

[2] The first eigenvector contains constant values.

To normalize $\boldsymbol{S}_{\text{ext}}$, we compute the vector of row sums

$$\boldsymbol{d}_{\text{ext}} = \begin{bmatrix} \boldsymbol{s}_r + \boldsymbol{p}_r \\ \boldsymbol{p}_c + \boldsymbol{P}^{\mathsf{T}} \boldsymbol{S}^{-1} \boldsymbol{p}_r \end{bmatrix}, \tag{4.4}$$

where $\boldsymbol{s}_r, \boldsymbol{p}_r \in \boldsymbol{R}^n$ contain the row sums of $\boldsymbol{S}$ and $\boldsymbol{P}$, respectively, and $\boldsymbol{p}_c \in \boldsymbol{R}^m$ contains the column sum of $\boldsymbol{P}$. In the case where $m$ is small compared to $n$, we have that $\boldsymbol{s}_r \approx \boldsymbol{s}_r + \boldsymbol{p}_r$, and thus $\boldsymbol{d}_{\text{ext}}$ can be approximated as

$$\boldsymbol{d}_{\text{ext}} \approx \begin{bmatrix} \boldsymbol{s}_r \\ \boldsymbol{p}_c + \boldsymbol{P}^{\mathsf{T}} \boldsymbol{S}^{-1} \boldsymbol{p}_r \end{bmatrix}. \tag{4.5}$$

This strategy, used in (O'Donnell and Westin, 2007a) for white matter fiber segmentation, allows preserving the previous embedding of base subjects. Let $\boldsymbol{D}_{\text{ext}}$ be the diagonal matrix with entries corresponding to $\boldsymbol{d}_{\text{ext}}$, the normalized adjacency matrix of the extended subject graph is calculated as $\overline{\boldsymbol{S}}_{\text{ext}} = \boldsymbol{D}_{\text{ext}}^{-\frac{1}{2}} \boldsymbol{S}_{\text{ext}} \boldsymbol{D}_{\text{ext}}^{-\frac{1}{2}}$. The extended embedding is then obtained following Nystrom's method as

$$\boldsymbol{X}_{\text{ext}} = \begin{bmatrix} \boldsymbol{U} \\ \overline{\boldsymbol{P}}^{\mathsf{T}} \boldsymbol{U} \Lambda^{-1} \end{bmatrix}, \tag{4.6}$$

where $\boldsymbol{U} \Lambda \boldsymbol{U}^{\mathsf{T}}$ is the eigendecomposition of $\overline{\boldsymbol{S}}$, and $\overline{\boldsymbol{P}}$ is the normalized submatrix in $\overline{\boldsymbol{S}}_{\text{ext}}$ corresponding to $\boldsymbol{P}$. Hence, the embedding of base subjects is the same as in Section 4.3.2.3, and new subjects are embedded as $\overline{\boldsymbol{P}}^{\mathsf{T}} \boldsymbol{U} \Lambda^{-1}$. Once more, a fingerprint of size $k$ is obtained by considering only the $k$ leading eigenvectors in matrix $\boldsymbol{U}$, ignoring the constant eigenvector.

### 4.3.3 Computational efficiency

Computational and memory requirements are key factors when performing large scale analyses. In this section, we evaluate these requirements for the main steps of the proposed framework. To highlight the efficiency of encoding images with local descriptors, we also compare our framework to a simple fingerprint using full images as features. Other aspects like scan resolution and image alignment requirements are discussed in Section 4.4. All experiments were performed on a 3.6 GHz processor with 32 GB RAM.

For the BoF representation of images, we extracted 3D SIFT features using a publicly available tool[3]. Computing these features took about $3$ seconds per image, and approximately $60$ minutes for all $945$ images, when processed sequentially. This runtime could however be reduced significantly by processing images in parallel. The feature matching routine (Muja and Lowe, 2009), for generating the subject proximity graph from the BoFs of all images, required around $5$ minutes to complete. In comparison, calculating the sum of squared distances (SSD) between full images took 1.7 seconds on average for a single pair, and 760,000 seconds for all (945 $\times$ 944)/2 = 446,040 pairs (with parallel computations). In terms of memory, each BoF file is approximately 400 KB in size, compared to 84 MB on average for a NIfTI volume file. In summary, the proposed framework is highly efficient in terms of computational and memory requirements compared to a baseline fingerprint using full images. Moreover, since computing the subject proximity graph has a complexity in $O(N \log N)$ where $N$ is the number of images, and because extending the manifold embedding can be done efficiently using the Nystrom method, our framework is scalable to large datasets.

### 4.3.4 Evaluation measures

To measure the link between fingerprint similarity and genetic proximity, we performed a rank retrieval analysis using the sibling information provided in the HCP dataset. In this analysis, we try to identify the twins/siblings of a given subject by comparing its fingerprint with that of all other subjects in the group. Another goal of this analysis is to provide a common platform for the quantitative comparison of individual modalities and their combination. Two standard performance metrics for rank retrieval are used to evaluate the fingerprints: mean recall@k and mean average precision (MAP) (Turpin and Scholer, 2006).

---

[3] http://www.matthewtoews.com/

Given a subject $i$, we rank all other subjects by the similarity (i.e., inverse of Euclidean distance) of their fingerprint to that of subject $i$. Denote as $\mathcal{T}_i$ the set of target siblings/twins of subject $i$. For instance, if the target group is non-twin siblings (NT), then $\mathcal{T}_i$ contains the siblings of subject $i$ that are not his/her twin. Moreover, let $\mathcal{S}_i^k$ be the set containing the $k$ subjects with fingerprints most similar to that of $i$ (i.e., the $k$ nearest neighbors of $i$). For a given value of $k$, we evaluate the retrieval performance using the measures of recall@k and precision@k:

$$(\text{recall@}k)_i \;=\; \frac{|\mathcal{T}_i \cap \mathcal{S}_i^k|}{|\mathcal{T}_i|}, \quad (\text{precision@}k)_i \;=\; \frac{|\mathcal{T}_i \cap \mathcal{S}_i^k|}{k}. \tag{4.7}$$

Mean recall@$k$, also known as sensitivity, evaluates the proportion of individuals that are genetically related to a given subject, which are within the $k$ individuals most similar to that subject (in terms of fingerprint distance). When analyzing the rank performance for a particular sibling type (i.e., monozygotic twin, dizygotic twin or non-twin sibling), we average values over the set of subjects which have at least one sibling of this type, i.e. the set $\{i, \text{ s.t. } |\mathcal{T}_i| > 0\}$.

We also evaluate performance with the average precision, which extends the above metrics by considering the rank of nearest neighbors:

$$\text{AveP}_i \;=\; \frac{1}{|\mathcal{T}_i|}\sum_{k=1}^{n}(\text{precision@}k)_i \times \text{rel}_i(k), \tag{4.8}$$

where $\text{rel}_i(k)$ is an indicator function with value equal to $1$ if the $k$-th nearest neighbor of $i$ is relevant (i.e., is in $\mathcal{T}_i$), and zero otherwise. The MAP is obtained by averaging AveP values over all subjects having at least one sibling of the target type.

Finally, we use the d-prime sensitivity index (Gale and Perkel, 2010) to obtain a quantitative measure of separability between the distribution of fingerprint distances corresponding to different sibling types. Let $\mu_1, \mu_2$ and $\sigma_1, \sigma_2$ be the means and standard deviations of compared distance distributions (e.g., distance between monozygotic twins versus between dizygotic twins). The d-prime index is computed as

$$\text{d-prime} \;=\; \frac{\mu_1 - \mu_2}{\sqrt{\frac{1}{2}\big(\sigma_1^2 \;+\; \sigma_2^2\big)}}. \tag{4.9}$$

In our experiments, we report absolute values of d-prime, higher values indicating better separability.

## 4.4 Experiments and results

A comprehensive set of experiments was conducted to analyze the proposed fingerprint and evaluate its usefulness in various applications. In the first experiment, we analyze the manifold embedding of subjects and measure the impact of manifold dimensionality on the fingerprint's ability to capture genetic proximity. We then perform a detailed rank retrieval analysis, in which fingerprints obtained from a single modality or combinations of multiple modalities are used to identify three types of genetically-related subject: monozygotic twins (MZ), dizygotic twins (DZ) and full siblings (FS). The driving hypothesis of this experiment is that individual modalities capture distinct properties of brain tissues, which can be effectively encoded in the fingerprint, and that combining different modalities can help describe the uniqueness of individual brains. Another goal of this experiment is to measure the relationship between the similarity of fingerprints, for different modality combinations, and genetic proximity.

In another experiment, we assess the impact of following factors on the proposed fingerprint: image alignment, scan resolution, inclusion of skull, and subject age. This is followed by a reproducibility analysis, performed with the restest scans of $42$ subjects, and a comparison with a baseline fingerprint using full images as features. The objective of these experiments is to demonstrate the robustness and performance of the proposed fingerprint, compared to a full image scan-based fingerprint.

We also present applications of the proposed framework for identifying retest scans, duplicate corrupt scans, and incorrectly-reported zygosity labels. In addition, we use the segmentation masks provided with the HCP data to identify cortical and subcortical brain regions where the distribution of feature correspondences between monozygotic twins is significantly different from dizygotic twins. In this analysis, we want to find brain regions which are more influenced by genetic promixity. Finally, we conduct a hemisphere asymmetry analysis using the feature correspondences for different types of siblings.

### 4.4.1 Manifold approximation analysis

To analyze the manifold approximation, we generated fingerprints by projecting the subject proximity graph onto a varying number of spectral components (i.e., leading eigenvectors of the normalized adjacency or Laplacian matrix). Fingerprints were normalized by converting each fingerprint to z-scores (centered to have mean 0 and scaled to have standard deviation 1). Figure 4.2 (top row) shows a representative 2D spectral embedding of subject proximity graphs obtained using T1w, FA, or both modalities (T1w+FA). As described in Section 4.3.2.2, modalities are combined by aggregating the feature correspondences in each modality when computing the pairwise subject similarities. In these plots, the position of each red dot corresponds to the 2D fingerprint of a subject. Additionally, in each plot, a single pair of MZ twins is highlighted using blue and cyan dots and their NT sibling highlighted using a green dot.

It can be seen that the distribution of subject embeddings in the manifold varies from T1w to FA, showing that these modalities encode different properties in the fingerprint. Differences between the distributions of FA and T1w+FA fingerprints are in part explained by the fact that spectral embeddings are defined up to a rotation or axis flipping. Moreover, we observe for all three modality combinations that genetically-related subjects are near to each other in the manifold, and that MZ twins are closer than their non-twin (full) sibling.

In Figure 4.2 (bottom row), we measure the impact of manifold dimensionality on the fingerprint obtained with T1w, FA or T1w+FA modalities. The left plot shows the eigenvalues (sorted by decreasing magnitude) of the subject proximity graph's normalized adjacency matrix, which reflect the amount of connectivity information captured by their corresponding eigenvector.

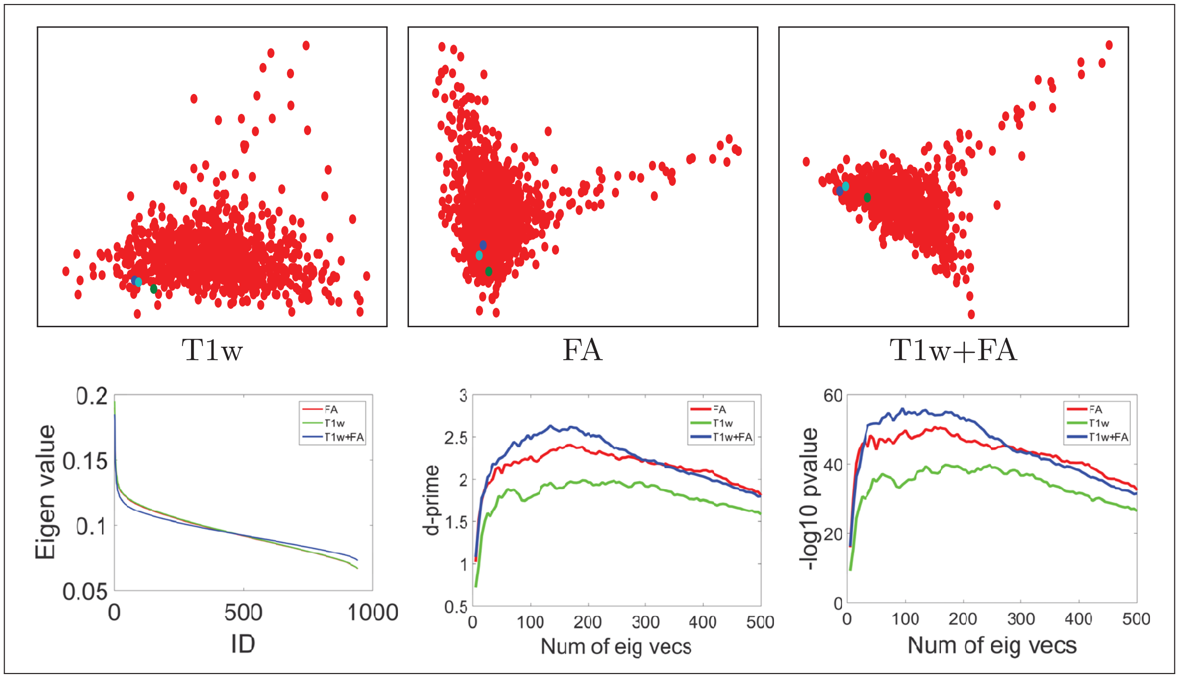


Figure 4.2  Compact fingerprint analysis. Top row: representative 2D spectral embedding visualization, blue and cyan dots show one pair of MZ twins and green dot shows their not twin (full) sibling; Bottom row: plots of eigenvalues (excluding the first), absolute d-prime, and -$\log_{10}$ (p-value) (unpaired t-test) for Euclidean distances between MZ pair vs DZ pair fingerprints with increasing number of eigenvectors.

This plot indicates that most information is encoded in the first leading eigenvectors and, thus, that a compact fingerprint is possible.

This hypothesis is further confirmed in the middle and right plots of the same row, which evaluate for an increasing number of spectral components (i.e., fingerprint size) how the distributions of distances between MZ fingerprints and between DZ fingerprints differ. The separability between these two distributions of fingerprint distances is measured in terms of d-prime (middle plot) and unpaired t-test p-values (in -$\log_{10}$ scale). In both measures, higher values correspond to a greater separability. For all three modality combinations, a peak separability is observed around $150$ eigenvectors, suggesting this value to be suitable for the fingerprint size. The decrease in separability for larger manifold dimensions is due to the fact that the added eigenvectors encode small variations of brain geometry which are not related to genetic proximity. Nevertheless, the difference between fingerprint distances in MZ pairs and in DZ pairs is significant with p-value $< 0.01$, for all tested manifold sizes and modality combinations.

Comparing the three modality combinations, the diffusion-based fingerprint using FA images provides a higher separability than the fingerprint generated from T1w, for all manifold sizes. However, the separability is increased further when combining both modalities in the fingerprint, in line with our hypothesis that multi-modal fingerprints are more discriminative than those based on a single modality.

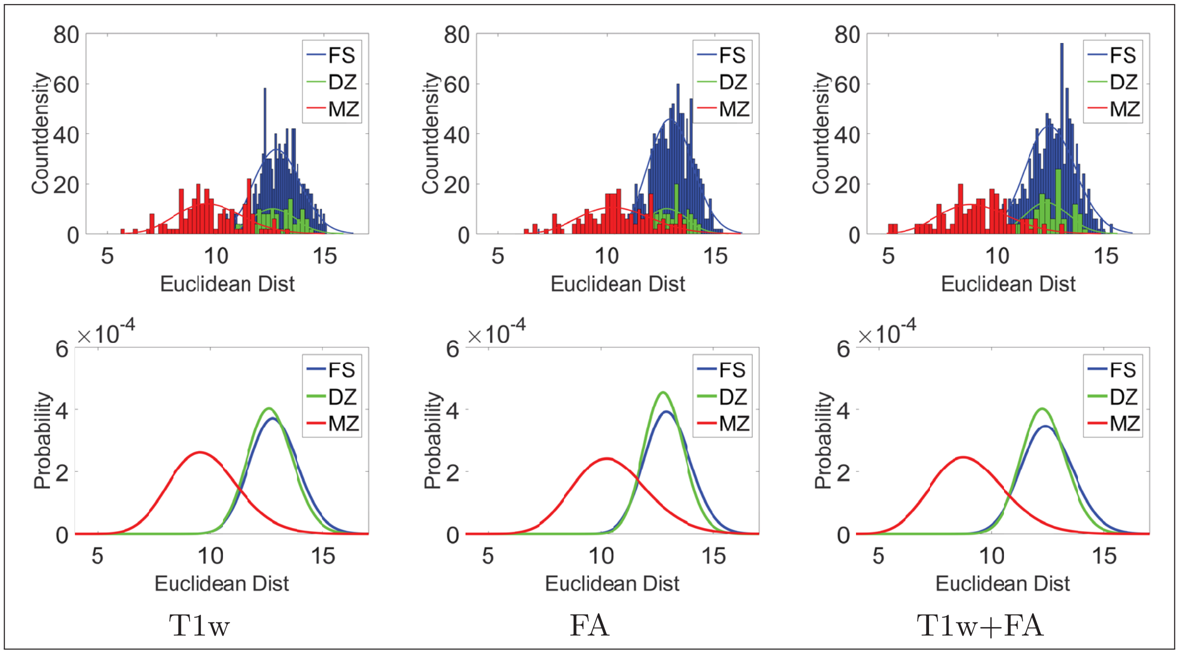


Figure 4.3 Compact fingerprint comparison for genetically-related subjects. Count-density histograms (top row) and probability-normalized curves (bottom row; gamma histogram fitting) of Euclidean distances between twin/sibling pair fingerprints using $150$ eigenvectors.

Finally, Figure 4.3 gives the count histograms and probability density curves (fitted) of Euclidean distances between fingerprints of different sibling types. To generate these results, and in all following experiments, we used a fingerprint of $150$ features (i.e., leading eigenvectors of the normalized adjacency matrix). It can be seen that the fingerprints of MZ twins, which share the most genetic material, are significantly more similar than those of DZ twins or full siblings (FS). This follows the results of various twin studies (Peper *et al.*, 2007; Thompson *et al.*, 2013), highlighting the relationship between genetic proximity and anatomical similarity.

### 4.4.2 Identification of genetically-related subjects

In this section, we use genetically verified labels of the HCP dataset to determine whether fingerprints generated using different modality combinations can identify genetically-related individuals within a group of subjects. For combining structural and diffusion modalities, we considered data at 1.25 mm resolution. For resting state fMRI, we utilize the connectivity matrices (netmats) as functional connectivity fingerprints, and obtain the subject proximity graph (manifold approximation) by computing pairwise Pearson correlation. The idea is to closely follow the functional connectivity fingerprint and similarity computation described in Finn et al. (Finn *et al.*, 2015) (the parcellation and dataset sizes are different). The multi-modal combinations involving rfMRI are obtained by a linear combination of the rfMRI subject proximity graph with the graph derived from BoFs. Combination weights were determined by grid search, and optimal values of evaluation measures are reported. For measures based on FreeSurfer, we used the unrestricted *csv* file, considering volume of sub-cortical structures, thickness and area measures for cortical regions. Each of the measures were converted to z-score across subjects, and then used as a fingerprint (volume measures are first divided by FS-IntraCranial-Vol). Subject proximity graph is approximated by computing the pairwise Pearson correlation. We refer the reader to Section 4.3.4 for details on the evaluation protocol and measures.

Table 4.2 reports the mean average precision (MAP) values obtained in a rank retrieval of three different siblings types (MZ, DZ and FS), using fingerprints generated from various combinations of the following modalities[4]: FA, MD, GFA, rfMRI netmat (partial correlation, ICA 100), and FreeSurfer volume, thickness and area measures (Vol+Thck+Area). To lighten the presen-

---

[4] DTI = FA+MD+RD+AD; rfMRI netmat = partial correlation and ICA-100

Table 4.2 Mean average precision (MAP) obtained with different modality combinations for the identification of genetically-related subjects: monozygotic twins (MZ), dizygotic twins (DZ) and full siblings (FS).

| **Experiment** | **Modality** | **Mean Avg Prec** | | |
|---|---|---|---|---|
| | | MZ | DZ | FS |
| sMRI | T1w | 0.886 | 0.160 | 0.128 |
| | T2w | 0.908 | 0.212 | 0.111 |
| dMRI | FA | 0.964 | 0.219 | 0.160 |
| | MD | 0.803 | 0.114 | 0.086 |
| | GFA | 0.968 | 0.234 | 0.161 |
| rfMRI | netmat | 0.968 | 0.352 | 0.205 |
| Modality Combination | T1w+T2w | 0.970 | 0.283 | 0.183 |
| | T1w+FA | 0.977 | 0.279 | 0.210 |
| | FA+MD | 0.978 | 0.259 | 0.198 |
| | T1w+rfMRI | 0.990 | 0.460 | 0.279 |
| | FA+rfMRI | 0.996 | 0.472 | 0.301 |
| | T1w+T2w+DTI | 0.994 | 0.392 | 0.270 |
| | T1w+T2w+FA+rfMRI | **0.997** | **0.546** | **0.371** |
| Skull Impact | T1w Skull | 0.990 | 0.305 | 0.230 |
| | T2w Skull | 0.980 | 0.310 | 0.164 |
| Alignment Impact | T1w MNI | 0.852 | 0.087 | 0.101 |
| | T2w MNI | 0.827 | 0.147 | 0.111 |
| Resolution Impact | T1w 1.25mm | 0.831 | 0.136 | 0.121 |
| | T2w 1.25mm | 0.879 | 0.173 | 0.132 |
| Baseline Comparison | T1w | 0.649 | 0.079 | 0.052 |
| | T2w | 0.520 | 0.069 | 0.038 |
| | FA | 0.707 | 0.076 | 0.049 |
| | Vol+Thck+Area (FreeSurfer) | 0.795 | 0.172 | 0.106 |
| Retest set | T1w | 0.915 | 0.137 | 0.130 |
| | T2w | 0.917 | 0.212 | 0.113 |
| | FA | 0.944 | 0.252 | 0.158 |
| Random | Rand | 0.005 | 0.005 | 0.006 |

tation, we only report mean average precision (MAP) values, however mean recall@k results can also be found in Supplement material (Figure 1 and Table 8). Moreover, detailed results obtained with dMRI based measures (DTI and GQI), rfMRI netmats, and FreeSurfer measures are described in Table 3, 4, and 5 of Supplement material, respectively. The statistical significance of differences between MAP distributions obtained for different modality combinations and sibling types is reported in Table 1 of Supplement material.

A rich and diverse set of observations can be drawn from Table 4.2. Comparing modalities, we observe that rfMRI netmat yields the highest MAP among all single-modality fingerprints, these improvements most significant for DZ and FS. For structure-based fingerprints, T1w and T2w provide similar performances across the different sibling types, slightly higher MAP values obtained for MZ and DZ using T2w. Similarly, for diffusion based fingerprints, FA and GFA provide similar performance, both of them outperforming MD. Furthermore, higher MAP values are obtained when combining multiple modalities, the combination of T1w, T2w, FA and rfMRI having the best performance for all sibling types. This applies for combinations within/across structural or diffusion modalities: T1w+T2w outperforms T1w and T2w, FA+MD performs better than FA and MD, T1w+FA outperforms T1w and FA, etc. Similarly, T1w+rfMRI outperforms T1w and rfMRI, and FA+rfMRI performs better than FA and rfMRI.

With respect to the tested sibling types, we observe a MAP values between $80.3\%$ and $99.7\%$ when identifying MZ twins, for all modalities and their combinations. This illustrates the high impact of genetic similarity on the structural and diffusion geometry of the brain, as well as on its functional connectivity. Comparing all sibling types, we see higher MAP values for MZ twins compared to DZ twins or full siblings, following the amount of genetic information shared between subjects of these groups (Polderman *et al.*, 2015). In contrast, performances obtained for DZ twins and full siblings are comparable, which reflects the fact that both sibling types have the same genetic proximity. In general, the differences between DZ twins and full siblings were found to be not significant in an unpaired t-test for single modalities, with T2w being the exception (Supplement material Table 1). Similar observations can also be drawn from mean recall@k plots and mean recall@10 values (Supplement material Figure 1 and Table 8), with combined modalities yielding higher recall values than individual ones. In this experiment, FA gives a higher recall than rfMRI for MZ identification, although this difference is not statistically significant. Comparing non-twin siblings, we observe higher MAP values for full sibling identification vs maternal half sibling (MHS) identification (Supplement material Table 6), confirming once again the impact of genetic proximity. However, no clear trend is found for full sibling identification vs paternal half sibling identification (PHS), which is mainly due to the limited sample size (i.e., the dataset contains only 5 PHS pairs).

Table 4.3 Relative informativeness of fingerprints from different modalities. Comparison between modalities or their combination for the task of identification of a given sibling type. The reported values are **relative percentages** of MZ/DZ twin identification for two modalities, with Mod1 representing successful identifications by modality 1 only. The total number of identification tasks is 238 and 126 for MZ and DZ respectively. Note: identification of twin 1 for twin 2 and vice-versa are considered two separate tasks. The identification is considered a success if the twin is identified within the 10 nearest neighbors of a subject (among 944 subjects).

| **Experiment** | **Mod1 vs Mod2** | **Identification % (MZ)** | | | | **Identification % (DZ)** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Both | Mod1 | Mod2 | None | Both | Mod1 | Mod2 | None |
| Single Modality | T1w vs T2w | 93.28 | 2.52 | **3.36** | 0.84 | 12.70 | 13.49 | **19.05** | 54.76 |
| | T1w vs FA | 95.38 | 0.42 | **3.78** | 0.42 | 14.29 | 11.90 | **27.78** | 46.03 |
| | T1w vs rfMRI | 93.28 | 2.52 | **4.20** | 0.00 | 7.14 | 19.05 | **45.24** | 28.57 |
| | FA vs rfMRI | 96.64 | **2.52** | 0.84 | 0.00 | 26.19 | 15.87 | **26.19** | 31.75 |
| | FA vs MD | 88.66 | **10.50** | 0.84 | 0.00 | 10.32 | **31.75** | 12.70 | 45.24 |
| Modality Combination | T1w vs All MRI | 95.80 | 0.00 | **4.20** | 0.00 | 21.43 | 4.76 | **60.32** | 13.49 |
| | T2w vs All MRI | 96.64 | 0.00 | **3.36** | 0.00 | 25.40 | 6.35 | **56.35** | 11.90 |
| | FA vs All MRI | 99.16 | 0.00 | **0.84** | 0.00 | 39.68 | 2.38 | **42.06** | 15.87 |
| | rfMRI vs All MRI | 97.48 | 0.00 | **2.52** | 0.00 | 49.21 | 3.17 | **32.54** | 15.08 |

**Note**: All MRI = T1w+T2w+FA+rfMRI

To quantify the informativeness of one modality versus another, Table 4.3 reports the relative percentage of MZ and DZ twins identified by both, a single, or none of the modalities[5]. Note that, for a given twin type, each row provides relative comparison between two modalities, with sum of row being 100%. The total number of identification tasks is 238 for MZ and 126 for DZ (the identification of twin 1 for twin 2 and vice-versa are considered two separate tasks). For each task, we consider the $k = 10$ nearest neighbors of a subject in terms of fingerprint distance. The identification is considered a success if the subject's twin is identified within these neighbors. When comparing the relative success rates of single modalities (top half of the table), we observe that FA identifies more twins uniquely than when using T1w or MD.

---

[5] Results for full siblings are reported in Table 7 of Supplement material.

This is particularly noticeable for DZ twins, where $27.78\%$ of DZ pairs were identified by the FA-based fingerprint but not the T1w-based ones. Yet, structural modalities still capture brain tissue properties that are not provided by dMRI, as shown by the $11.90\%$ of all DZ pairs that are identified using T1w but not with FA. Similar observations can be drawn when comparing rfMRI with structural and diffusion modalities. For example, rfMRI identifies $45.24\%$ of DZ pairs that are not identified using T1w within $10$ neighbors, while T1w identifies $19.05\%$ unique DZ pairs.

As with the results in Table 4.2, we see that combining multiple modalities leads to a more discriminative fingerprint. For example, $4.20\%$ of MZ and $60.30\%$ of DZ twins are identified by fingerprints generated from all modalities (i.e., All MRI=T1w+T2w+FA+rfMRI) but not from fingerprints based only on T1w. Reversely, all MZ twins identified with T1w are also found using T1w+T2w+FA+rfMRI, and only $4.76\%$ of DZ twins are identified uniquely with T1w. This last result could be explained by the fact that subjects can have local similarities due to factors not related to genetics.

### 4.4.3 Impact of various factors

Factors like image alignment, scan resolution, skull inclusion and subject age, can be confounds when analyzing the proposed fingerprint. In the following sub-sections, we measure the impact of these factors on the fingerprint's ability to find genetically-related subjects.

#### 4.4.3.1 Image alignment

Population-level analyses usually require aligning images to a common space or segmenting them into regions of interest, two steps which can be computationally expensive.

Table 4.2 (sMRI vs alignment impact rows) reports the retrieval performance obtained for fingerprints generated from T1w and T2w images in MNI space (0.7 mm resolution, data provided by the HCP with affine alignment to MNI template). For all sibling types, MNI space-aligned fingerprints (denoted as MNI in the table) obtained lower MAP values than fingerprints using native space data. This observation, which is consistent across T1w/T2w modalities and all sibling types, indicates that image alignment is not required for our fingerprint. Note that similar results were obtained using full images as fingerprints (analyzed in the following section), with lower MAP for affine-aligned images.

#### 4.4.3.2 Scan resolution

Scan resolution is another important factor in multi-modal and multi-subject analyses, for example, sMRI data usually offer higher resolutions compared to dMRI.

Table 4.2 (sMRI vs resolution impact rows) shows that MAP values for the MZ/DZ twin identification task decrease when going from 0.7 mm to 1.25mm resolution, for both T1w- and T2w-based fingerprints. This is due in part to the reduced number of SIFT features extracted from 1.25 mm resolution images, compared to 0.7 mm resolution ones. However, this is not the case for FS identification tasks, where contrasting trends are observer for T1w and T2w. Moreover, differences in MAP values for the two resolutions are not significant when running an unpaired t-test with p-value $< 0.01$, for any sibling type (see Supplement material). These results suggest the robustness of our framework to varying scan resolutions.

#### 4.4.3.3 Inclusion of skull

Since skull size and shape is strongly influenced by genetics, including skull information in fingerprints could increase their discriminative power. In this experiment, we measure the usefulness of skull tissues for identifying pairs of MZ, DZ and FS subjects (facial features are not analyzed).

Table 4.2 reports the performances of fingerprints based on T1w and T2w image, with or without skull stripping. For both T1w and T2w, as well as all sibling types, including the skull in images improves MAP values. These results are significant with p-value $< 0.01$ in an unpaired t-test (see Table 2 of Supplement material). Hence, skull tissues provides additional feature correspondences which help identify twins and non-twin siblings. However, we should mention that skull stripping is essential to most neuroimaging analyses, and our objective here is only to measure the informativeness of skull tissues on the proposed fingerprint. An extended skull-inclusion analysis, including T1w-by-T2w MRI ratio images (myelin content) and modality combinations are reported in Supplement material Table 10.

#### 4.4.3.4 Subject age

In twin studies, the age of subjects can be a confound when comparing between different sibling types. For instance, DZ twins and FS siblings share the same amount of genetic material, yet DZ twins could be more similar due to their same age. The HCP data used in this study was acquired in the age range of 22–36, which corresponds to the plateau/saturation in brain and white matter development (Kochunov *et al.*, 2015; Van Essen *et al.*, 2012). Nevertheless, we analyze whether age differences in non-twin siblings is a contributing factor on performance.

Toward this goal, we divided FS sibling pairs in two groups based on the median age difference of 3 years, and measured the MAP in each group for fingerprints generated from T1w, T2w, and FA. Similarly, we also evaluated the impact of absolute age on performance of MZ/DZ. In this case, we divided subjects (not subject pairs) in two groups based on the median subject age of 29 years. As shown in Supplement material Table 9, in general no significant differences in MAP are observed across these groups. In summary, using the HCP dataset, we found no significant impact of subject age on the proposed fingerprint.

### 4.4.4 Comparison to baseline fingerprint

We compared the performance of our fingerprint to a baseline using full images as features. In this baseline, the similarity of two fingerprints is measured as the sum of squared distances (SSD) between intensities of corresponding voxels. Table 4.2 gives the MAP obtained using this baseline, for T1w, T2w, and FA images in native subject space. For MZ twin identification, the baseline using FA performs better than T1w or T2w, which is consistent with the results of the proposed fingerprint. However, we see that our fingerprint performs consistently better than the baseline, with MAP improvements of $0.237$ in T1w, $0.388$ in T2w, and $0.257$ in FA, for the task of identifying MZ twins. These improvements are significant in a one-sided unpaired t-test with p-value $< 0.01$ (see Supplement material Table 2). Note that we also tested a similar baseline created from MNI aligned images, however this led to lower MAP values.

In addition, we used Freesurfer derived measures of sub-cortical volumes, and thickness and area of cortical regions as other baseline fingerprints (see Supplement material Table 5 for detailed analysis on FreeSurfer measures). Higher MAP values are obtained for MZ twin identification using our fingerprint vs Vol+Thck+Area FreeSurfer (0.886 vs 0.795, p-value $< 0.01$). However, no significant difference is observed for DZ and FS identification.

In summary, while much more compact and efficient (see Section 4.3.3), our fingerprint based on local features is significantly more informative than a voxel-based representation. It also captures additional information on brain morphology, compared to simple measures of cortical volume, thickness and area, outperforming this baseline on all identification tasks.

### 4.4.5 Results reproducibility

To test the reproducibility of the results, we re-ran the same analysis after replacing the T1w, T2w and FA images of $42$ subjects with their retest data. Table 4.2 gives the MAP values obtained following this process. We observe small differences in MAP, compared to fingerprints using the original data, however, these are not significant (see Supplement material Table 2).

We note that the majority of retest subjects available in the HCP data are MZ twins. Since we do not observe significant differences for identifying this type of twins, it shows that the results are reproducible. The small differences in MAP values for DZ twins and FS siblings could be attributed to slight changes in the ordering of retest subjects' nearest neighbors.

### 4.4.6 Applications

In this section, we demonstrate the usefulness of our fingerprint on three different applications: 1) the correction of erroneous zygosity labels, 2) the detection of retest and duplicate scans, 3) the visualization and analysis of local feature correspondences for different modalities, sibling types and neuroanatomical regions.

#### 4.4.6.1 Zygosity label correction

The Q3 release of the HCP dataset contained self-reported zygosity labels for twins. In the HCP 1200 release, which contains genetically verified zyosity labels, it was found that many self-reported DZ twins were actually MZ twins. In light of this problem, we first evaluate if the proposed framework can be used to identify the twins in large dataset whose self-reported zygosity differs from their true zygosity.

In earlier experiments, we found higher MAP values for MZ twins. Such siblings were always found within the $10$ nearest neighbors of a subject (i.e., a mean recall@k of $100\%$ was obtained for $k \leq 10$, Supplement material Table 8), regardless of the modality combination used for the fingerprint. Conversely, a lower percentage of DZ twins could be identified in these lists of nearest neighbors. Based on this idea, we find incorrectly reported MZ candidates as the DZ twins which are within the $10$ nearest neighbors of a subject.

Table 4.4 reports the relative percentage of DZ-to-MZ twins ($56$ in total) correctly identified by the proposed fingerprint, the baseline using full images, both or none of these methods, for T1w, T2w and FA modalities. The results show that our fingerprint can identify most incorrectly self-reported MZ twins, with a detection rate of $92.86\%$ for T1w, $100.00\%$ for T2w, and $100.00\%$ for FA. For all modalities, over $32\%$ of cases were identified uniquely by our fingerprint. In contrast, no DZ-to-MZ twins were identified uniquely by the baseline fingerprint. In conclusion, the proposed fingerprint can be used effectively to detect misreported zygosity labels.

Table 4.4 Analysis of self-reported zygosity to genetically verified zygosity detection. Relative percentage of DZ-to-MZ twin identifications by the proposed framework and the full-image baseline. Total number of identification tasks is $56$. Identification is considered a success if the twin is identified within the $10$ nearest neighbors of a subject.

| **Modality** | **Identification** $\%$ | | | |
|---|---|---|---|---|
| | Both | Proposed | Base | None |
| T1w | $60.71$ | **32.15** | $0.00$ | $7.14$ |
| T2w | $55.36$ | **44.64** | $0.00$ | $0.00$ |
| FA | $64.29$ | **35.71** | $0.00$ | $0.00$ |

#### 4.4.6.2 Retest and duplicate scan identification

To analyze our fingerprint's ability to detect repeat scans of the same subjects (acquired after a time gap), we used the data of 945 subjects + 42 retest subjects, and considered the task of identifying repeat scan in a rank retrieval analysis.

Following the same evaluation protocol as for identifying MZ/DZ/FS siblings, we obtained a MAP value of 1 for fingerprints generated from T1w, T2w or FA. Thus, in all cases, the single most similar fingerprint to that of a subject corresponded to this subject's retest data. Moreover, when considering the number of local feature correspondences in the subject similarity (i.e., $\sum_{m \in \boldsymbol{M}} |\boldsymbol{B}_i^m \cap \boldsymbol{B}_j^m|$ in Eq (4.1)), we observed more correspondences for the retest data of a subject than for the subject's MZ twin.

Duplicate scans in a dataset, for example resulting from noise corruption, renaming or other manual errors, can introduce bias in analyses. Therefore, we also assessed if our fingerprint could detect duplicate scans of the same subject, corrupted by noise. For this experiment, we introduced duplicate scans for $42$ T1w images, to which was added random noise (uniformly distributed random numbers in the $[-a, a]$ range, where $a \in \{20, 60, 100, 150, 200, 400\}$; the mean and stdev of image intensities are respectively 720 and 185). Running a rank retrieval analysis using these duplicates as target, we again obtained an MAP value of 1, for all tested noise levels. As in the retest scan identification task, the number of local feature correspondences was higher with corrupted duplicates than with images of MZ twins. Compared to retest scans, the number of feature correspondences was nearly half for corrupted duplicated, suggesting that noise can reduce correspondences to some extent. Overall, the results of this experiment demonstrate that our fingerprint can preserve brain characteristics over different scans of a subject.

#### 4.4.6.3 Local feature correspondence analysis

To understand the advantages and limitations of a BoF-based fingerprint compared to voxel-wise or shape-based methods, we perform an in-depth analysis of local feature correspondences between subjects. In order to compare our findings with those of related fingerprint studies like Brainprint (Wachinger *et al.*, 2015a), we limit our analysis to genetically-related subjects from HCP and to structural MRI modalities. Other applications of BoF representations for neuro-image analysis have been well studied in the literature (Toews *et al.*, 2010; Toews and Wells, 2013; Toews *et al.*, 2015).

We start with a qualitative visualization of pairwise feature correspondences between subjects of different sibling types. The distribution of correspondences in these modalities is then analyzed using the segmentation maps (WM parcellation) files provided with HCP data. Furthermore, we also report cortical and subcortical regions having significantly different match distributions across sibling types, these regions having a closer relationship to genetic proximity. Finally, we perform a lateral asymmetry analysis in which the distribution of correspondences in hemispheres are compared. Since fMRI is not required for these analyses, we considered all subjects in the HCP twin dataset having genetically verified labels (only 945 out of 1010 subjects have rfMRI netmats data), giving a total of 139 MZ pairs, 72 DZ pairs, and 1214 full sibling pairs.

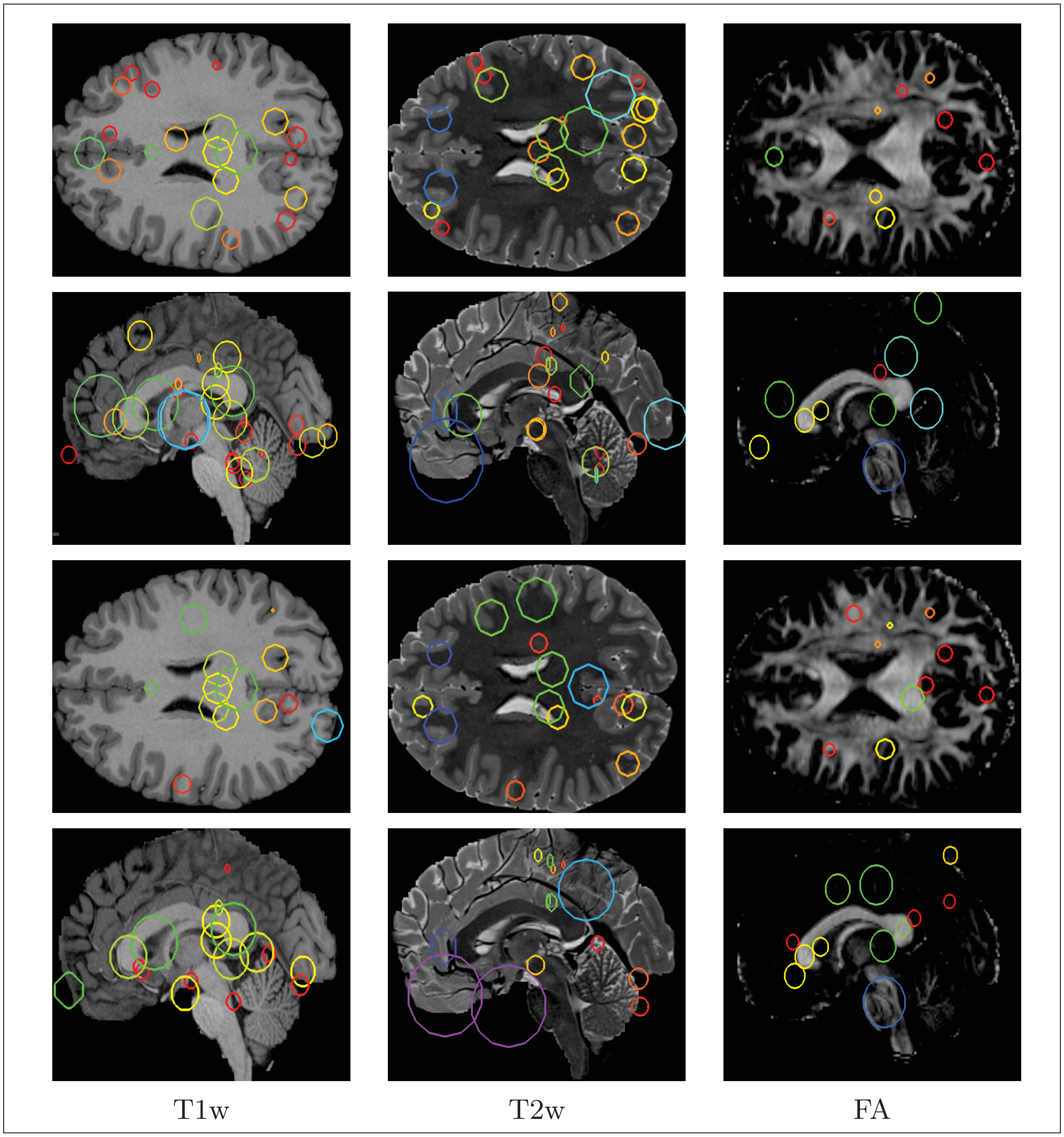


Figure 4.4 Example of feature correspondences for a subject and his/her MZ twin (rows 1-2), and the subject's full sibling (FS) (rows 3-4). Scale space is represented using circle radius (for the visible slice).

### Scale-space visualization of features correspondences

Analyzing local feature correspondences between sibling pairs provides information in terms of their location as well as scale. In 3D SIFT features, scale coincides with the variance of a

Gaussian blur kernel for which the corresponding voxel in the blurred image is a local extrema (Lowe, 2004, 1999). It thus corresponds to a certain degree with the size of structures in which these features are located.

Figure 4.4 gives a scale-space visualization of features matched between a subject and his/her MZ twin, as well as the subject's non-twin (full) sibling, for T1w, T2w and FA images (See Supplement material Figure 2 for DZ and non-twin (full) sibling). The scale information is represented using the radius of circles. Note that circles represent the intersection of 3D spheres with the visible slice and, thus, non-intersecting features are hidden in this 2D visualization.

It can be seen that different image modalities generally result in distinct, complementary feature correspondences throughout the brain. In T1w and T2w images, features are mainly located in the frontal lobe, corpus callosum and cerebellum. Smaller-scale features are also visible along various cortical regions, as well as in subcortical structures near the basal ganglia. Moreover, images based on diffusion measures have less correspondences than in structural modalities. These correspondences are located mostly inside or near to white matter: larger-scale features in the corpus-callosum, and smaller-scale ones in the brain stem and along white matter bundles. The distribution of features in prominent brain regions is further analyzed in the next section.

Comparing different sibling types, we see a greater number of correspondences between MZ twins than between DZ twins or full siblings. This observation, which is easier to visualize in T1w and T2w images, is consistent with other analyses on twin datasets. In terms of feature location and scale, we observe a slightly higher number of correspondences in the frontal cortex for MZ twins, however, no obvious pattern can be drawn from one set of representative plots.

### Region-wise analysis of feature correspondences

Here, we analyze the distribution of feature correspondences across atlas-defined neuroanatomical regions, measured over the entire group of subjects. For each scan, segmentation labels were obtained from the Freesurfer-processed data, using LUT table for label descriptions.

Figure 4.5 shows the box plot distributions of feature correspondences between pairs of MZ, DZ and full siblings, for T1w and T2w images. Feature match counts are reported for five broad regions: non-white matter subcortex (`s-cort`), left/right cortex (`crtx-lh/rh`) and left/right white matter (`wm-lh/rh`). Note that mapping local features to a finer cortical parcellation is difficult due to the limited thickness of the cortex. Subcortical regions are further analyzed below.

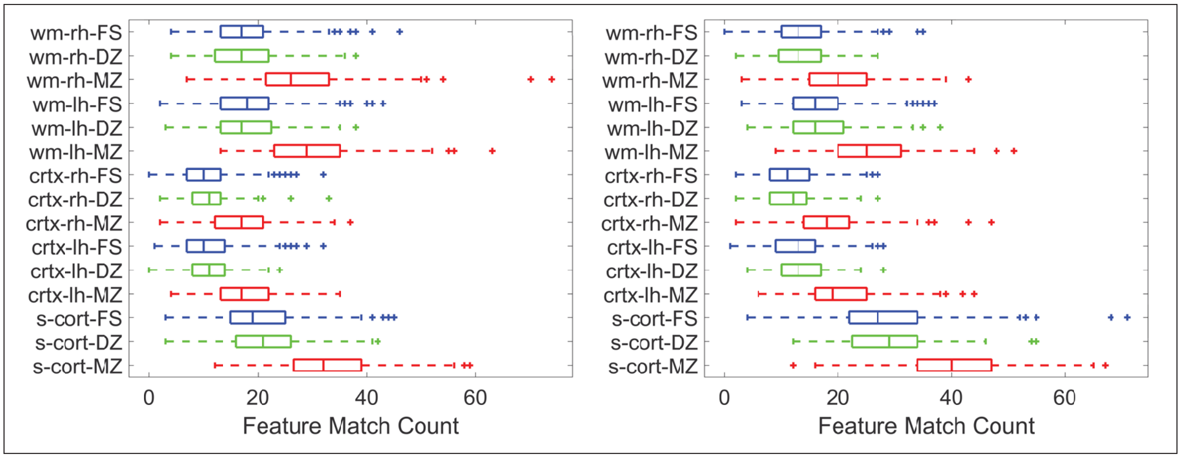


Figure 4.5 Box plot comparison between MZ, DZ, and FS for pairwise feature correspondence counts for T1w (left) and T2w (right) for major structures. Red, green and blue correspond to MZ, DZ, and FS respectively.

Comparing across sibling types, we observe a higher number of feature correspondences for MZ pairs across all five regions and both T1w and T2w modalities. This confirms once again that the local features employed in our fingerprint captures brain characteristics related to genetic proximity. Analyzing the region-wise distribution of feature correspondences, all five regions are well represented. Since the number of local features in a region is proportional to its size, it is not surprising that the cortex has the least correspondences. Yet, such features are also produced by intensity variations (i.e., edges), thus explaining why many correspondences are found in the cortex. Finally, when comparing T1w and T2w modalities, we see small differences in the match counts, however these are not statistically significant.

Table 4.5 Differences in feature match counts obtained for different sibling types in various brain regions, using T1w and T2w. We report Holm–Bonferroni corrected p-values (-$\log_{10}$ scale) measured using an unpaired t-test. Significant results with corrected p-value $< 0.01$ are highlighted using bold font.

| **Label** | **T1w** | | **T2w** | |
|---|---|---|---|---|
| | MZ vs DZ | MZ vs FS | MZ vs DZ | MZ vs FS |
| subcortical | **29.31** | **50.31** | **26.06** | **39.41** |
| Crtx-LH | **22.85** | **35.17** | **23.37** | **38.87** |
| Crtx-RH | **21.48** | **39.76** | **25.38** | **37.73** |
| WM-LH | **37.64** | **62.83** | **27.88** | **47.38** |
| WM-RH | **23.38** | **36.80** | **21.88** | **32.62** |
| L-Lat-Vent | **5.84** | **11.34** | **5.31** | **7.50** |
| R-Lat-Vent | **4.21** | **10.72** | **3.99** | **7.57** |
| L-VentralDC | 1.49 | **6.06** | **5.16** | **2.98** |
| R-VentralDC | 0.45 | 0.60 | 0.00 | 0.01 |
| R-Cerebellum-WM | **3.98** | **15.40** | 0.00 | 0.57 |
| L-Cerebellum-WM | **4.82** | **11.11** | **2.33** | **6.55** |
| R-Putamen | 0.48 | 0.51 | **2.34** | 1.24 |
| L-Putamen | 0.87 | 0.35 | 0.06 | 0.30 |
| L-Cerebellum-Crtx | **5.74** | **6.26** | **5.58** | **13.81** |
| L-Thalamus-Proper | **2.71** | **4.03** | 0.37 | 0.01 |
| 4th-Ventricle | 1.49 | **2.24** | 1.86 | **3.91** |
| L-Hippocampus | **3.23** | **3.76** | **4.61** | **5.85** |
| CC-Anterior | 1.83 | 0.51 | 0.40 | 0.71 |
| R-Cerebellum-Crtx | **5.96** | **11.93** | **3.38** | **7.57** |
| 3rd-Ventricle | 0.45 | 0.51 | 0.37 | 0.43 |

To identify regions showing a strong relationship to genetic proximity, Table 4.5 gives the p-values (-$\log_{10}$ scale) of an unpaired t-test comparing the mean number of correspondences between subjects of a given sibling type versus another sibling type (e.g., MZ vs DZ). Significance values are provided for the five major regions described above, as well for 15 prominent subcortical structures matching the analysis by Wachinger et al. (Wachinger *et al.*, 2015a). To account for multiple comparisons (i.e., one for each tested region), reported p-values have been corrected using the Holm-Bonferroni procedure (Holm, 1979). Moreover, to account for age and size bias in this analysis, we selected FS pairs with less than 3 years age difference, and matched the number of FS pairs to MZ pairs using a simple bipartite matching based on age.

From Table 4.5, we observe significant differences between MZ twins and DZ-twins/full-siblings (i.e., corrected p-value $< 0.01$), for all five major regions and for both T1w and T2w images. In subcortical structures of T1w images, cerebellum white matter and cortex (left and right), lateral ventricles (left and right), left hippocampus and left thalamus proper have a significantly different number feature correspondences in MZ twins than in DZ twins or FS subjects. Comparing results obtained with T1w and T2w, the same structures are significant across both modalities, differences in significance reflecting the complimentary of these modalities.

### Hemisphere asymmetry analysis

In our last experiment, we analyze the symmetry of feature match counts across brain hemispheres, for major structures. Toward this goal, we considered only right-handed (RH) subjects, and limited sibling pairs to subjects with same gender (i.e., a male and his brother, or a female and her sister). For non-twin siblings, we also restricted our analysis to subject pairs with less than $3$ years of age difference.

Table 4.6 gives the results of two-sided unpaired t-tests comparing the feature match counts between cortical or white matter regions (Freesurfer LUT labels) in left- and right- hemispheres. To analyze gender effects, we also report results individually for RH male siblings and RH female siblings. Overall, we observe significant asymmetry in white matter regions (with corrected p-value $< 0.01$) of MZ twins, the highest significance values obtained for T2w images. No clear pattern is found across sibling types, although hemispherical differences are generally higher in MZ twins than in DZ twins or full siblings. Likewise, no conclusion can be drawn

Table 4.6 Hemisphere asymmetry analysis. For a given modality and twin type, we compare feature match count differences across hemisphere for major structures. Differences are reported as Holm–Bonferroni corrected p-values ($-\log_{10}$ scale) measured using an unpaired t-test, significant results (corrected p-value $< 0.01$) highlighted using bold font.

| **Modality** | **Type** | **RH Female** | | **RH Male** | | **RH Pairs** | |
|---|---|---|---|---|---|---|---|
| | | Crtx | WM | Crtx | WM | Crtx | WM |
| T1w | MZ | 0.95 | 1.20 | 0.15 | **2.57** | 0.87 | **2.73** |
| | DZ | 1.05 | 0.78 | 0.35 | 0.99 | 1.12 | 0.06 |
| | FS | 1.71 | 0.39 | 0.84 | 0.11 | 1.89 | 0.09 |
| T2w | MZ | 1.95 | **9.13** | 1.52 | **5.77** | **3.00** | **13.97** |
| | DZ | 1.06 | **3.60** | 1.29 | 1.23 | 1.93 | **4.06** |
| | FS | 1.04 | 1.22 | 1.23 | **5.35** | 1.90 | **5.84** |

when comparing results for male and female sibling pairs, with significance values varying across different sibling types and modalities.

The asymmetry of function in the brain, for example the hemispheric specializations of language and motor functions, has been extensively studied (Toga and Thompson, 2003). Similarly, studies have analyzed anatomical brain asymmetries based on voxel-based morphometry, sulci and other brain features (Wachinger *et al.*, 2015a). The multi-modal and multi-region analysis presented in this work extends previous studies of brain asymmetry in the literature by considering sibling types. Accounting for various confounds, including gender, genetics, handedness and age, this analysis has shown a greater asymmetry in feature correspondences between MZ twins than DZ twins and full siblings, mostly found in white matter regions and T2w images. Moreover, differences in asymmetry appear to be directional.

## 4.5 Discussion

In this section, we summarize the findings of this study and emphasize their link to previous investigations. We also highlight its limitations and discuss additional considerations.

### Identification of genetically-related subjects

Our experiments on the task of identifying genetically-related subjects led to various useful observations. We established that the proposed fingerprint, generated from individual modalities or their combination, respects the genetic relationships between siblings, with MZ twins being more similar than DZ twins or full siblings (Peper *et al.*, 2007; Thompson *et al.*, 2013).

3D SIFT features (i.e., keypoints) coincide with the local extrema of a difference of Gaussians function applied in scale space. These features typically lie in high-contrast regions of an image, for instance due to the boundaries between white matter and grey matter (see Figure 4.4). More generally, these features represent blob-like structures of varying size and location, which are robust and discriminative for finding correspondences across images. With respect to a voxel-wise full image comparison, the proposed BoF-based fingerprint offers a more compact representation of brain geometry, which is less sensitive to differences in image alignment and contrast. Likewise, compared to standard morphological measures like cortical thickness or sub-cortical region volume/area, our representation may capture a broader range of geometrical brain characteristics, for example distinctive cortical folding patterns only present in a subset of the population.

Analyzing the manifold approximation, we also showed that a discriminative fingerprint could be obtained with only $150$ spectral components (i.e., leading eigenvectors of the normalized adjacency matrix of the subject proximity graph). When compared to a baseline using full images as features, this compact fingerprint yielded significantly better performances, for all modalities and sibling types. This illustrates the high efficiency of our fingerprint and its advantages for comparing large groups of subjects. Moreover, while Laplacian eigenmaps were used to embed the subject proximity graph, the proposed framework is generic and other approaches (e.g., see (Bengio *et al.*, 2013)) can be employed for this task.

The comparison of fingerprints obtained from structural MRI, diffusion MRI, and resting state fMRI highlighted the informativeness and complementarity of these modalities. Among individual modalities, resting state fMRI based fingerprint performed best for DZ/FS identification and had similar performance to FA/GFA for MZ twin identification. As mentioned in Finn et al. (Finn *et al.*, 2015), this could be due to the discriminative power of connectivity profiles, which is a result of integrating imaging data over a relatively long period of time (4800 volumes, and 4 runs of 15 minutes each). The inter-individual variability reflected by the connectivity profile (rfMRI netmat) is dominated by the spatial topography (spatial variability in the location of functional regions across individuals) rather than the coupling strength. We refer readers to the work of Bijsterbosch et al. (Bijsterbosch *et al.*, 2018) to understand the specific contribution of functional coupling and spatial topography in rfMRI netmats. Moreover, while the MAP values for FA/GFA are similar to rfMRI based MZ twin identification, mean recall@10 and relative identification $\%$ showed that FA performs slightly better than rfMRI ($2.54\%$ unique MZ pair identification as opposed to $0.84\%$ pairs).

Comparing structural and diffusion MRI modalities, we found fingerprints based on FA/GFA to outperform those derived from T1w or T2w. We hypothesize this is caused by the pronounced contrast/magnitude changes in FA maps occurring at the boundary between grey matter and white matter (e.g., endpoints of fiber bundles). As shown qualitatively in Figure 4.4, this leads to a more evenly-distributed set of feature matches.

Another interesting observation is the higher MAP values obtained for the identification of DZ twins compared to full siblings, although both sibling types have the same genetic similarity. In Supplemental material Table 1, this difference is found to be significant for several modality combinations (e.g., T1w+T2w+FA+rfMRI, p-value $< 0.001$). While our experiments accounted for group size and age differences by matching DZ subject pairs with FS pairs having the smallest age differences (one to three years difference), remaining age differences may explain this observation. Further investigation is however required to fully validate this hypothesis.

This work is motivated by the recent increase in multi-modal brain studies. For instance, multi-modal MRI has been shown useful for the analysis of neurodegenrative disorders (Calhoun and Sui, 2016) as well as for identifying subjects with schizophrenia (Sui *et al.*, 2014). Results of this study demonstrate the usefulness of combining multiple modalities in a brain fingerprint. The improvements due to multi-modal combination, for all the twin/sibling types, can be attributed to more comprehensive characterization considering structure, white matter architecture, and functional connectivity. Thus, better performances were obtained with a combined set of modalities than with these modalities alone. Our results are consistent with previous studies underlining the benefit of a multi-modal fusion (Calhoun and Sui, 2016; Groves *et al.*, 2012). As a note, we have focused on major observations only, the comprehensive analysis is open to various other observations including comparison of DTI vs GQI measures, inclusion of T1w-by-T2w MRI ratio images, FreeSurfer measures based identification, etc.

### Applicability of the proposed fingerprint

Our factor impact analysis demonstrated the robustness of the proposed fingerprint to the non-alignment of images. Since image alignment is key for most population level analysis (Dubois and Adolphs, 2016), by alleviating this requirement, the proposed fingerprint may help save computational costs and avoids errors introduced during alignment. Experiments have also shown that scan resolution (from 0.7mm to 1.25mm) does not have a significant impact on results, although using lower resolution images reduces the number detected features. Data acquired from multiple sites or scanners often need to be brought to same resolution, introducing small errors during interpolation and re-sampling. The proposed fingerprint may thus be of help for multi-site studies, and pave the way to resolution-independent analyses. Lastly, using retest scans led to no significant changes in results, further validating the robustness of our fingerprint to image acquisition. However, a detailed longitudinal analysis with longer between-scan times would be required to fully confirm this claim.

The proposed rank retrieval analysis based on MAP provides a principled approach for comparing different brain fingerprints, which could be utilized in future studies. In this work, we used the proposed fingerprint to find incorrectly reported zygosity labels and identify retest/duplicate scans of the same subjects. Hence, our fingerprint could serve as efficient and reliable tool for detecting inconsistent information in large cohorts. Another potential application could be to provide physicians with related cases in clinical settings like MCI diagnostic assistance (Gao *et al.*, 2015).

While various twin studies have analyzed genetic influences based on volume, cortical thickness, surface area, and morphometry (Wachinger *et al.*, 2015a), this is the first work to use local features and manifold approximation for this problem. Analyzing the distribution of features correspondences across brain regions, in images of different modalities, reveals many interesting insights. Results identify various neuroanatomical regions (e.g., cerebellum, lateral ventricles, ventral diencephalon, hippocampus and thalamus proper) having significantly different match counts in MZ twins than DZ twins or full siblings. These findings relate to those reported in (Wachinger *et al.*, 2015a), which were obtained on a different dataset (mean subject of age of $56$ years, compared to a median of $29$ years in the HCP dataset). Another key aspect of our analysis is the size of the subject cohort, larger than that of related studies (Peper *et al.*, 2007).

### Additional considerations

In this work, we used a rank retrieval analysis to evaluate the relation between fingerprint similarity and genetic proximity. Mean recall@k and mean average precision (MAP) were employed to measure sensitivity, specificity, and relative informativeness of fingerprints generated from different modality combinations. However, estimating heritability directly, for instance using the approach described in (Ge *et al.*, 2016), would provide a better quantification of genetic influence on fingerprint features. In (Elliott *et al.*, 2017), Elliott et al. considered the data of over $8,000$ from the UK Bio-bank (Sudlow *et al.*, 2015) to determine the heritability of multi-modal brain imaging phenotypes. Similarly, Colclough et al. report in (Colclough *et al.*, 2017) the heritability of multi-modal functional connectivity profiles using 800 HCP subjects. An extensive analysis is required to asses the heritability of the proposed fingerprint and relate our findings to those in these recent studies.

Moreover, when building the subject proximity graph, we assumed the independence of feature correspondences across modalities. However, a deeper analysis could be carried out to investigate false feature correspondences and correlation between features correspondences across modality. As mentioned before, other manifold embedding methods like Locally Linear Embedding (LLE) (Roweis and Saul, 2000) could also be employed for this step.

In this study we analyzed data from sMRI, dMRI and rfMRI. However, the proposed framework is generic and could be extended to other modalities like task-fMRI, PET-MRI and quantitative T1/T2 maps. Finally, this study focused on comparing and combining different modalities for identifying genetically-related subjects, misreported zygosity labels and duplicate/restest scans. An interesting extension of this work would to be to assess whether our fingerprint can be used as a biomarker to identify subjects with cognitive or neurological disorders. Publicly available data, for instance from the ADNI dataset (Toews *et al.*, 2010) or Parkinson's Progression Markers Initiative (PPMI) dataset (Marek *et al.*, 2011), could be used for this analysis.

## 4.6 Conclusion

We presented a brain fingerprint, based on manifold approximation, for the multi-modal analysis of genetically-related subjects. In a rank retrieval analysis, mean recall@$k$ and mean average precision were used to measure the relation between fingerprint similarity and genetic proximity, as well as the contribution/complementarity of information from different MRI modalities. Results indicated that a compact fingerprint of only 150 features could identify genetically-related subjects better than a baseline using full images as features. Our experiments also showed that each modality provides complementary information which can uniquely identify some sibling pairs. Furthermore, we demonstrated the benefit of considering multiple modalities in the fingerprint, combined modalities leading to a better performance than considering these modalities separately. Moreover, our analysis demonstrated the robustness of the proposed fingerprint to various factors, including image alignment, scan resolution and subject age. The reproducibility of results was also confirmed using retest scans from the HCP dataset, showing our fingerprint to be robust to variability in image acquisition.

The usefulness of our fingerprint was assessed on the tasks of identifying incorrectly reported zygosity and retest/duplicate scans in large dataset. Results of this experiment highlighted the effectiveness of our fingerprint, with MAP values near 100% for all test cases. Moreover, analyzing the distribution of features correspondences across the brain revealed neuroanatomical regions (e.g., cerebellum, lateral ventricles, ventral diencephalon, hippocampus and thalamus proper) with significantly different match counts in MZ twins compared to DZ twins or full siblings. This work could be extended by further investigating the differences, in terms of feature location and similarity, between dizygotic twins and non-twin siblings. A deeper analysis of aging effects could also be performed, for instance, using longitudinal data. Such analysis would help understand the effect of neuroplasticity on individual brain characteristics.

## 4.7 Supplement results

### 4.7.1 Mean Average Precision results

#### 4.7.1.1 Significance testing across twin/sibling types

Table 4.7 reports the unpaired t-test results for twin/sibling vs twin/sibling comparisons for individual modalities or multi-modal combinations using average precision values. For this analysis, the samples sizes are matched using simple bipartite matching, for each comparison (number of MZ pairs and FS pairs are matched to DZ pairs), based on age, and are fixed across modalities.

Table 4.7 Significance testing across twin/sibling types for a given modality, using the distribution of average precision (AveP) values obtained for the the task of twin/sibling identification. We report -$\log_{10}$ p-values for unpaired t-test for MZvsDZ, MZvsFS, and DZvsFS for each modality. Note the samples sizes are matched for each comparison (number of MZ pairs and FS pairs are matched to DZ pairs) based on age, and are fixed across modalities. **p-values $\leq$ 0.01** are in bold.

| **Experiment** | **Modality** | **Significance** | | |
|---|---|---|---|---|
| | | MZ vs DZ | MZ vs FS | DZ vs FS |
| sMRI | T1w | **51.63** | **57.14** | 0.30 |
| | T2w | **43.72** | **72.73** | **2.68** |
| dMRI | FA | **56.56** | **69.82** | 1.14 |
| | MD | **48.54** | **50.56** | 0.38 |
| | GFA | **47.14** | **66.66** | 1.27 |
| rfMRI | netmat | **34.32** | **47.08** | 1.26 |
| Modality Combination | T1w+T2w | **41.55** | **77.29** | **3.06** |
| | T1w+FA | **45.46** | **58.03** | 1.50 |
| | FA+MD | **48.01** | **58.55** | 1.00 |
| | T1w+rfMRI | **27.86** | **43.21** | **2.93** |
| | FA+rfMRI | **26.41** | **41.69** | **2.57** |
| | T1w+T2w+DTI | **36.50** | **50.19** | **3.09** |
| | T1w+T2w+FA+rfMRI | **24.39** | **37.86** | **3.70** |
| Skull Impact | T1w Skull | **41.72** | **53.70** | 1.51 |
| | T2w Skull | **37.98** | **65.90** | **2.86** |
| Alignment Impact | T1w MNI | **62.52** | **64.75** | 0.41 |
| | T2w MNI | **44.29** | **51.40** | 0.89 |
| Resolution Impact | T1w 1.25mm | **48.45** | **54.77** | 1.60 |
| | T2w 1.25mm | **50.29** | **61.66** | 1.22 |
| Baseline Comparison | T1w | **26.92** | **29.65** | 0.98 |
| | T2w | **19.22** | **22.22** | 1.51 |
| | FA | **40.26** | **42.93** | 1.35 |
| | Vol+Thck+Area | **40.54** | **49.54** | 1.36 |
| Retest set | T1w | **65.07** | **70.14** | 0.33 |
| | T2w | **50.88** | **81.68** | **2.22** |
| | FA | **47.37** | **68.56** | 1.91 |
| Random | Rand | 0.08 | 0.11 | 0.02 |

We observe that the average precision values for mono zygotic twins are statistically significant compared to di-zygotic twins or full siblings. Comparing di-zygotic twins and full siblings, we observe that in general for single modalities they are not statistically significant and for multi-modal combinations they are statistically significant. Since DZ and FS pairs differ in terms of age (DZ pair subjects have same age, while FA pair subjects may have different age also) and to some extent shared environment (due to age difference). The observed change in statistical significance from single to multi-modal combination could be due to complimentary information provided by additional modalities enhancing the differences also (in addition to increasing the similarities within pairs as reflected by increased MAP values).

#### 4.7.1.2 Significance testing: Modality vs Modality comparisons

Table 4.8 reports significance results (-$\log_{10}$ p-values) for modality vs modality comparisons, for a given sibling type, based on the average precision values. The results support the observations drawn from MAP values and mean recall@k plots. In general, multi-modal combinations lead to better non-twin sibling identification as compared to improvements in other sibling types. For example, comparing T1w vs T1w+FA, we observe that average precision values for non-twin identification are statistically significant for all three sibling types, with highest improvement for non-twin siblings (-$\log_{10}$ p-value = $9.83$).

#### 4.7.1.3 DTI vs HARDI dMRI measures

Table 4.9 compares the performance of Diffusion Tensor Imaging (DTI) (Alexander *et al.*, 2007) and Generalized Q-Ball Imaging (GQI) (Yeh *et al.*, 2010) based indices obtained in native space (skull stripped) for the task of genetically related subject identification. Four widely used DTI based measures were extracted to characterize white matter micro-structure: fractional anisotropy (FA), axial diffusivity (AD), mean diffusivity (MD) and radial diffusivity (RD). The interpretation of these measures are discussed in (Alexander *et al.*, 2007). Utilizing the high angular resolution of HCP data, following GQI (Yeh *et al.*, 2010) based measures were also extracted: Generalized Fractional Anisotropy (GFA), Quantitative Anisotropy (QA), Normalized Quantitative Anisotropy (NQA0), isotropic component of the ODF (ISO), Restricted Diffusion Imaging (RDI), and Non-Restricted Diffusion Imaging (NRDI). GFA extends the standard FA measure to orientation distribution functions (ODF) based on spherical harmon-

Table 4.8 Modality vs modality comparison and contrast. Comparisons between two modalities (or their combinations), for a given sibling type, for the task of identification of a given sibling type. $-\log_{10}$ p-values are reported for unpaired t-test performed using the distribution of average precision values. **p-values $\leq$ 0.01** are in bold. (All MRI= T1w+T2w+FA+rfMRI)

| **Experiment** | **Modality vs Modality** | **Significance** | | |
|---|---|---|---|---|
| | | MZ | DZ | FS |
| Single | T1w v T2w | 0.43 | 0.66 | 1.01 |
| | T1w v FA | **3.64** | 0.85 | **2.18** |
| | T1w v rfMRI | **3.92** | **4.60** | **9.12** |
| | FA v rfMRI | 0.12 | **2.45** | **3.20** |
| | FA v GFA | 0.12 | 0.14 | 0.02 |
| | FA v MD | **9.36** | **2.47** | **11.05** |
| Modality Combination | T1w v T1w+T2w | **4.27** | **2.29** | **5.42** |
| | T1w v T1w+FA | **5.23** | **2.27** | **9.94** |
| | FA v T1w+FA | 0.52 | 0.79 | **3.73** |
| | FA v FA+MD | 0.55 | 0.47 | **2.47** |
| | T1w v All MRI | **8.37** | **15.48** | **62.20** |
| | T2w v All MRI | **6.45** | **10.91** | **72.99** |
| | FA v All MRI | **2.69** | **11.46** | **44.24** |
| | rfMRI v All MRI | 1.98 | **3.96** | **26.36** |
| Skull Impact | T1w v T1w Skull | **7.25** | **3.04** | **14.24** |
| | T2w v T2w Skull | **3.93** | 1.38 | **5.38** |
| Alignment Impact | T1w v T1w MNI | 0.69 | 1.54 | **2.02** |
| | T2w v T2w MNI | **2.57** | 0.96 | 0.01 |
| Resolution Impact | T1w v T1w 1.25mm | 1.26 | 0.32 | 0.31 |
| | T2w v T2w 1.25mm | 0.63 | 0.46 | 1.31 |
| Baseline Comparison | T1w v Base | **11.10** | 1.91 | **15.46** |
| | T2w v Base | **26.38** | **4.10** | **16.71** |
| | FA v Base | **16.88** | **4.49** | **27.09** |
| | T1w v VTA FreeSurfer | **2.57** | 0.12 | 1.43 |
| Retest set | T1w v T1w Retest | 0.62 | 0.28 | 0.07 |
| | T2w v T2w Retest | 0.16 | 0.00 | 0.09 |
| | FA v FA Retest | 0.67 | 0.37 | 0.08 |

ics. NQA is a scaled version of quantitative anisotropy, which is calculated from the peak orientations on a spin distribution function (SDF). More information about these measures can be found in (Yeh *et al.*, 2010) and DSI studio documentation [6]. The reconstruction was performed using DSI studio toolbox. DTI reconstruction option was used for computing: FA,

---

[6] http://dsi-studio.labsolver.org/Manual/Reconstruction

Table 4.9 Mean average precision (MAP) table comparing diffusion MRI based measures: DTI and GQI based indices (native space) for the task of genetically related subject identification.

| **Experiment** | **Modality** | **Mean Avg Prec** | | |
|---|---|---|---|---|
| | | MZ | DZ | FS |
| DTI | FA | 0.964 | 0.219 | 0.160 |
| | AD | 0.701 | 0.112 | 0.083 |
| | MD | 0.803 | 0.114 | 0.086 |
| | RD | 0.840 | 0.141 | 0.113 |
| DTI Combination | FA+MD | 0.978 | 0.259 | 0.198 |
| | FA+MD+RD+AD | 0.971 | 0.332 | 0.232 |
| GQI | GFA | 0.968 | 0.234 | 0.161 |
| | QA | 0.929 | 0.165 | 0.141 |
| | NQA | 0.899 | 0.177 | 0.152 |
| | ISO | 0.820 | 0.113 | 0.103 |
| | RDI | 0.828 | 0.147 | 0.138 |
| | NRDI | 0.856 | 0.165 | 0.108 |
| GQI Combination | GFA+NQA+RDI | **0.991** | **0.333** | **0.258** |

MD, RD, and AD. While GFA, QA, NQA, ISO, RDI (rdi02L), and NRDI (nrdi02L) were computed using GQI reconstruction option. “rdi02L” quantified the restricted diffusion within “0.2 L”, where L is the diffusion distance. “nrdi02” quantifies non-restricted diffusion with displacement greater than “0.2 L”.

Overall we observe that, first, FA and GFA perform best among DTI and GQI measures respectively. Second, comparing across DTI and GQI, GFA perfroms better than FA, this could be due to the fact that GFA is generalization of FA to high angular resolution data. Third, combination of DTI or GQI measures perform better than individual measures with GQI combination reaching highest MAP values. Various other observations can be drawn from Table 4.9, for example, NRDI performs better than RDI, and RD performs better than MD and AD.

#### 4.7.1.4 rfMRI network matrices as fingerprints

Table 4.10 reports mean average precision results for functional connectivity profiles as fingerprints. We use the netmats provided on HCP website, and compare the performance across correlation (full vs partial) and node sizes, for the task of identification of genetically related subjects. Similar to the analysis reported in Finn et al. (Finn *et al.*, 2015), we use the Pearson Correlation to compute similarity between subjects' netmats. The pairwise similarity matrix serves as subject proximity graph.

Table 4.10 Functional Connectivity Profiles as fingerprint: Impact of Nodes and correlation. Mean average precision (MAP) table comparing resting state fMRI network matrices (netmats) for the task of genetically related subject identification. Manifold approximation is obtained in the form of a subject proximity graph by computing similarity between netmats using the Pearson Correlation (Finn *et al.*, 2015).

| **Experiment** | **Modality** | **Mean Avg Prec** | | |
|---|---|---|---|---|
| | | MZ | DZ | FS |
| rfMRI full corr | ICA 15 | 0.349 | 0.084 | 0.040 |
| | ICA 25 | 0.470 | 0.091 | 0.045 |
| | ICA 50 | 0.597 | 0.115 | 0.060 |
| | ICA 100 | 0.706 | 0.141 | 0.082 |
| | ICA 200 | 0.779 | 0.183 | 0.106 |
| | ICA 300 | 0.824 | 0.241 | 0.116 |
| rfMRI partial corr | ICA 15 | 0.462 | 0.057 | 0.045 |
| | ICA 25 | 0.645 | 0.098 | 0.070 |
| | ICA 50 | 0.859 | 0.242 | 0.128 |
| | ICA 100 | **0.968** | **0.352** | **0.205** |
| | ICA 200 | 0.968 | 0.340 | 0.205 |
| | ICA 300 | 0.869 | 0.179 | 0.117 |

We draw following major observations: first, partial correlation netmats works better than full correlation netmats. Second, compared to small number of nodes (15 and 25), high resolution parcellations (more nodes) give better MAP values (Finn *et al.*, 2015). Third, for partial correlation netmats, we observe a peak in mean average precision around 100 nodes. The maximum of MAP values around 100 nodes could reflect that increasing nodes doesn't necessarily add additional information and may be adding noise to individual profiles. We have used rfMRI netmats based on partial correlation and 100 ICA nodes in this work.

#### 4.7.1.5 FreeSurfer derived measures as fingerprint

Table 4.11 shows the MAP values obtained in a rank retrieval of three different siblings types (MZ, DZ and FS), using fingerprints generated from FreeSurfer measures. We use the unrestricted csv file and consider volume of sub-cortical structures, thickness and area measures for cortical parcellations. Each of the measures are converted to zscore across subjects, and then used as a fingerprint (volume measures are ICV corrected by first dividing by ICV). Subject proximity graph is approximated by computing the pairwise Pearson correlation.

Table 4.11 Freesurfer derived measures as fingerprint. Mean average precision (MAP) table comparing freesurfer derived measures: Vol, Thck, Area, Vol+Thck+Area as fingerprints for the task of genetically related subject identification. The measures are obtained from the csv file provided by HCP (each measure converted to zscore column-wise i.e. across subjects, and NxN subject proximity graph computed using pairwise Pearson correlation, Vol measures are ICV corrected.)

| **Experiment** | **Modality** | **Mean Avg Prec** | | |
|---|---|---|---|---|
| | | MZ | DZ | FS |
| Fressurfer Meas | Vol | 0.654 | 0.116 | 0.068 |
| | Thck | 0.356 | 0.085 | 0.053 |
| | Area | 0.367 | 0.126 | 0.062 |
| | Vol+Thck+Area | **0.795** | **0.172** | **0.106** |

We observe the MAP values follow a similar trend as reported for proposed fingerprints, with MZ twins having higher values as compared to DZ or FS. Comparing across measures, sub-cortical volumes show a better performance when compared to cortical thickness or area measures. However, the combination of Volume, Thickness and Area measures is most discriminative.

#### 4.7.1.6 Comparison of sibling types: FS, MHS, PHS

Table 4.12 shows MAP values for the task of identification of sibling types. We report results for 546 full-sibling (FS) pairs, 39 maternal half sibling (MHS) pairs, and 5 paternal half sibling (PHS) pairs. While the MAP values are higher than the Random ordering based identification,

Table 4.12 Non-twin sibling comparison. Mean average precision (MAP) table comparing different modalities for the task of non-twin sibling identification (full-siblings (FS), maternal half-siblings (MHS), paternal half-siblings(PHS)).

| **Experiment** | **Modality** | **Mean Avg Prec** | | |
|---|---|---|---|---|
| | | FS | MHS | PHS |
| sMRI | T1w | 0.128 | 0.008 | 0.051 |
| | T2w | 0.111 | 0.022 | 0.053 |
| dMRI | FA | 0.160 | 0.015 | 0.131 |
| | MD | 0.086 | 0.041 | 0.178 |
| | GFA | 0.161 | 0.020 | 0.231 |
| rfMRI | netmat | 0.205 | 0.063 | 0.037 |
| Modality Combination (1.25mm) | T1w+T2w | 0.183 | 0.012 | 0.132 |
| | T1w+FA | 0.210 | 0.015 | 0.237 |
| | FA+MD | 0.198 | 0.044 | 0.270 |
| | T1w+rfMRI | 0.279 | 0.059 | 0.054 |
| | FA+rfMRI | 0.301 | 0.058 | 0.168 |
| | T1w+T2w+DTI | 0.270 | **0.072** | **0.439** |
| | T1w+T2w+FA+rfMRI | **0.371** | 0.066 | 0.350 |
| Random | Rand | 0.006 | 0.004 | 0.005 |

Table 4.13 Relative informativeness of fingerprints from two modalities. Percentage of full sibling (FS) identification for two modalities. Total number of identification tasks is 1092. We consider 10 nearest neighbors, and if the sibling (one for a given task) is identified within these neighbors, identification is considered a success (among 944 subjects). (All MRI= T1w+T2w+FA+rfMRI)

| **Mod1 vs Mod2** | **Identification %** | | | |
|---|---|---|---|---|
| | Both | Mod1 | Mod2 | None |
| T1w vs T2w | 9.07 | **15.75** | 12.55 | 62.64 |
| T1w vs FA | 12.09 | 12.73 | **18.59** | 56.59 |
| T1w vs rfMRI | 11.36 | 13.46 | **25.55** | 49.63 |
| FA vs rfMRI | 12.91 | 17.77 | **23.99** | 45.33 |
| FA vs MD | 8.24 | **22.44** | 7.88 | 61.45 |
| T1w vs All MRI | 19.41 | 5.40 | **40.84** | 34.34 |
| T2w vs All MRI | 16.67 | 4.95 | **43.59** | 34.80 |
| FA vs All MRI | 25.64 | 5.04 | **34.62** | 34.71 |
| rfMRI vs All MRI | 34.89 | 2.01 | **25.37** | 37.73 |

due to small sample size, specifically for PHS, it could be difficult to draw major inferences. Still we observe that similar to identification of genetically related subjects the multi-modal combinations have higher MAP values. Also, FS identification results are higher than MHS, while PHS results seems to be better than MHS, but it could be due to limited sample size.

### 4.7.2 Relative informativeness of fingerprints

To compare relative informativeness of one modality vs another for full-siblings, Table 4.13 reports the relative percentage of full-siblings identified by both, a single, or none of the modalities. The total number of identification tasks is $1092$. We consider $k = 10$ nearest neighbors of a subject in terms of fingerprint distance. While comparing single modalities we observe that each modality identifies certain (relative) percentage of full-siblings, with rfMRI performing relatively better than both T1w and FA. As with the MAP results, we see that combining multiple modalities leads to more discriminative fingerprints for full siblings also. For example, combination of all MRI modalities (All MRI= T1w+T2w+FA+rfMRI) identifies $40.84\%$, $43.59\%$, $34.62\%$, and $25.37\%$ of full siblings not identified (relative) from fingerprints based only on T1w, T2w, FA, and rfMRI respectively. Please note, each row represents relative percentages.

### 4.7.3 Mean recall@k results

#### 4.7.3.1 Mean recall@k plots

Figure 4.6 shows mean recall@$k$, for $k = 1, \ldots, 50$. This measure, also known as sensitivity, evaluates the proportion of individuals that are genetically related to a given subject, which are within the $k$ individuals most similar to that subject (in terms of fingerprint distance). To account for chance, we provide the mean recall@k values obtained using a random ranking of subjects. We compare different MRI modalities (top row), structural MRI based modalities (second row), diffusion MRI based DTI measures (third row) and GQI measures (fourth row), and reproducibility (last row).

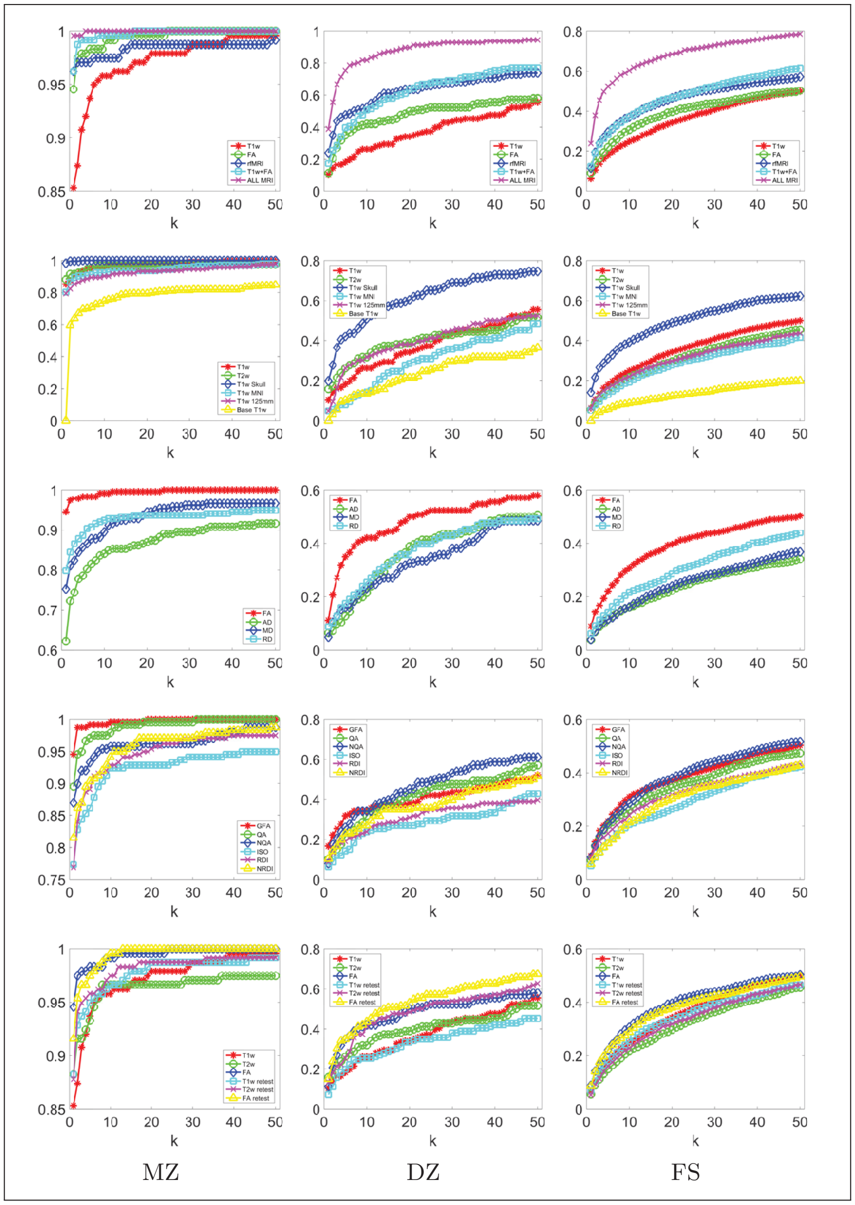


Figure 4.6 Twin/Sibling identification. Mean recall@$k$ plots for MZ,DZ and full-siblings (FS). Plots for comparisons across modalities (top row); sMRI comparisons (second row); dMRI DTI comparisons (third row); dMRI GQI comparisons (fourth row); and reproducibility (last row). Mean recall@$50$ for **random** ranking of subjects are: MZ-$0.063$ ; DZ-$0.056$ ; FA-$0.055$. Note: y-axis represents mean recall@$k$, and it's range is varied for plots for better visualization. (All MRI= T1w+T2w+FA+rfMRI)

The results substantiate the observations drawn using Mean Average Precision and provide more detailed information on changes in sensitivity with increasing $k$. For example, multi-modality combination (All MRI = T1w+T2w+FA+rfMRI) achieves mean recall values of $0.818$

for $k = 10$, and $0.944$ for $k = 50$. Similarly, for full sibling we obtain mean recall values of $0.603$ for $k = 10$, and $0.786$ for $k = 50$.

#### 4.7.3.2 Fingerprint comparisons using mean recall@10

Table 4.14 compares various fingerprints including Fiberprint (Kumar *et al.*, 2017b), functional connectivity based fingerprint (rfMRI netmat), Freesurfer measures based fingerprint, and proposed BoF and multi-modal fingerprints using mean recall@10. The results reflect the improvement provided by multi-modal combination, with T1w+T2w+FA+rfMRI reaching mean recall of $100\%$ for MZ, $81.7\%$ for DZ, and $60.3\%$ for full siblings. The observations are in line with the Mean Average Precision values while providing another standard measure for comparing various fingerprints. Comparing Fiberprint which captures fiber geometry, a very specific information about brain connectivity, we observe that it performs poor w.r.t fingerprints capturing richer information and over multiple scans (e.g. rfMRI captures information over several minutes).

Table 4.14 Comparison across fingerprints. Mean recall@10 values. Note: the data size and processing may vary across the fingerprints, however, the values are on HCP data with more than 850 subjects in each case, thus providing an approximate comparison.

| **Experiment** | **Fingerprint** | **Mean Recall@10** | | |
|---|---|---|---|---|
| | | MZ | DZ | FS |
| Single | Fiberprint | 0.500 | 0.213 | 0.063 |
| | rfMRI netmat | 0.975 | 0.524 | 0.369 |
| | FA BoF | 0.992 | 0.421 | 0.307 |
| | T1w BoF | 0.958 | 0.262 | 0.248 |
| | T2w BoF | 0.966 | 0.317 | 0.216 |
| | T1w Baseline | 0.748 | 0.135 | 0.088 |
| Combination | T1w+T2w+FA | 1.000 | 0.579 | 0.429 |
| | T1w+T2w+DTI | 1.000 | 0.683 | 0.469 |
| | T1w+T2w+FA+rfMRI | **1.000** | **0.818** | **0.603** |
| Fressurfer Meas | Vol | 0.819 | 0.246 | 0.147 |
| | Thck | 0.567 | 0.135 | 0.099 |
| | Area | 0.588 | 0.238 | 0.135 |
| | Vol+Thck+Area | 0.882 | 0.325 | 0.196 |
| Random | Rand | 0.021 | 0.032 | 0.015 |

### 4.7.4 Impact of age

Table 4.15 reports results for the task of genetically related subject identification (MAP values) and significance testing, after dividing twins/siblings into two groups based on median age (MZ/DZ) or median age difference (full siblings).

Table 4.15 Impact of age along with significance testing (right half of the table, with corresponding hypothesis in last columns). Mean avg precision (MAP) table comparing different modalities for the task of genetically-related subject identification, with MZ/DZ divided into 2 groups based on **median age** of $29$, and FS divided into 2 groups based on **median age difference** of $3$ years. In both case, set1 correspond to sibling pairs with age (difference) below or equal to the median, and set2 to those above the median. Right side shows -$\log_{10}$ p-values for unpaired t-test, with p-values $< 0.01$ in bold font.

| **Modality** | **Mean Avg Prec** | | | **Significance** | | | **Hypothesis** |
|---|---|---|---|---|---|---|---|
| | MZ | DZ | FS | MZ | DZ | FS | |
| T1w | 0.886 | 0.160 | 0.128 | 0.63 | 0.31 | 0.07 | set1 vs set2 |
| Age set1 | 0.908 | 0.177 | 0.130 | 0.32 | 0.15 | 0.04 | T1w vs set1 |
| Age set2 | 0.865 | 0.140 | 0.127 | 0.30 | 0.17 | 0.04 | T1w vs set2 |
| T2w | 0.908 | 0.212 | 0.111 | 0.79 | 0.12 | 0.47 | set1 vs set2 |
| Age set1 | 0.932 | 0.203 | 0.117 | 0.41 | 0.06 | 0.22 | T2w vs set1 |
| Age set2 | 0.885 | 0.223 | 0.104 | 0.36 | 0.07 | 0.25 | T2w vs set2 |
| FA | 0.964 | 0.219 | 0.160 | 0.60 | 0.59 | **2.64** | set1 vs set2 |
| Age set1 | 0.976 | 0.248 | 0.186 | 0.31 | 0.26 | 1.01 | FA vs set1 |
| Age set2 | 0.952 | 0.183 | 0.132 | 0.28 | 0.32 | 1.21 | FA vs set2 |

### 4.7.5 Extension to T1w-by-T2w ratio images (myelin content)

We extend our analysis to T1w-by-T2w ratio images (provided with HCP data), which provides an estimate of myelin content. For comparison we use T1w and T2w images with skull. Table 4.16 reports the MAP values comparing different modalities for the task of genetically related subject identification. We observe lower MAP values for T1w-by-T2w, compared to T1w Skull or T2w Skull. Also, the combination of T1w,T2w and T1w-by-T2w (with skull) outperforms single modalities. This analysis, while including myelin content information, validates our multi-modal combination hypothesis in a separate setting (skull included).

Table 4.16 Impact of skull inclusion and extension to T1w/T2w MRI ratio images (myelin content). Mean average precision (MAP) table comparing different modalities for the task of genetically related subject identification. Facial features are not analyzed, and all modalities have 0.7mm scan resolution.

| **Experiment** | **Modality** | **Mean Avg Prec** | | |
|---|---|---|---|---|
| | | MZ | DZ | FS |
| sMRI | T1w | 0.990 | 0.305 | 0.230 |
| | T2w | 0.980 | 0.310 | 0.164 |
| Myelin | T1w-by-T2w | 0.883 | 0.154 | 0.079 |
| Baseline Comparison | T1w MNI | 0.615 | 0.123 | 0.050 |
| Modality Combination | T1w+T2w | 0.998 | 0.445 | 0.285 |
| | T1w+T1w-by-T2w | 1.000 | 0.363 | 0.250 |
| | T1w+T2w+T1w-by-T2w | **1.000** | **0.468** | **0.300** |
| Random | Rand | 0.005 | 0.003 | 0.007 |

### 4.7.6 Scale-space visualization of features correspondences

Figure 4.7 shows features correspondences for a subject and his/her DZ twin, and with another non-twin (full) sibling. The scale information is represented using the circles' radius.

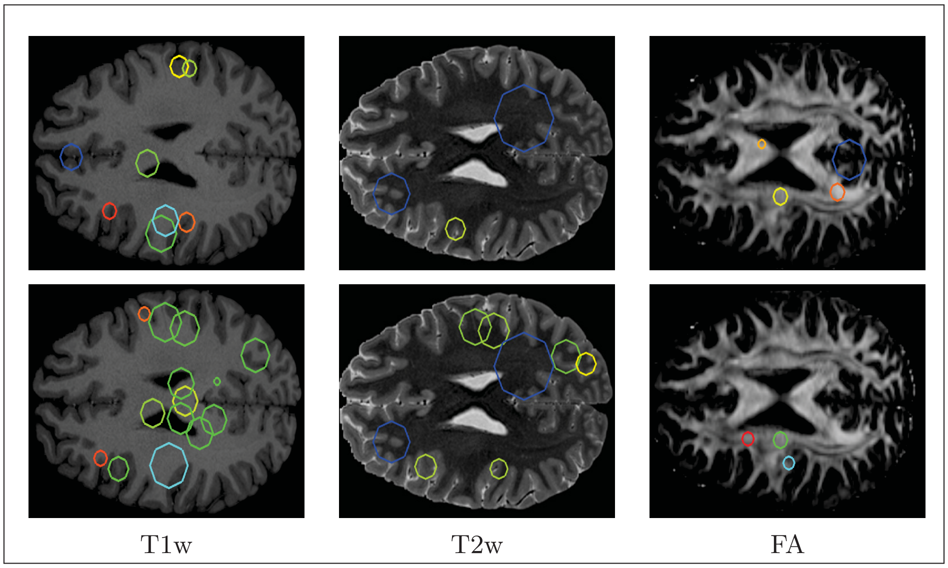


Figure 4.7 Example of feature correspondences for a subject and his/her DZ twin (top row), and the subject's non-twin sibling (FS) (bottom row). Scale space is represented using circle radius (for the visible slice).

## 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. We thank the reviewers whose comments/suggestions helped improve and clarify this manuscript. We also thank researchers processing Human Connectome Project data for providing the pre-processed data and rfMRI netmats.

## Code availability

Matlab scripts of the proposed analysis are available at https://github.com/kkumar-iitkgp-livia/Multi_Modal_Brain_Fingerprinting.git.

# CHAPTER 5

# CONCLUSION

> *"Important thing in science is not so much to obtain new facts as to discover new ways of thinking about them"*
>
> Sir William Bragg

This last chapter provides a summary of the thesis' contributions and recommendations for addressing the limitations of this work.

## 5.1 Summary of contributions

In Chapter 2, we proposed a general framework based on kernel dictionary learning and sparsity priors for white matter fiber analysis. The proposed framework uses an implicit embedding of streamlines and thus can be employed with any fiber similarity measure or computational models for fiber representation. Dictionary bundles are encoded as a combination of training streamlines and kernels are used to model the non-linear relationships between streamlines and bundles. Comparisons against a variety of state-of-the-art streamline clustering approaches using expert-labeled data, as well as subjects from the HCP and MIDAS dataset, demonstrate the usefulness of having a soft assignment. Results also show that our framework is suitable for scenarios where streamlines are not clearly separated, bundles overlap, or when there is important inter-individual variability. Furthermore, experiments indicate that using group sparsity and manifold regularization priors improves clustering by adding robustness to the input number of clusters and incorporating anatomical constraints. The benefits of the proposed approach in cases of inter-individual variability were showcased for the automated segmentation of streamlines from new subjects.

**Impact:** The findings of this chapter has the potential to impact neuroscience studies on diffusion tractography analysis, as well as pattern recognition applications requiring the unsupervised clustering of 3D curves. The proposed streamline segmentation framework extends current approaches by addressing the many challenges related to this task, including streamline length and inter-subject variability. It also offers a flexible way to encode various properties of white matter, such as anatomical priors and microstructure information.

Chapter 3, presented a novel brain fingerprint, called Fiberprint, for the compact characterization of white matter fiber geometry. 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 a limited number of streamlines 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 and that the fingerprint encodes information related to genetics.

**Impact:** The proposed fingerprint is the first to characterize individual differences in white matter fiber geometry. It offers a powerful technique to explore individual differences in terms of white matter connectivity and its relationship to genetics. By including along-tract information on microstructure, it could also be used to define novel biomarkers for detecting and tracking the progression of neurological diseases like Parkinson's.

Finally, Chapter 4 proposes a first multi-modal brain fingerprint, that combines T1/T2-weighted MRI, diffusion MRI, and resting state fMRI for the compact characterization of individual subjects. A comprehensive analysis using multi-modal data from 945 HCP subjects demonstrated the computational efficiency of the proposed fingerprint, as well as its robustness to various factors including image alignment, scan resolution, and subject age. The reproducibility of the results was also confirmed using retest scans from the HCP dataset, showing our fingerprint to be robust to variability in image acquisition. Furthermore, the retrieval analysis presented in the chapter indicated that a compact fingerprint of only 150 features could identify genetically-related subjects better than a baseline using full images as features. Results also showed that each modality provides complementary information which can uniquely identify some sibling pairs. The usefulness of our fingerprint was assessed on the tasks of identifying incorrectly reported zygosity and retest/duplicate scans in a large dataset. Results of this experiment highlighted the effectiveness of our fingerprint, with 100% retrieval performance for all test cases. Finally, analyzing the distribution of features correspondences across the brain revealed various neuroanatomical regions (e.g., cerebellum, lateral ventricles, ventral diencephalon, hippocampus and thalamus proper) with a significantly higher similarity in MZ twins compared to DZ twins or full siblings.

**Impact:** This work constitutes the first study to compare and contrast the contribution of individual modalities towards fingerprint generation. As such, it lays the foundation for future analyses of brain differences in multi-modal MRI. It can thus lead to new insights on the variability of both brain structure and function, which could contribute to the development of personalized treatment strategies.

## 5.2 Limitations and recommendations

Partial volume effects and other tractography-related effects (Maier-Hein *et al.*, 2017), such as fiber tracking or registration errors, could impact the evaluation of our streamline clustering methods and the proposed Fiberprint. Moreover, as highlighted in (Jones *et al.*, 2013), caution must be used 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 (Maier-Hein *et al.*, 2017; Yeh and Tseng, 2011).

One of the main advantages of the proposed kernel-based framework (as well as Fiberprint) is that it alleviates the need for an explicit streamline representation. Previous attempts at utilizing dictionary learning and sparse coding for streamline clustering might have been hindered by this. Employing kernels also provides flexibility and enables the extension to other streamline similarity measures, which can incorporate a richer set of characteristics such as along-tract diffusivity (Kumar *et al.*, 2017d; Charon and Trouvé, 2013; Charlier *et al.*, 2014).

Another key element of our study is the anatomical interpretation of clustering results. The streamlines generated from diffusion tractography provide a macro-scale inference of the underlying fibers (Jones *et al.*, 2013; Maier-Hein *et al.*, 2017). As such, the clustering and Fiberprint for a given distance/similarity measure focuses primarily on the geometric aspect of streamlines. Although we considered end-points proximity in our manifold regularization prior, additional information such as structural or functional parcellation could be incorporated to improve the anatomic plausibility of the final clustering and Fiberprint (O'Donnell *et al.*, 2013; Siless *et al.*, 2018).

Another limitation of the proposed streamline clustering methods stems from their optimization techniques, which are based on the ADMM algorithm. While ADMM facilitates solving a complex problem (e.g., combining several regularization terms) through a process of decomposition, its convergence rate is below that of other optimization approaches. An alternative could be to use techniques based on accelerated gradient descent (Nesterov *et al.*, 2007) like Nesterov's method. Moreover, techniques combining ADMM optimization with deep learning, such as ADMM-Net (Sun *et al.*, 2016), could also be explored as a way to improve computations and reduce the burden of parameter tuning.

The brain fingerprints are motivated by the fact that brain characteristics are largely determined by genetic factors that are often unique to individuals (Thompson *et al.*, 2013). In this work, we used a rank retrieval analysis to evaluate the relation between fingerprint similarity and genetic proximity. However, estimating heritability directly, for instance using the approach described in (Ge *et al.*, 2016), would provide a better quantification of genetic influence on fingerprint features. In (Elliott *et al.*, 2017), the data of over 8,000 subjects from the UK Biobank (Sudlow *et al.*, 2015) was considered to determine the heritability of multi-modal brain imaging phenotypes. Similarly, Colclough et al. (Colclough *et al.*, 2017) the heritability of multi-modal functional connectivity profiles using 800 HCP subjects. An extensive analysis is required to assess the heritability of the proposed fingerprint and relate our findings to these recent studies.

An interesting extension of this work would be to assess whether our fingerprints can be used as a biomarker to identify subjects with cognitive or neurological disorders. Publicly available data, for instance from the ADNI dataset (Mueller *et al.*, 2005; Toews *et al.*, 2010) or Parkinson's Progression Markers Initiative (PPMI) dataset (Marek *et al.*, 2011), could be used for this analysis. Although aging effects were considered in our analysis, 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.

# APPENDIX I

## PUBLICATIONS DURING PH.D. STUDY

- **Kumar Kuldeep**, Desrosiers Christian, Siddiqi Kaleem, Colliot Olivier, Toews Matthew. (2017). Fiberprint: A subject fingerprint based on sparse code pooling for white matter fiber analysis. Neuroimage. 158: 242-259.

- **Kumar Kuldeep**, Chauvin Laurent, Toews Matthew, Colliot Olivier, Desrosiers Christian. (2018). Multi-modal brain fingerprinting: a manifold approximation based framework. (BioRxiv, and under review Neuroimage)

- **Kumar Kuldeep**, Siddiqi Kaleem, Desrosiers Christian. (2018). White matter fiber analysis using kernel dictionary learning and sparsity priors. (ArXiv, Submitted to Pattern Recognition)

- **Kumar Kuldeep**, Gori Pietro, Charlier Benjamin, Durrleman Stanley, Colliot Olivier, Desrosiers Christian. (2017). White matter fiber segmentation using functional varifolds. International Workshop on Mathematical Foundations of Computational Anatomy, MICCAI 2017. Graphs in Biomedical Image Analysis, Computational Anatomy and Imaging Genetics. Springer, Cham. 92-100.

- **Kumar Kuldeep**, Chauvin Laurent, Toews Matthew, Colliot Olivier, Desrosiers Christian. (2017). Multi-modal analysis of genetically-related subjects using SIFT descriptors in brain MRI. Workshop on Computational Diffusion MRI, MICCAI 2017. Book Series: Mathematics and Visualization. (pp. 219-228). Springer.

- Chauvin Laurent, **Kumar Kuldeep**, Desrosiers Christian, De Guise Jacques, Toews Matthew. (2017). Diffusion Orientation Histograms (DOH) for Diffusion Weighted Image Analysis. Workshop on Computational Diffusion MRI, MICCAI 2017. Book Series: Mathematics and Visualization. (pp. 91-99). Springer.

- **Kumar Kuldeep**, Desrosiers Christian, Toews Matthew, Chaddad Ahmad. (2016). Spatially constrained sparse regression for the data-driven discovery of neuroimaging biomarkers. 23rd IEEE International Conference on Pattern Recognition (ICPR), 2016. (pp. 2162-2167).

- Zhang Mingli, **Kumar Kuldeep**, Desrosiers Christian. (2016). A weighted total variation approach for the atlas-based reconstruction of brain MR data. 2016 IEEE International Conference on Image Processing (ICIP), pp. 4329-4333. IEEE, 2016.

- **Kumar Kuldeep**, Desrosiers Christian. (2016). A sparse coding approach for the efficient representation and segmentation of white matter fibers. IEEE 13th International Symposium on Biomedical Imaging (ISBI), 2016. (pp. 915-919).

- **Kumar Kuldeep**, Desrosiers Christian, Siddiqi Kaleem. (2015). Brain Fiber Clustering Using Non-negative Kernelized Matching Pursuit. International Workshop on Machine Learning in Medical Imaging, MICCAI 2015. (pp. 144-152). Springer, Cham.

- **Kumar Kuldeep**, Desrosiers Christian. (2014). Group Sparse Kernelized Dictionary Learning for the Clustering of White Matter Fibers. Workshop on Sparsity Techniques in Medical Imaging, MICCAI 2014.

### Select Abstracts/Posters

- **Kumar Kuldeep**, Chauvin Laurent, Toews Matthew, Colliot Olivier, Desrosiers Christian. (2018). Analysis of genetically related subjects using multi-modal brain fingerprints. 1st Symposium on Applications of Artificial Intelligence in Medicine (IAM 2018). (**Best poster award**)

- **Kumar Kuldeep**, Desrosiers Christian, Siddiqi Kaleem, Colliot Olivier, Toews Matthew. (2017). Fiberprint: Identifying subjects and twins using fiber geometry based brain fingerprint. Medical Imaging meets NIPS, NIPS 2017.

- **Kumar Kuldeep**, Chauvin Laurent, Toews Matthew, Colliot Olivier, Desrosiers Christian. (2017). Multi-modal brain fingerprinting: a bag of features and manifold approximation based twin analysis. Montreal Artificial Intelligence & Neuroscience (MAIN 2017).

### Select Publication Coverage

- *L'empreinte cérébrale : une image des connexions du cerveau*
  SUBSTANCE ÉTS 2017/11/14
- *Fiberprint: Human Brain Wiring Shows Unique Fingerprint*
  SUBSTANCE ÉTS 2017/11/14
- *Fiberprint: A subject fingerprint based on sparse code pooling for white matter fiber analysis*
  NewsRx LLC 2017/10/12

# APPENDIX II

## CODE AVAILABILITY

- Matlab scripts of the Multi-modal brain fingerprinting analysis are available at https://github.com/kkumar-iitkgp-livia/Multi_Modal_Brain_Fingerprinting.git.

# APPENDIX III

## COPYRIGHT PERMISSIONS FOR FIGURES USED IN LITERATURE REVIEW

- **Copyright permissions/License for Figure 1.1**:

  “ *This file is licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license. Attribution: Quasar Jarosz at English Wikipedia Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled GNU Free Documentation License.*”

- **Copyright permissions/License for Figure 1.2**:

  “*This file is licensed under the Creative Commons Attribution 3.0 Unported license.*”

- **Copyright permissions/License for Figure 1.3**:

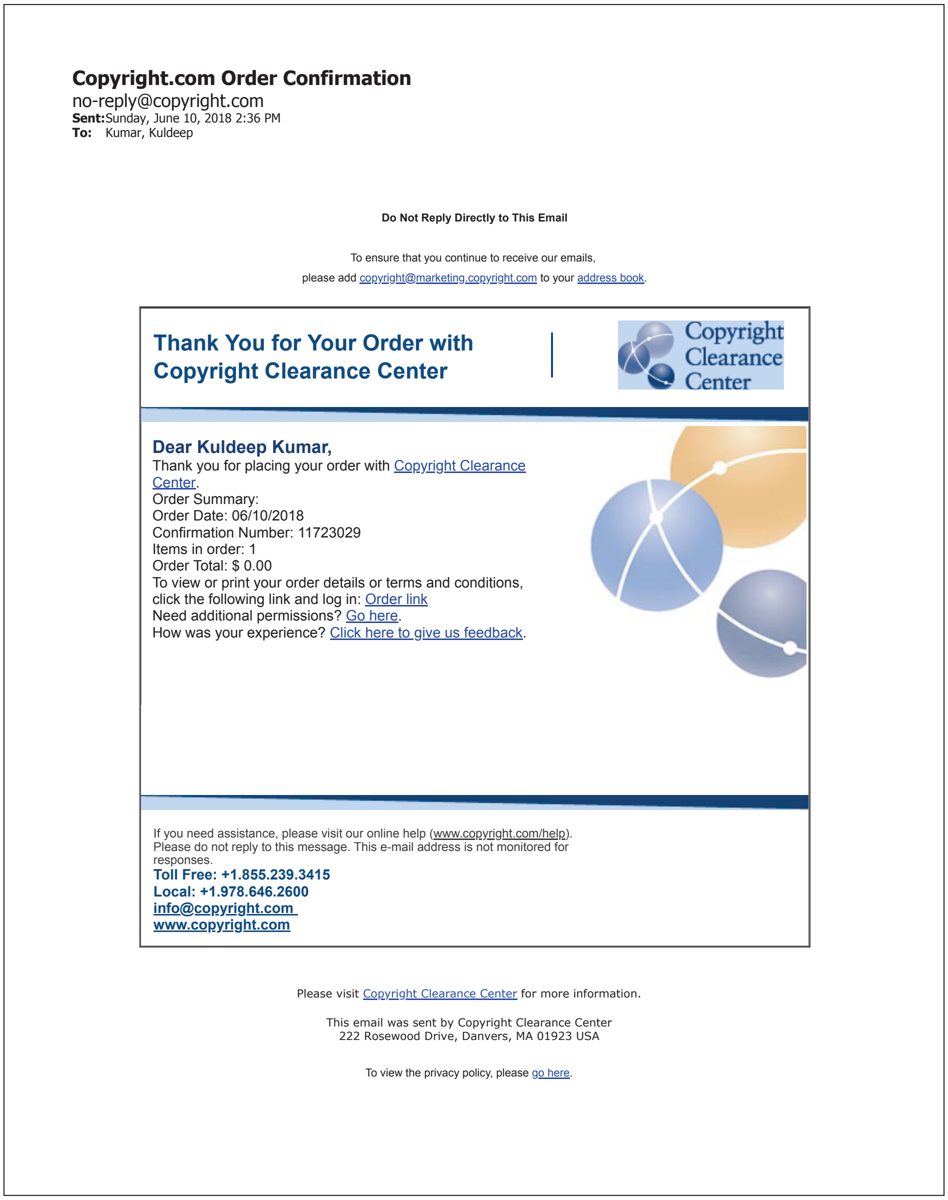
  “© *2015 The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, provided the original author and source are credited.*”

- **Copyright permissions/License for Figure 1.4**: see Fig. A.III.1

- **Copyright permissions/License for Figure 1.5 (right column)**: see Fig. A.III.2

- **Copyright permissions/License for Figure 1.6 (top row)**:

  “*Creative Commons Attribution-NonCommercial-No Derivatives License (CC BY NC ND) This article is published under the terms of the Creative Commons Attribution-NonCommercial-No Derivatives License (CC BY NC ND). For non-commercial purposes you may copy and distribute the article, use portions or extracts from the article in other works, and text or data mine the article, provided you do not alter or modify the article without permission from Elsevier. You may also create adaptations of the article for your*



**Copyright.com Order Confirmation**
no-reply@copyright.com
**Sent:**Sunday, June 10, 2018 2:36 PM
**To:** Kumar, Kuldeep

**Do Not Reply Directly to This Email**

To ensure that you continue to receive our emails,
please add copyright@marketing.copyright.com to your address book.

**Thank You for Your Order with Copyright Clearance Center**

**Dear Kuldeep Kumar,**
Thank you for placing your order with Copyright Clearance Center.
Order Summary:
Order Date: 06/10/2018
Confirmation Number: 11723029
Items in order: 1
Order Total: $ 0.00
To view or print your order details or terms and conditions, click the following link and log in: Order link
Need additional permissions? Go here.
How was your experience? Click here to give us feedback.

If you need assistance, please visit our online help (www.copyright.com/help).
Please do not reply to this message. This e-mail address is not monitored for responses.
**Toll Free: +1.855.239.3415**
**Local: +1.978.646.2600**
**info@copyright.com**
**www.copyright.com**

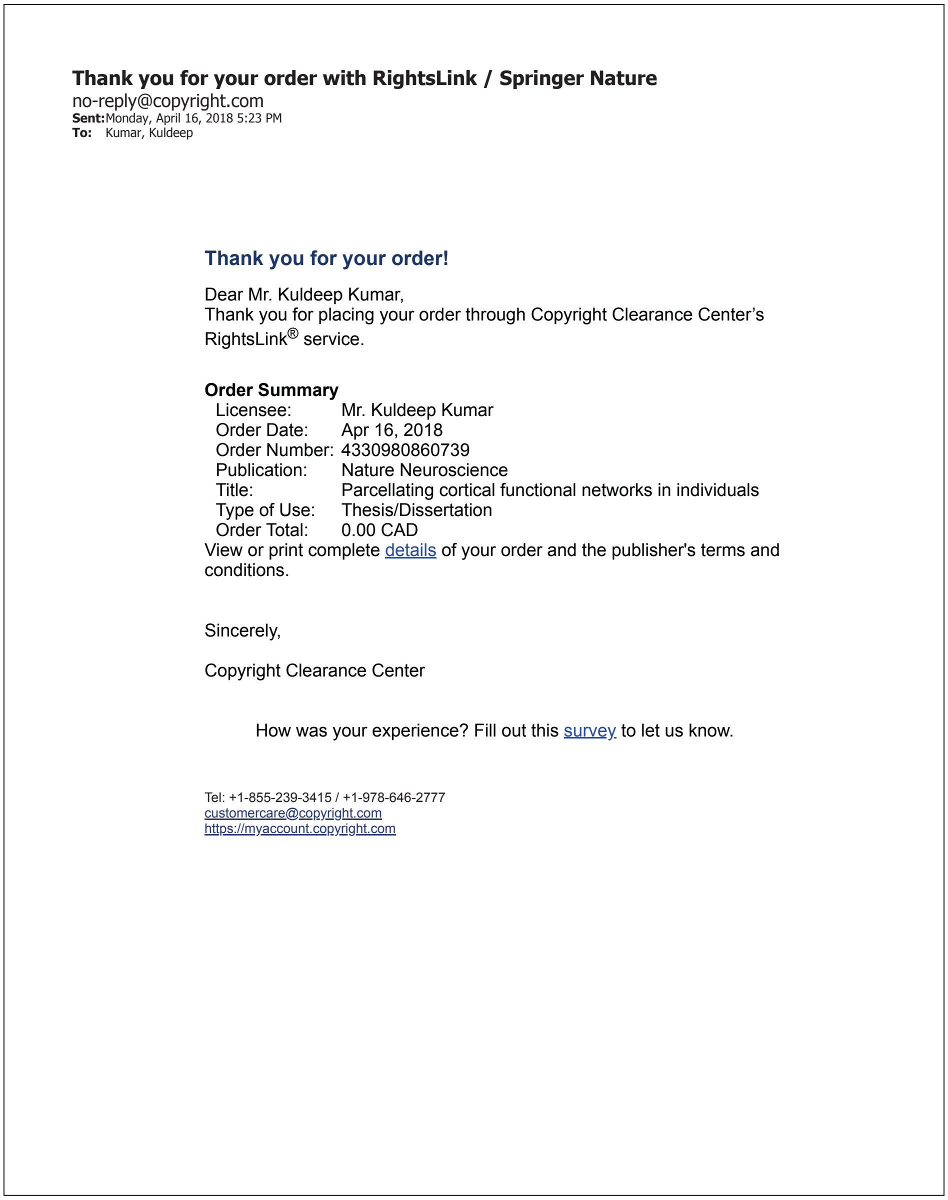
Please visit Copyright Clearance Center for more information.

This email was sent by Copyright Clearance Center
222 Rosewood Drive, Danvers, MA 01923 USA

To view the privacy policy, please go here.

Figure A.III.1 Copyright permissions for Figure 1.4

*own personal use only, but not distribute these to others. You must give appropriate credit to the original work, together with a link to the formal publication through the relevant DOI, and a link to the Creative Commons user license above. If changes are permitted, you must indicate if any changes are made but not in any way that suggests the licensor endorses you or your use of the work. Permission is not required for this non-commercial use. For commercial use please continue to request permission via Rightslink.”*



**Thank you for your order with RightsLink / Springer Nature**
no-reply@copyright.com
**Sent:**Monday, April 16, 2018 5:23 PM
**To:** Kumar, Kuldeep

**Thank you for your order!**

Dear Mr. Kuldeep Kumar,
Thank you for placing your order through Copyright Clearance Center's RightsLink® service.

**Order Summary**

| | |
|---|---|
| Licensee: | Mr. Kuldeep Kumar |
| Order Date: | Apr 16, 2018 |
| Order Number: | 4330980860739 |
| Publication: | Nature Neuroscience |
| Title: | Parcellating cortical functional networks in individuals |
| Type of Use: | Thesis/Dissertation |
| Order Total: | 0.00 CAD |

View or print complete details of your order and the publisher's terms and conditions.

Sincerely,

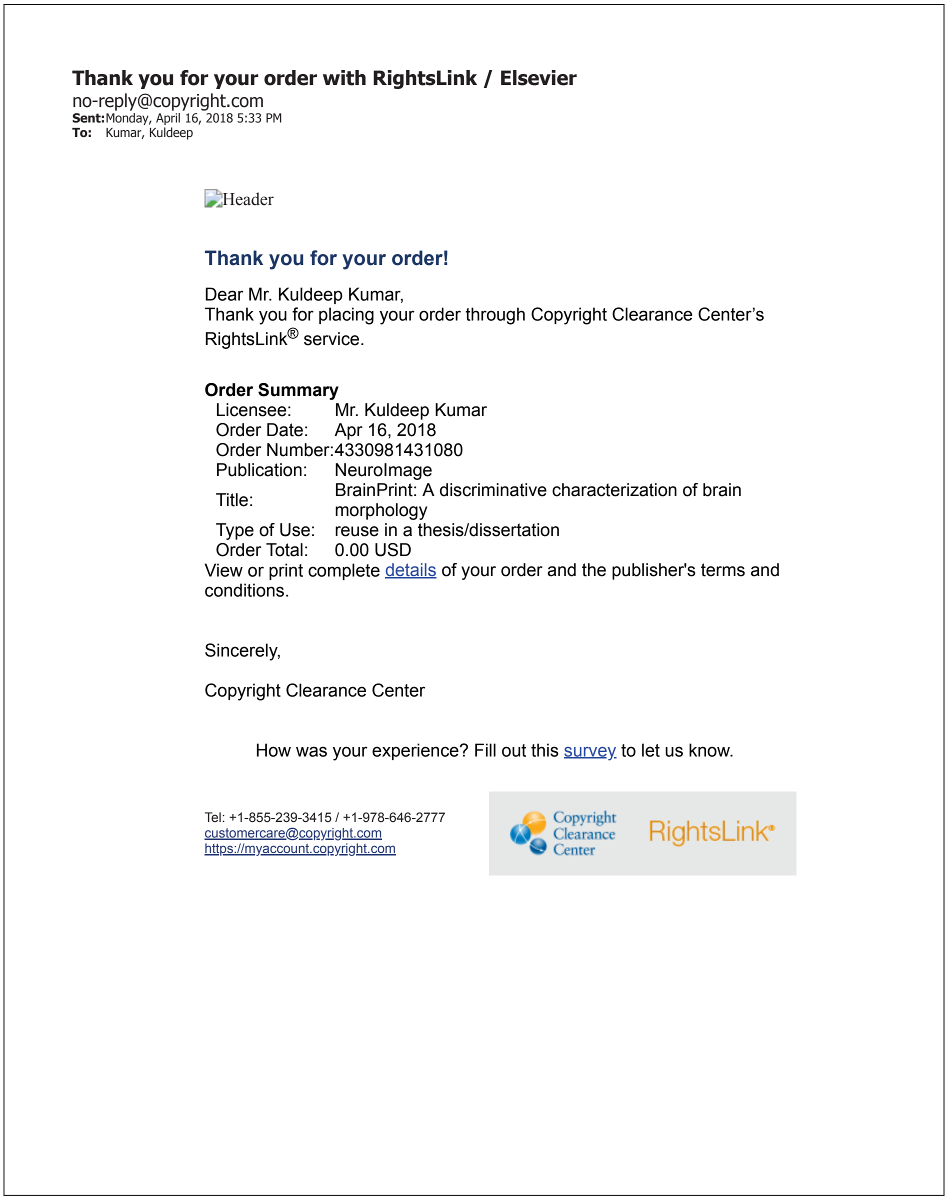
Copyright Clearance Center

How was your experience? Fill out this survey to let us know.

Tel: +1-855-239-3415 / +1-978-646-2777
customercare@copyright.com
https://myaccount.copyright.com

Figure A.III.2 Copyright permissions for Figure 1.5 (right column)

- **Copyright permissions/License for Figure 1.6 (middle row)**:
  "*This is an open access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the Creative Commons CC0 public domain dedication.*"
- **Copyright permissions/License for Figure 1.6 (bottom row)**: see Fig. A.III.3



**Thank you for your order with RightsLink / Elsevier**
no-reply@copyright.com
**Sent:** Monday, April 16, 2018 5:33 PM
**To:** Kumar, Kuldeep

Header

**Thank you for your order!**

Dear Mr. Kuldeep Kumar,
Thank you for placing your order through Copyright Clearance Center's RightsLink® service.

**Order Summary**

| | |
|---|---|
| Licensee: | Mr. Kuldeep Kumar |
| Order Date: | Apr 16, 2018 |
| Order Number: | 4330981431080 |
| Publication: | NeuroImage |
| Title: | BrainPrint: A discriminative characterization of brain morphology |
| Type of Use: | reuse in a thesis/dissertation |
| Order Total: | 0.00 USD |

View or print complete details of your order and the publisher's terms and conditions.

Sincerely,

Copyright Clearance Center

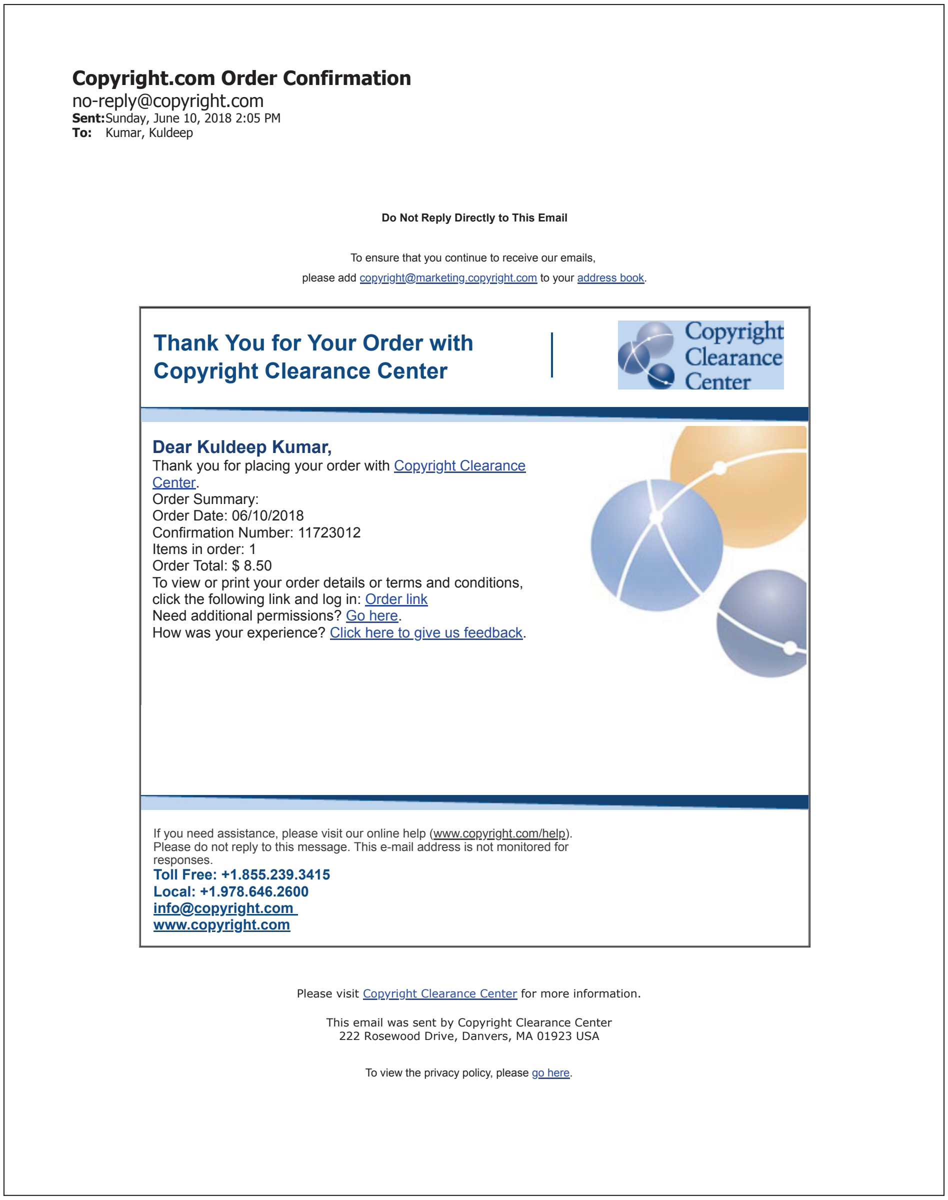
How was your experience? Fill out this survey to let us know.

Tel: +1-855-239-3415 / +1-978-646-2777
customercare@copyright.com
https://myaccount.copyright.com

Copyright Clearance Center RightsLink®

Figure A.III.3 Copyright permissions for Figure 1.6 (bottom row)

- **Copyright permissions/License for Figure 1.7 (top row)**: see Fig. A.III.4
- **Copyright permissions/License for Figure 1.7 (middle row)**:
  "*Copyright: © 2012 Garyfallidis, Brett, Correia, Williams and Nimmo-Smith. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in other forums, provided the*



**Copyright.com Order Confirmation**
no-reply@copyright.com
**Sent:**Sunday, June 10, 2018 2:05 PM
**To:** Kumar, Kuldeep

**Do Not Reply Directly to This Email**

To ensure that you continue to receive our emails,
please add copyright@marketing.copyright.com to your address book.

**Thank You for Your Order with Copyright Clearance Center**

**Dear Kuldeep Kumar,**
Thank you for placing your order with Copyright Clearance Center.
Order Summary:
Order Date: 06/10/2018
Confirmation Number: 11723012
Items in order: 1
Order Total: $ 8.50
To view or print your order details or terms and conditions, click the following link and log in: Order link
Need additional permissions? Go here.
How was your experience? Click here to give us feedback.

If you need assistance, please visit our online help (www.copyright.com/help).
Please do not reply to this message. This e-mail address is not monitored for responses.
**Toll Free: +1.855.239.3415**
**Local: +1.978.646.2600**
**info@copyright.com**
**www.copyright.com**

Please visit Copyright Clearance Center for more information.

This email was sent by Copyright Clearance Center
222 Rosewood Drive, Danvers, MA 01923 USA

To view the privacy policy, please go here.

Figure A.III.4 Copyright permissions for Figure 1.7 (top row)

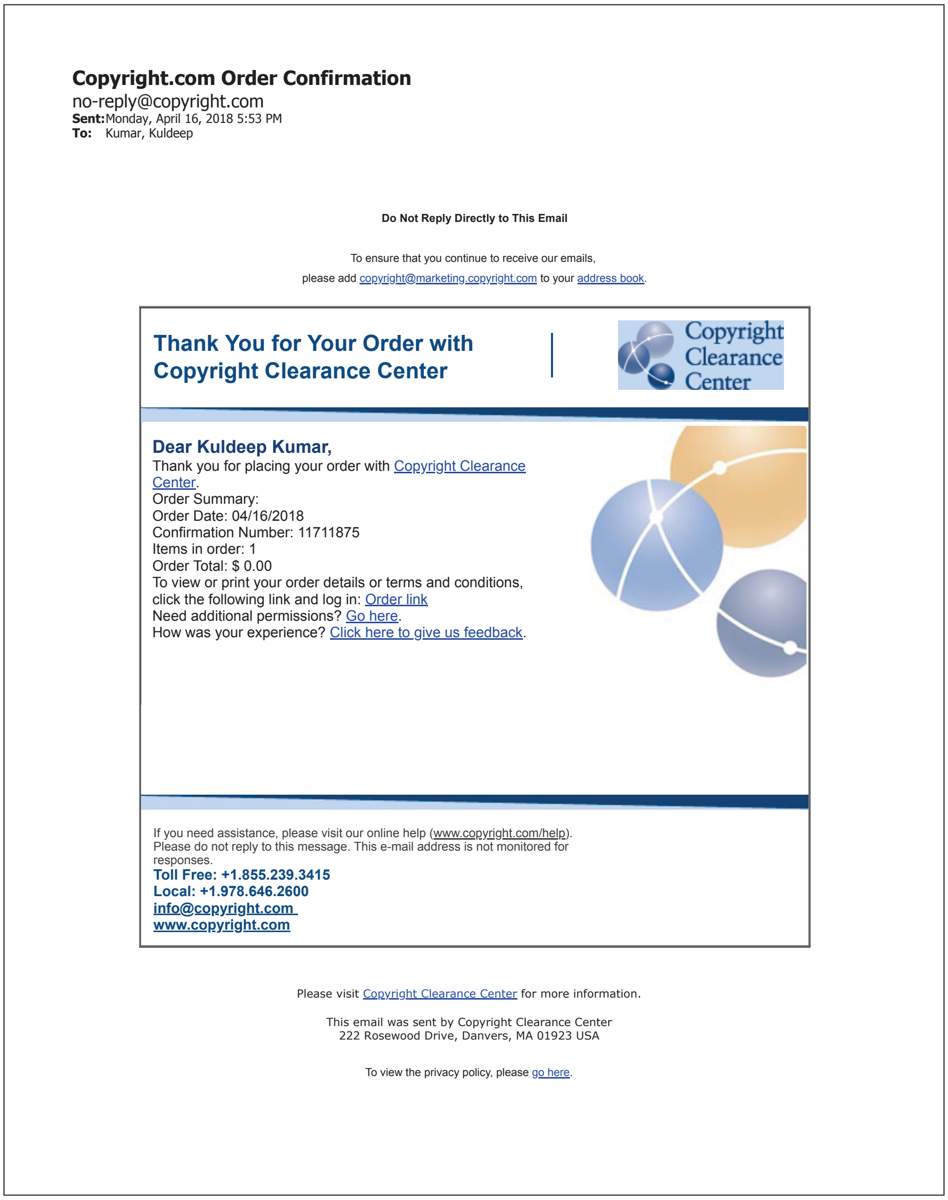
*original authors and source are credited and subject to any copyright notices concerning any third-party graphics etc.”*

- **Copyright permissions/License for Figure 1.7 (bottom row)**:

  “*Copyright: © 2013 Ros et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, dis-*

*tribution, and reproduction in any medium, provided the original author and source are credited.”*

- **Copyright permissions/License for Figure 1.9**: see Fig. A.III.5



**Copyright.com Order Confirmation**
no-reply@copyright.com
**Sent:** Monday, April 16, 2018 5:53 PM
**To:** Kumar, Kuldeep

**Do Not Reply Directly to This Email**

To ensure that you continue to receive our emails,
please add copyright@marketing.copyright.com to your address book.

**Thank You for Your Order with Copyright Clearance Center**

**Dear Kuldeep Kumar,**
Thank you for placing your order with Copyright Clearance Center.
Order Summary:
Order Date: 04/16/2018
Confirmation Number: 11711875
Items in order: 1
Order Total: $ 0.00
To view or print your order details or terms and conditions, click the following link and log in: Order link
Need additional permissions? Go here.
How was your experience? Click here to give us feedback.

If you need assistance, please visit our online help (www.copyright.com/help). Please do not reply to this message. This e-mail address is not monitored for responses.
**Toll Free: +1.855.239.3415**
**Local: +1.978.646.2600**
**info@copyright.com**
**www.copyright.com**

Please visit Copyright Clearance Center for more information.

This email was sent by Copyright Clearance Center
222 Rosewood Drive, Danvers, MA 01923 USA

To view the privacy policy, please go here.

Figure A.III.5  Copyright permissions for Figure 1.9

# BIBLIOGRAPHY

Aharon, Michal, Michael Elad, and Alfred Bruckstein. 2006. “k-svd: An algorithm for designing overcomplete dictionaries for sparse representation”. *Signal Processing, IEEE Transactions on*, vol. 54, n° 11, p. 4311–4322.

Alexander, Andrew L, Jee Eun Lee, Mariana Lazar, and Aaron S Field. 2007. “Diffusion tensor imaging of the brain”. *Neurotherapeutics*, vol. 4, n° 3, p. 316–329.

Alexander, Daniel C, Andrada Ianus, and Aurobrata Ghosh. 2018. Advanced diffusion models. *Quantitative MRI of the Brain*, p. 157–178. CRC Press.

Alexandroni, Guy, Yana Podolsky, Hayit Greenspan, Tal Remez, Or Litany, Alexander Bronstein, and Raja Giryes. 2017. “White Matter Fiber Representation Using Continuous Dictionary Learning”. In *International Conference on Medical Image Computing and Computer-Assisted Intervention.* p. 566–574. Springer.

Aljabar, Paul, Daniel Rueckert, and William R Crum. 2008. “Automated morphological analysis of magnetic resonance brain imaging using spectral analysis”. *NeuroImage*, vol. 43, n° 2, p. 225–235.

Aljabar, Paul, Robin Wolz, Latha Srinivasan, S Counsell, James P Boardman, Maria Murgasova, Valentina Doria, Mary A Rutherford, A David Edwards, Joseph V Hajnal, et al. 2010. “Combining morphological information in a manifold learning framework: application to neonatal MRI”. In *International Conference on Medical Image Computing and Computer-Assisted Intervention.* p. 1–8. Springer.

Aljabar, Paul, Robin Wolz, and Daniel Rueckert. 2012. “Manifold learning for medical image registration, segmentation, and classification”. *Machine Learning in Computer-Aided Diagnosis: Medical Imaging Intelligence and Analysis: Medical Imaging Intelligence and Analysis*, p. 351.

Amunts, Katrin, Aleksandar Malikovic, Hartmut Mohlberg, Thorsten Schormann, and Karl Zilles. 2000. “Brodmann’s areas 17 and 18 brought into stereotaxic space—where and how variable?”. *Neuroimage*, vol. 11, n° 1, p. 66–84.

Amunts, Katrin, Christoph Ebell, Jeff Muller, Martin Telefont, Alois Knoll, and Thomas Lippert. 2016. “The Human Brain Project: creating a European research infrastructure to decode the human brain”. *Neuron*, vol. 92, n° 3, p. 574–581.

Aranda, Ramon, Alonso Ramirez-Manzanares, and Mariano Rivera. 2015. “Sparse and Adaptive Diffusion Dictionary (SADD) for recovering intra-voxel white matter structure”. *Medical image analysis*, vol. 26, n° 1, p. 243–255.

Arbabshirani, Mohammad R, Sergey Plis, Jing Sui, and Vince D Calhoun. 2017. “Single subject prediction of brain disorders in neuroimaging: promises and pitfalls”. *NeuroImage*, vol. 145, p. 137–165.

Armstrong, Blair C, Maria V Ruiz-Blondet, Negin Khalifian, Kenneth J Kurtz, Zhanpeng Jin, and Sarah Laszlo. 2015. “Brainprint: Assessing the uniqueness, collectability, and permanence of a novel method for ERP biometrics”. *Neurocomputing*, vol. 166, p. 59–67.

Ashburner, J. 2000. “Computational Neuroanatomy”. PhD thesis, University College London. </spm/doc/theses/john/>.

Assemlal, Haz-Edine, David Tschumperlé, Luc Brun, and Kaleem Siddiqi. 2011. “Recent advances in diffusion MRI modeling: Angular and radial reconstruction”. *Medical image analysis*, vol. 15, n° 4, p. 369–396.

Auría, Anna, Alessandro Daducci, J-P Thiran, and Yves Wiaux. 2015. “Structured sparsity for spatially coherent fibre orientation estimation in diffusion MRI”. *NeuroImage*, vol. 115, p. 245–255.

Bandettini, Peter A. 2012. “Twenty years of functional MRI: the science and the stories”. *Neuroimage*, vol. 62, n° 2, p. 575–588.

Barch, Deanna M, Gregory C Burgess, Michael P Harms, Steven E Petersen, Bradley L Schlaggar, Maurizio Corbetta, Matthew F Glasser, Sandra Curtiss, Sachin Dixit, Cindy Feldt, et al. 2013. “Function in the human connectome: task-fMRI and individual differences in behavior”. *Neuroimage*, vol. 80, p. 169–189.

Bartels, Richard H. and GW Stewart. 1972. “Solution of the matrix equation AX+ XB= C [F4]”. *Communications of the ACM*, vol. 15, n° 9, p. 820–826.

Basser, Peter J, James Mattiello, and Denis LeBihan. 1994. “MR diffusion tensor spectroscopy and imaging.”. *Biophysical journal*, vol. 66, n° 1, p. 259.

Basser, Peter J, Sinisa Pajevic, Carlo Pierpaoli, Jeffrey Duda, and Akram Aldroubi. 2000. “In vivo fiber tractography using DT-MRI data”. *Magnetic resonance in medicine*, vol. 44, n° 4, p. 625–632.

Baumann, Michael, Mechthild Krause, Jens Overgaard, Jurgen Debus, Saren M Bentzen, Juliane Daartz, Christian Richter, Daniel Zips, and Thomas Bortfeld. 2016. “Radiation oncology in the era of precision medicine”. *Nature Reviews Cancer*, vol. 16, n° 4, p. 234–250.

Bay, Herbert, Tinne Tuytelaars, and Luc Van Gool. 2006. “Surf: Speeded up robust features”. *Computer vision–ECCV 2006*, p. 404–417.

Bay, Herbert, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. 2008. “Speeded-up robust features (SURF)”. *Computer vision and image understanding*, vol. 110, n° 3, p. 346–359.

Beckmann, Christian F. 2012. “Modelling with independent components”. *Neuroimage*, vol. 62, n° 2, p. 891–901.

Behrens, TEJ, H Johansen Berg, Saad Jbabdi, MFS Rushworth, and MW Woolrich. 2007. “Probabilistic diffusion tractography with multiple fibre orientations: What can we gain?”. *Neuroimage*, vol. 34, n° 1, p. 144–155.

Belkin, Mikhail and Partha Niyogi. 2003. “Laplacian eigenmaps for dimensionality reduction and data representation”. *Neural computation*, vol. 15, n° 6, p. 1373–1396.

Bengio, Yoshua, Jean-françcois Paiement, Pascal Vincent, Olivier Delalleau, Nicolas L Roux, and Marie Ouimet. 2004. “Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering”. In *Advances in neural information processing systems*. p. 177–184.

Bengio, Yoshua, Aaron Courville, and Pascal Vincent. 2013. “Representation learning: A review and new perspectives”. *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, n° 8, p. 1798–1828.

Berman, Jeffrey I, SungWon Chung, Pratik Mukherjee, Christopher P Hess, Eric T Han, and Roland G Henry. 2008. “Probabilistic streamline< i> q</i>-ball tractography using the residual bootstrap”. *Neuroimage*, vol. 39, n° 1, p. 215–222.

Bijsterbosch, Janine Diane, Mark W Woolrich, Matthew F Glasser, Emma C Robinson, Christian F Beckmann, David C Van Essen, Samuel J Harrison, and Stephen M Smith. 2018. “The relationship between spatial configuration and functional connectivity of brain regions”. *eLife*, vol. 7, p. e32992.

Bilgic, Berkin, Itthi Chatnuntawech, Kawin Setsompop, Stephen F Cauley, Anastasia Yendiki, Lawrence L Wald, and Elfar Adalsteinsson. 2013. “Fast dictionary-based reconstruction for diffusion spectrum imaging”. *IEEE transactions on medical imaging*, vol. 32, n° 11, p. 2022–2033.

Biswal, Bharat B. 2012. “Resting state fMRI: a personal history”. *Neuroimage*, vol. 62, n° 2, p. 938–944.

Bloch, Felix. 1946. “Nuclear induction”. *Physical review*, vol. 70, n° 7-8, p. 460.

Booth, Brian G and Ghassan Hamarneh. 2011. “Exact integration of diffusion orientation distribution functions for graph-based diffusion MRI analysis”. In *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*. p. 935–938. IEEE.

Booth, Brian G and Ghassan Hamarneh. 2012. “Multi-region competitive tractography via graph-based random walks”. In *Mathematical Methods in Biomedical Image Analysis (MMBIA), 2012 IEEE Workshop on*. p. 73–78. IEEE.

Boyd, Stephen, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. 2011. “Distributed optimization and statistical learning via the alternating direction method of multipliers”. *Foundations and Trends® in Machine learning*, vol. 3, n° 1, p. 1–122.

Brosch, Tom, Roger Tam, Alzheimer's Disease Neuroimaging Initiative, et al. 2013. “Manifold learning of brain MRIs by deep learning”. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. p. 633–640. Springer.

Brown, Timothy T. 2017. “Individual differences in human brain development”. *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 8, n° 1-2.

Brun, Anders, Hans Knutsson, Hae-Jeong Park, Martha E Shenton, and Carl-Fredrik Westin. 2004. Clustering fiber traces using normalized cuts. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2004*, p. 368–375. Springer.

Bu, Lu-Lu, Ke Yang, Wei-Xi Xiong, Feng-Tao Liu, Boyd Anderson, Ye Wang, and Jian Wang. 2016. “Toward precision medicine in Parkinson's disease”. *Annals of translational medicine*, vol. 4, n° 2.

Buckley, Chris and Ellen M Voorhees. 2000. “Evaluating evaluation measure stability”. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. p. 33–40. ACM.

Bullitt, E, D Zeng, et al. 2005. “Vessel tortuosity and brain tumor malignancy: a blinded study1”. *Academic radiology*, vol. 12, n° 10, p. 1232–1240.

Bürgel, Uli, Katrin Amunts, Lars Hoemke, Hartmut Mohlberg, Joachim M Gilsbach, and Karl Zilles. 2006. “White matter fiber tracts of the human brain: three-dimensional mapping at microscopic resolution, topography and intersubject variability”. *Neuroimage*, vol. 29, n° 4, p. 1092–1105.

Bzdok, Danilo and Andreas Meyer-Lindenberg. 2017. “Machine learning for precision psychiatry”. *arXiv preprint arXiv:1705.10553*.

Bzdok, Danilo and BT Thomas Yeo. 2017. “Inference in the age of big data: Future perspectives on neuroscience”. *NeuroImage*, vol. 155, p. 549 - 564.

Calhoun, Vince D and Jing Sui. 2016. “Multimodal fusion of brain imaging data: A key to finding the missing link (s) in complex mental illness”. *Biological psychiatry: cognitive neuroscience and neuroimaging*, vol. 1, n° 3, p. 230–244.

Calhoun, Vince D, Stephan M Lawrie, Janaina Mourao-Miranda, and Klaas E Stephan. 2017. “Prediction of individual differences from neuroimaging data”. *NeuroImage*, vol. 145, n° Pt B, p. 135.

Çetingül, H Ertan, Margaret J Wright, Paul M Thompson, and René Vidal. 2014. “Segmentation of high angular resolution diffusion MRI using sparse Riemannian manifold clustering”. *IEEE transactions on medical imaging*, vol. 33, n° 2, p. 301–317.

Chamberland, Maxime, Kevin Whittingstall, David Fortin, David Mathieu, and Maxime Descoteaux. 2014. “Real-time multi-peak tractography for instantaneous connectivity display”. *Frontiers in neuroinformatics*, vol. 8, p. 59.

Chamberland, Maxime, Gabriel Girard, Michaël Bernier, David Fortin, Maxime Descoteaux, and Kevin Whittingstall. 2017. “ On the Origin of Individual Functional Connectivity Variability: The Role of White Matter Architecture ”. *Brain connectivity*, vol. 7, n° 8, p. 491–503.

Chan, Micaela Y, Fahd H Alhazmi, Denise C Park, Neil K Savalia, and Gagan S Wig. 2017. “ Resting-state network topology differentiates task signals across the adult life span ”. *Journal of Neuroscience*, vol. 37, n° 10, p. 2734–2745.

Chance, Steven A and Timothy J Crow. 2007. “ Distinctively human: cerebral lateralisation and language in Homo sapiens ”. *J Anthropol Sci*, vol. 85, p. 83–100.

Charlier, Benjamin, Nicolas Charon, and Alain Trouvé. 2014. “ The fshape framework for the variability analysis of functional shapes ”. *Foundations of Computational Mathematics*, p. 1–71.

Charon, Nicolas and Alain Trouvé. 2013. “ The varifold representation of nonoriented shapes for diffeomorphic registration ”. *SIAM Journal on Imaging Sciences*, vol. 6, n° 4, p. 2547–2580.

Chen, Yihua, Maya R Gupta, and Benjamin Recht. 2009. “ Learning kernels from indefinite similarities ”. In *ICML 2009*. p. 145–152. ACM.

Chiang, Ming-Chang, Marina Barysheva, David W Shattuck, Agatha D Lee, Sarah K Madsen, Christina Avedissian, Andrea D Klunder, Arthur W Toga, Katie L McMahon, Greig I De Zubicaray, et al. 2009. “ Genetics of brain fiber architecture and intellectual performance ”. *Journal of Neuroscience*, vol. 29, n° 7, p. 2212–2224.

Chiang, Ming-Chang, Katie L McMahon, Greig I de Zubicaray, Nicholas G Martin, Ian Hickie, Arthur W Toga, Margaret J Wright, and Paul M Thompson. 2011. “ Genetics of white matter development: a DTI study of 705 twins and their siblings aged 12 to 29 ”. *Neuroimage*, vol. 54, n° 3, p. 2308–2317.

Chou, Kun-Hsien, Yawei Cheng, I-Yun Chen, Ching-Po Lin, and Woei-Chyn Chu. 2011. “ Sex-linked white matter microstructure of the social and analytic brain ”. *Neuroimage*, vol. 54, n° 1, p. 725–733.

Chung, H-W, M-C Chou, and C-Y Chen. 2011. “ Principles and limitations of computational algorithms in clinical diffusion tensor MR tractography ”. *American Journal of Neuroradiology*, vol. 32, n° 1, p. 3–13.

Cichocki, Andrzej. 2013. “ Tensor Decompositions: A New Concept in Brain Data Analysis? ”. *arXiv preprint arXiv:1305.0395*.

Colby, John B, Lindsay Soderberg, Catherine Lebel, Ivo D Dinov, Paul M Thompson, and Elizabeth R Sowell. 2012. “ Along-tract statistics allow for enhanced tractography analysis ”. *Neuroimage*, vol. 59, n° 4, p. 3227–3242.

Colclough, Giles L, Stephen M Smith, Thomas E Nichols, Anderson M Winkler, Stamatios N Sotiropoulos, Matthew F Glasser, David C Van Essen, and Mark W Woolrich. 2017. “ The heritability of multi-modal connectivity in human brain activity ”. *Elife*, vol. 6.

Collins, Francis S and Harold Varmus. 2015. “ A new initiative on precision medicine ”. *New England Journal of Medicine*, vol. 372, n° 9, p. 793–795.

Conturo, Thomas E, Nicolas F Lori, Thomas S Cull, Erbil Akbudak, Abraham Z Snyder, Joshua S Shimony, Robert C McKinstry, Harold Burton, and Marcus E Raichle. 1999. “ Tracking neuronal fiber pathways in the living human brain ”. *Proceedings of the National Academy of Sciences*, vol. 96, n° 18, p. 10422–10427.

Corouge, Isabelle, Sylvain Gouttard, and Guido Gerig. 2004. “ Towards a shape model of white matter fiber bundles using diffusion tensor MRI ”. In *Biomedical Imaging: Nano to Macro, 2004. IEEE International Symposium on*. p. 344–347. IEEE.

Corouge, Isabelle, P Thomas Fletcher, Sarang Joshi, Sylvain Gouttard, and Guido Gerig. 2006. “ Fiber tract-oriented statistics for quantitative diffusion tensor MRI analysis ”. *Medical Image Analysis*, vol. 10, n° 5, p. 786–798.

Côté, Marc-Alexandre, Gabriel Girard, Arnaud Boré, Eleftherios Garyfallidis, Jean-Christophe Houde, and Maxime Descoteaux. 2013. “ Tractometer: towards validation of tractography pipelines ”. *Medical image analysis*, vol. 17, n° 7, p. 844–857.

Cousineau, Martin, Pierre-Marc Jodoin, Eleftherios Garyfallidis, Marc-Alexandre Côté, Félix C Morency, Verena Rozanski, Marilyn Grand’Maison, Barry J Bedell, and Maxime Descoteaux. 2017. “ A test-retest study on Parkinson’s PPMI dataset yields statistically significant white matter fascicles ”. *NeuroImage: Clinical*, vol. 16, p. 222–233.

Csurka, Gabriella, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. 2004. “ Visual categorization with bags of keypoints ”. In *Workshop on statistical learning in computer vision, ECCV*. p. 1–2. Prague.

Daducci, Alessandro, Dimitri Van De Ville, Jean-Philippe Thiran, and Yves Wiaux. 2014. “ Sparse regularization for fiber ODF reconstruction: From the suboptimality of $l_2$ and $l_1$ priors to $l_0$ ”. *Medical Image Analysis*, vol. 18, n° 6, p. 820–833.

Daducci, Alessandro, Alessandro Dal Palù, Alia Lemkaddem, and Jean-Philippe Thiran. 2015. “ COMMIT: convex optimization modeling for microstructure informed tractography ”. *IEEE transactions on medical imaging*, vol. 34, n° 1, p. 246–257.

Daducci, Alessandro, Alessandro Dal Palú, Maxime Descoteaux, and Jean-Philippe Thiran. 2016. “ Microstructure informed tractography: pitfalls and open challenges ”. *Frontiers in neuroscience*, vol. 10, p. 247.

Damasio, Hanna, 1995. *Human brain anatomy in computerized images.* Oxford university press.

Davidson, Ian, Sean Gilpin, Owen Carmichael, and Peter Walker. 2013. “ Network discovery via constrained tensor analysis of fMRI data ”. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. p. 194–202. ACM.

de Reus, Marcel A and Martijn P Van den Heuvel. 2013. “ The parcellation-based connectome: limitations and extensions ”. *Neuroimage*, vol. 80, p. 397–404.

de Schotten, Michel Thiebaut, Alberto Bizzi, Flavio Dell’Acqua, Matthew Allin, Muriel Walshe, Robin Murray, Steven C Williams, Declan GM Murphy, Marco Catani, et al. 2011. “ Atlasing location, asymmetry and inter-subject variability of white matter tracts in the human brain with MR diffusion tractography ”. *Neuroimage*, vol. 54, n° 1, p. 49–59.

Descoteaux, Maxime. 2008. “ High angular resolution diffusion MRI: from local estimation to segmentation and tractography ”. PhD thesis, Université Nice Sophia Antipolis.

Descoteaux, Maxime. 2015. “ High angular resolution diffusion imaging (hardi) ”. *Wiley Encyclopedia of Electrical and Electronics Engineering*.

Descoteaux, Maxime, Elaine Angelino, Shaun Fitzgibbons, and Rachid Deriche. 2007. “ Regularized, fast, and robust analytical Q-ball imaging ”. *Magnetic Resonance in Medicine*, vol. 58, n° 3, p. 497–510.

Dhillon, Inderjit S, Yuqiang Guan, and Brian Kulis. 2004. “ Kernel k-means: spectral clustering and normalized cuts ”. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. p. 551–556. ACM.

Ding, Chris, Tao Li, Wei Peng, and Haesun Park. 2006. “ Orthogonal nonnegative matrix t-factorizations for clustering ”. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. p. 126–135. ACM.

Ding, Zhaohua, John C Gore, and Adam W Anderson. 2003. “ Classification and quantification of neuronal fiber pathways using diffusion tensor MRI ”. *Magnetic Resonance in Medicine*, vol. 49, n° 4, p. 716–721.

Dosenbach, Nico UF, Binyam Nardos, Alexander L Cohen, Damien A Fair, Jonathan D Power, Jessica A Church, Steven M Nelson, Gagan S Wig, Alecia C Vogel, Christina N Lessov-Schlaggar, et al. 2010. “ Prediction of individual brain maturity using fMRI ”. *Science*, vol. 329, n° 5997, p. 1358–1361.

Dubois, Julien and Ralph Adolphs. 2016. “ Building a science of individual differences from fMRI ”. *Trends in cognitive sciences*, vol. 20, n° 6, p. 425–443.

Durrleman, Stanley. 2010. “ Statistical models of currents for measuring the variability of anatomical curves, surfaces and their evolution ”. PhD thesis, Université Nice Sophia Antipolis.

Durrleman, Stanley, Pierre Fillard, Xavier Pennec, Alain Trouvé, and Nicholas Ayache. 2009. “ A statistical model of white matter fiber bundles based on currents ”. In *International Conference on Information Processing in Medical Imaging*. p. 114–125. Springer.

Durrleman, Stanley, Pierre Fillard, Xavier Pennec, Alain Trouvé, and Nicholas Ayache. 2011. “Registration, atlas estimation and variability analysis of white matter fiber bundles modeled as currents”. *NeuroImage*, vol. 55, n° 3, p. 1073–1090.

El Kouby, Vincent, Yann Cointepas, Cyril Poupon, Denis Rivière, Narly Golestani, J-B Poline, Denis Le Bihan, and J-F Mangin. 2005. Mr diffusion-based inference of a fiber bundle model from a population of subjects. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2005*, p. 196–204. Springer.

Elad, Michael, Mario AT Figueiredo, and Yi Ma. 2010. “On the role of sparse and redundant representations in image processing”. *Proceedings of the IEEE*, vol. 98, n° 6, p. 972–982.

Elliott, Lloyd, Kevin Sharp, Fidel Alfaro-Almagro, Gwenaelle Douaud, Karla Miller, Jonathan Marchini, and Stephen Smith. 2017. “The genetic basis of human brain structure and function: 1,262 genome-wide associations found from 3,144 GWAS of multimodal brain imaging phenotypes from 9,707 UK Biobank participants”. *BioRxiv*, p. 178806.

Engan, K, SO Aase, and J Hakon Husoy. 1999. “Method of optimal directions for frame design”. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*. p. 2443–2446. IEEE.

Feng, Yuanjing, Ye Wu, Yogesh Rathi, and Carl-Fredrik Westin. 2015. “Sparse deconvolution of higher order tensor for fiber orientation distribution estimation”. *Artificial intelligence in medicine*, vol. 65, n° 3, p. 229–238.

Finn, Emily S and R Todd Constable. 2016. “Individual variation in functional brain connectivity: implications for personalized approaches to psychiatric disease”. *Dialogues in clinical neuroscience*, vol. 18, n° 3, p. 277.

Finn, Emily S, Xilin Shen, Dustin Scheinost, Monica D Rosenberg, Jessica Huang, Marvin M Chun, Xenophon Papademetris, and R Todd Constable. 2015. “Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity”. *Nature neuroscience*.

Finn, Emily S, Dustin Scheinost, Daniel M Finn, Xilin Shen, Xenophon Papademetris, and R Todd Constable. 2017. “Can brain state be manipulated to emphasize individual differences in functional connectivity?”. *NeuroImage*.

Fischl, Bruce. 2012. “FreeSurfer”. *Neuroimage*, vol. 62, n° 2, p. 774–781.

Fleming, Stephen M, Rimona S Weil, Zoltan Nagy, Raymond J Dolan, and Geraint Rees. 2010. “Relating introspective accuracy to individual differences in brain structure”. *Science*, vol. 329, n° 5998, p. 1541–1543.

Fortin, David, Camille Aubin-Lemay, Arnaud Boré, Gabriel Girard, Jean-Christophe Houde, Kevin Whittingstall, and Maxime Descoteaux. 2012. “Tractography in the study of the human brain: a neurosurgical perspective”. *The Canadian Journal of Neurological Sciences*, vol. 39, n° 6, p. 747–756.

Foulkes, Lucy and Sarah-Jayne Blakemore. 2018. “Studying individual differences in human adolescent brain development”. *Nature neuroscience*, p. 1.

Fowlkes, Charless, Serge Belongie, Fan Chung, and Jitendra Malik. 2004. “Spectral grouping using the Nystrom method”. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, n° 2, p. 214–225.

Frackowiak, Richard and Henry Markram. 2015. “The future of human cerebral cartography: a novel approach”. *Phil. Trans. R. Soc. B*, vol. 370, n° 1668, p. 20140171.

Fraguas, D, CM Díaz-Caneja, MC O’donovan, RE Gur, C Arango, et al. 2017. “Mental disorders of known aetiology and precision medicine in psychiatry: a promising but neglected alliance”. *Psychological medicine*, vol. 47, n° 2, p. 193–197.

Frangou, Sophia, Danai Dima, and Jigar Jogia. 2017. “Towards person-centered neuroimaging markers for resilience and vulnerability in Bipolar Disorder”. *Neuroimage*, vol. 145, p. 230–237.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010. “A note on the group lasso and a sparse group lasso”. *arXiv preprint arXiv:1001.0736*.

Gale, Samuel D and David J Perkel. 2010. “A basal ganglia pathway drives selective auditory responses in songbird dopaminergic neurons via disinhibition”. *The Journal of Neuroscience*, vol. 30, n° 3, p. 1027–1037.

Gao, Yue, Ehsan Adeli-M, Minjeong Kim, Panteleimon Giannakopoulos, Sven Haller, and Dinggang Shen. 2015. “Medical image retrieval using multi-graph learning for MCI diagnostic assistance”. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. p. 86–93. Springer.

Garyfallidis, Eleftherios, Matthew Brett, Marta Morgado Correia, Guy B Williams, and Ian Nimmo-Smith. 2012. “Quickbundles, a method for tractography simplification”. *Frontiers in neuroscience*, vol. 6.

Garyfallidis, Eleftherios, Matthew Brett, Bagrat Amirbekian, Ariel Rokem, Stefan Van Der Walt, Maxime Descoteaux, and Ian Nimmo-Smith. 2014. “Dipy, a library for the analysis of diffusion MRI data”. *Frontiers in neuroinformatics*, vol. 8, p. 8.

Garyfallidis, Eleftherios, Omar Ocegueda, Demian Wassermann, and Maxime Descoteaux. 2015. “Robust and efficient linear registration of white-matter fascicles in the space of streamlines”. *NeuroImage*, vol. 117, p. 124–140.

Ge, Tian, Martin Reuter, Anderson M Winkler, Avram J Holmes, Phil H Lee, Lee S Tirrell, Joshua L Roffman, Randy L Buckner, Jordan W Smoller, and Mert R Sabuncu. 2016. “Multidimensional heritability analysis of neuroanatomical shape”. *Nature communications*, vol. 7, p. 13291.

Geevarghese, Ruben, Daniel E Lumsden, Natasha Hulse, Michael Samuel, and Keyoumars Ashkan. 2015. “ Subcortical Structure Volumes and Correlation to Clinical Variables in Parkinson’s Disease ”. *Journal of Neuroimaging*, vol. 25, n° 2, p. 275–280.

Gerber, Samuel, Tolga Tasdizen, P Thomas Fletcher, Sarang Joshi, Ross Whitaker, Alzheimers Disease Neuroimaging Initiative, et al. 2010. “ Manifold modeling for brain population analysis ”. *Medical image analysis*, vol. 14, n° 5, p. 643–653.

Gerig, Guido, Sylvain Gouttard, and Isabelle Corouge. 2004. “ 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.* p. 4421–4424. IEEE.

Geschwind, Norman and Albert M Galaburda. 1985. “ Cerebral lateralization: Biological mechanisms, associations, and pathology: I. A hypothesis and a program for research ”. *Archives of neurology*, vol. 42, n° 5, p. 428–459.

Girard, Gabriel, Kevin Whittingstall, Rachid Deriche, and Maxime Descoteaux. 2014. “ Towards quantitative connectivity analysis: reducing tractography biases ”. *Neuroimage*, vol. 98, p. 266–278.

Glasser, Matthew F, Stamatios N Sotiropoulos, J Anthony Wilson, Timothy S Coalson, Bruce Fischl, Jesper L Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R Polimeni, et al. 2013. “ The minimal preprocessing pipelines for the Human Connectome Project ”. *Neuroimage*, vol. 80, p. 105–124.

Glasser, Matthew F, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, et al. 2016. “ A multi-modal parcellation of human cerebral cortex ”. *Nature*, vol. 536, n° 7615, p. 171–178.

Goldman, Sylvie, Liam M O’Brien, Pauline A Filipek, Isabelle Rapin, and Martha R Herbert. 2013. “ Motor stereotypies and volumetric brain alterations in children with Autistic Disorder ”. *Research in autism spectrum disorders*, vol. 7, n° 1, p. 82–92.

Gómez-Robles, Aida, William D Hopkins, and Chet C Sherwood. 2013. “ Increased morphological asymmetry, evolvability and plasticity in human brain evolution ”. In *Proc. R. Soc. B.* p. 20130575. The Royal Society.

Gong, Gaolang, Yong He, and Alan C Evans. 2011. “ Brain connectivity gender makes a difference ”. *The Neuroscientist*, vol. 17, n° 5, p. 575–591.

Gordon, Evan M, Timothy O Laumann, Babatunde Adeyemo, Adrian W Gilmore, Steven M Nelson, Nico UF Dosenbach, and Steven E Petersen. 2017a. “ Individual-specific features of brain systems identified with resting state functional correlations ”. *NeuroImage*, vol. 146, p. 918–939.

Gordon, Evan M, Timothy O Laumann, Babatunde Adeyemo, and Steven E Petersen. 2017b. “ Individual variability of the system-level organization of the human brain ”. *Cerebral Cortex*, vol. 27, n° 1, p. 386–399.

Gordon, Evan M, Timothy O Laumann, Adrian W Gilmore, Dillan J Newbold, Deanna J Greene, Jeffrey J Berg, Mario Ortega, Catherine Hoyt-Drazen, Caterina Gratton, Haoxin Sun, et al. 2017c. “ Precision functional mapping of individual human brains ”. *Neuron*, vol. 95, n° 4, p. 791–807.

Gori, Pietro, Olivier Colliot, Linda Marrakchi-Kacem, Yulia Worbe, Fabrizio De Vico Fallani, Mario Chavez, Cyril Poupon, Andreas Hartmann, Nicholas Ayache, and Stanley Durrleman. 2016. “ Parsimonious approximation of streamline trajectories in white matter fiber bundles ”. *IEEE Transactions on Medical Imaging*, vol. 35, n° 12, p. 2609–2619.

Gori, Pietro, Olivier Colliot, Linda Marrakchi-Kacem, Yulia Worbe, Cyril Poupon, Andreas Hartmann, Nicholas Ayache, and Stanley Durrleman. 2017. “ A Bayesian framework for joint morphometry of surface and curve meshes in multi-object complexes ”. *Medical Image Analysis*, vol. 35, p. 458–474.

Grabner, Roland H, Daniel Ansari, Gernot Reishofer, Elsbeth Stern, Franz Ebner, and Christa Neuper. 2007. “ Individual differences in mathematical competence predict parietal brain activation during mental calculation ”. *Neuroimage*, vol. 38, n° 2, p. 346–356.

Gramfort, Alexandre, Cyril Poupon, and Maxime Descoteaux. 2014. “ Denoising and fast diffusion imaging with physically constrained sparse dictionary learning ”. *Medical image analysis*, vol. 18, n° 1, p. 36–49.

Gray, Henry, 1878. *Anatomy of the human body*, volume 8. Lea & Febiger.

Griffa, Alessandra, Philipp S Baumann, Jean-Philippe Thiran, and Patric Hagmann. 2013. “ Structural connectomics in brain diseases ”. *Neuroimage*, vol. 80, p. 515–526.

Groves, Adrian R, Stephen M Smith, Anders M Fjell, Christian K Tamnes, Kristine B Walhovd, Gwenaëlle Douaud, Mark W Woolrich, and Lars T Westlye. 2012. “ Benefits of multimodal fusion analysis on a large-scale dataset: life-span patterns of inter-subject variability in cortical morphometry and white matter microstructure ”. *Neuroimage*, vol. 63, n° 1, p. 365–380.

Gu, Jenny and Ryota Kanai. 2014. “ What contributes to individual differences in brain structure? ”. *Frontiers in human neuroscience*, vol. 8, p. 262.

Guevara, Pamela, Cyril Poupon, Denis Rivière, Yann Cointepas, Maxime Descoteaux, Bertrand Thirion, and J-F Mangin. 2011. “ Robust clustering of massive tractography datasets ”. *Neuroimage*, vol. 54, n° 3, p. 1975–1993.

Guevara, Pamela, Delphine Duclap, Cyril Poupon, Linda Marrakchi-Kacem, Pierre Fillard, Denis Le Bihan, Marion Leboyer, Josselin Houenou, and J-F Mangin. 2012. “ Automatic fiber bundle segmentation in massive tractography datasets using a multi-subject bundle atlas ”. *Neuroimage*, vol. 61, n° 4, p. 1083–1099.

Hagmann, Patric, Lisa Jonasson, Philippe Maeder, Jean-Philippe Thiran, Van J Wedeen, and Reto Meuli. 2006. “Understanding diffusion mr imaging techniques: From scalar diffusion-weighted imaging to diffusion tensor imaging and beyond1”. *Radiographics*, vol. 26, n° suppl 1, p. S205–S223.

Halpern, Diane F, 2013. *Sex differences in cognitive abilities*. Psychology press.

Hampel, Harald, SE O'Bryant, Stanley Durrleman, Erfan Younesi, Katrine Rojkova, Valentina Escott-Price, Jean-Christophe Corvol, Karl Broich, Bruno Dubois, Simone Lista, et al. 2017. “A precision medicine initiative for Alzheimer's disease: the road ahead to biomarker-guided integrative disease modeling”. *Climacteric*, vol. 20, n° 2, p. 107–118.

Handl, Julia, Joshua Knowles, and Douglas B Kell. 2005. “Computational cluster validation in post-genomic data analysis”. *Bioinformatics*, vol. 21, n° 15, p. 3201–3212.

Hassibi, Arjang, Aydin Babakhani, and Ali Hajimiri. 2009. “A spectral-scanning nuclear magnetic resonance imaging (MRI) transceiver”. *IEEE Journal of Solid-State Circuits*, vol. 44, n° 6, p. 1805–1813.

Herculano-Houzel, Suzana. 2009. “The human brain in numbers: a linearly scaled-up primate brain”. *Frontiers in human neuroscience*, vol. 3, p. 31.

Hofmann, T, B Schölkopf, and AJ Smola. 2008. “Kernel methods in machine learning”. *The annals of statistics*, p. 1171–1220.

Holm, Sture. 1979. “A simple sequentially rejective multiple test procedure”. *Scandinavian journal of statistics*, p. 65–70.

Horien, Corey, Stephanie Noble, Emily S Finn, Xilin Shen, Dustin Scheinost, and R Todd Constable. 2018. “Considering factors affecting the connectome-based identification process: Comment on Waller et al.”. *NeuroImage*, vol. 169, p. 172–175.

Huettel, Scott A. 2012. “Event-related fMRI in cognition”. *Neuroimage*, vol. 62, n° 2, p. 1152–1156.

Huo, Xiaoming, Xuelei Sherry Ni, and Andrew K Smith. 2007. “A survey of manifold-based learning methods”. *Recent advances in data mining of enterprise data*, p. 691–745.

Iglesias, Juan Eugenio, Paul M Thompson, Cheng-Yi Liu, and Zhuowen Tu. 2012. “Fast approximate stochastic tractography”. *Neuroinformatics*, vol. 10, n° 1, p. 5–17.

Itahashi, Takashi, Takashi Yamada, Motoaki Nakamura, Hiromi Watanabe, Bun Yamagata, Daiki Jimbo, Seiji Shioda, Miho Kuroda, Kazuo Toriizuka, Nobumasa Kato, et al. 2015. “Linked alterations in gray and white matter morphology in adults with high-functioning autism spectrum disorder: a multimodal brain imaging study”. *NeuroImage: Clinical*, vol. 7, p. 155–169.

Jacobs, Michael A, Tamer S Ibrahim, and Ronald Ouwerkerk. 2007. “MR imaging: brief overview and emerging applications”. *Radiographics*, vol. 27, n° 4, p. 1213–1229.

Jahanshad, Neda, Derrek P Hibar, April Ryles, Arthur W Toga, Katie L McMahon, Greig I De Zubicaray, Narelle K Hansell, Grant W Montgomery, Nicholas G Martin, Margaret J Wright, et al. 2012. “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)*. p. 542–545. IEEE.

Jahanshad, Neda, Peter V Kochunov, Emma Sprooten, René C Mandl, Thomas E Nichols, Laura Almasy, John Blangero, Rachel M Brouwer, Joanne E Curran, Greig I de Zubicaray, et al. 2013. “Multi-site genetic analysis of diffusion images and voxelwise heritability analysis: A pilot project of the ENIGMA–DTI working group”. *Neuroimage*, vol. 81, p. 455–469.

Jbabdi, Saad, Stamatios N Sotiropoulos, Suzanne N Haber, David C Van Essen, and Timothy E Behrens. 2015. “Measuring macroscopic brain connections in vivo”. *Nature neuroscience*, vol. 18, n° 11, p. 1546.

Jegelka, Stefanie, Suvrit Sra, and Arindam Banerjee. 2009. “Approximation algorithms for tensor clustering”. In *Algorithmic learning theory*. p. 368–383. Springer.

Jenkinson, Mark, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. 2012. “Fsl”. *Neuroimage*, vol. 62, n° 2, p. 782–790.

Jiang, Yu-Gang, Chong-Wah Ngo, and Jun Yang. 2007. “Towards optimal bag-of-features for object categorization and semantic video retrieval”. In *Proceedings of the 6th ACM international conference on Image and video retrieval*. p. 494–501. ACM.

Jin, Yan, Yonggang Shi, Liang Zhan, Junning Li, Greig I de Zubicaray, Katie L McMahon, Nicholas G Martin, Margaret J Wright, and Paul M Thompson. 2012. Automatic population hardi white matter tract clustering by label fusion of multiple tract atlases. *Multimodal Brain Image Analysis*, p. 147–156. Springer.

Jin, Yan, Yonggang Shi, Liang Zhan, Boris A Gutman, Greig I de Zubicaray, Katie L McMahon, Margaret J Wright, Arthur W Toga, and Paul M Thompson. 2014. “Automatic clustering of white matter fibers in brain diffusion MRI with an application to genetics”. *NeuroImage*, vol. 100, p. 75–90.

Jonasson, L, P Hagmann, JP Thiran, VJ Wedeen, et al. 2005. “Fiber tracts of high angular resolution diffusion MRI are easily segmented with spectral clustering”. In *International Society for Magnetic Resonance in Medicine*. p. 1127–1137.

Jones, Derek K, Thomas R Knösche, and Robert Turner. 2013. “White matter integrity, fiber count, and other fallacies: the do’s and don’ts of diffusion MRI”. *Neuroimage*, vol. 73, p. 239–254.

Kanai, Ryota and Geraint Rees. 2011. “The structural basis of inter-individual differences in human behaviour and cognition”. *Nature Reviews Neuroscience*, vol. 12, n° 4, p. 231.

Kanchibhotla, Sri C, Karen A Mather, Wei Wen, Peter R Schofield, John BJ Kwok, and Perminder S Sachdev. 2013. “Genetics of ageing-related changes in brain white matter integrity–A review”. *Ageing research reviews*, vol. 12, n° 1, p. 391–401.

Kane, Michael J and Randall W Engle. 2002. “The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective”. *Psychonomic bulletin & review*, vol. 9, n° 4, p. 637–671.

Khoje, Suchitra and Shrikant Bodhe. 2012. “Performance Comparison of Fourier Transform and Its Derivatives as Shape Descriptors for Mango Grading.”. *International Journal of Computer Applications*, vol. 53.

Kim, Hyunsoo and Haesun Park. 2007. “Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis”. *Bioinformatics*, vol. 23, n° 12, p. 1495–1502.

Kim, Sungho, Kuk-Jin Yoon, and In So Kweon. 2008. “Object recognition using a generalized robust invariant feature and Gestalt’s law of proximity and similarity”. *Pattern Recognition*, vol. 41, n° 2, p. 726–741.

Kochunov, P, David C Glahn, J Lancaster, Paul M Thompson, V Kochunov, B Rogers, P Fox, John Blangero, and DE Williamson. 2011. “Fractional anisotropy of cerebral white matter and thickness of cortical gray matter across the lifespan”. *Neuroimage*, vol. 58, n° 1, p. 41–49.

Kochunov, P, DE Williamson, J Lancaster, P Fox, J Cornell, J Blangero, and DC Glahn. 2012. “Fractional anisotropy of water diffusion in cerebral white matter across the lifespan”. *Neurobiology of aging*, vol. 33, n° 1, p. 9–20.

Kochunov, Peter, Neda Jahanshad, Emma Sprooten, Thomas E Nichols, René C Mandl, Laura Almasy, Tom Booth, Rachel M Brouwer, Joanne E Curran, Greig I de Zubicaray, et al. 2014. “Multi-site study of additive genetic effects on fractional anisotropy of cerebral white matter: comparing meta and megaanalytical approaches for data pooling”. *NeuroImage*, vol. 95, p. 136–150.

Kochunov, Peter, Neda Jahanshad, Daniel Marcus, Anderson Winkler, Emma Sprooten, Thomas E Nichols, Susan N Wright, L Elliot Hong, Binish Patel, Timothy Behrens, et al. 2015. “Heritability of fractional anisotropy in human white matter: a comparison of Human Connectome Project and ENIGMA-DTI data”. *Neuroimage*, vol. 111, p. 300–311.

Kohannim, Omid, Neda Jahanshad, Meredith N Braskie, Jason L Stein, Ming-Chang Chiang, April H Reese, Derrek P Hibar, Arthur W Toga, Katie L McMahon, Greig I de Zubicaray, et al. 2012. “Predicting white matter integrity from multiple common genetic variants”. *Neuropsychopharmacology*, vol. 37, n° 9.

Kong, Xiangnan and Philip S Yu. 2014. “Brain network analysis: a data mining perspective”. *ACM SIGKDD Explorations Newsletter*, vol. 15, n° 2, p. 30–38.

Kruskal, Joseph B and Myron Wish, 1978. *Multidimensional scaling*, volume 11. Sage.

Kudo, Hiroyuki, Taizo Suzuki, and Essam A Rashed. 2013. “Image reconstruction for sparse-view CT and interior CT—introduction to compressed sensing and differentiated back-projection”. *Quantitative imaging in medicine and surgery*, vol. 3, n° 3, p. 147.

Kumar, Kuldeep and Christian Desrosiers. 2014. “Group Sparse Kernelized Dictionary Learning for the Clustering of White Matter Fibers”.

Kumar, Kuldeep and Christian Desrosiers. 2016. “A sparse coding approach for the efficient representation and segmentation of white matter fibers”. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. p. 915–919. IEEE.

Kumar, Kuldeep, Christian Desrosiers, and Kaleem Siddiqi. 2015. “Brain Fiber Clustering Using Non-negative Kernelized Matching Pursuit”. In *International Workshop on Machine Learning in Medical Imaging*. p. 144–152. Springer.

Kumar, Kuldeep, Laurent Chauvin, Mathew Toews, Olivier Colliot, and Christian Desrosiers. 2017a. “Multi-modal analysis of genetically-related subjects using SIFT descriptors in brain MRI”. In *Workshop on Computational Diffusion MRI, CDMRI 2017, MICCAI Workshop*.

Kumar, Kuldeep, Christian Desrosiers, Kaleem Siddiqi, Olivier Colliot, and Matthew Toews. 2017b. “Fiberprint: A subject fingerprint based on sparse code pooling for white matter fiber analysis”. *NeuroImage*, vol. 158, p. 242 - 259.

Kumar, Kuldeep, Christian Desrosiers, Kaleem Siddiqi, Olivier Colliot, and Matthew Toews. 2017c. “Fiberprint: A subject fingerprint based on sparse code pooling for white matter fiber analysis”. *NeuroImage*, vol. 158, p. 242–259.

Kumar, Kuldeep, Pietro Gori, Benjamin Charlier, Stanley Durrleman, Olivier Colliot, and Christian Desrosiers. 2017d. White matter fiber segmentation using functional varifolds. *Graphs in Biomedical Image Analysis, Computational Anatomy and Imaging Genetics*, p. 92–100. Springer.

Laub, Julian and Klaus-Robert Müller. 2004. “Feature discovery in non-metric pairwise data”. *The Journal of Machine Learning Research*, vol. 5, p. 801–818.

Laumann, Timothy O, Evan M Gordon, Babatunde Adeyemo, Abraham Z Snyder, Sung Jun Joo, Mei-Yen Chen, Adrian W Gilmore, Kathleen B McDermott, Steven M Nelson, Nico UF Dosenbach, et al. 2015. “Functional system and areal organization of a highly sampled individual human brain”. *Neuron*, vol. 87, n° 3, p. 657–670.

Lauterbur, PC et al. 1973. “Image formation by induced local interactions: examples employing nuclear magnetic resonance”.

Lazar, M, DM Weinstein, et al. 2003. “White matter tractography using diffusion tensor deflection”. *Human brain mapping*, vol. 18, n° 4, p. 306–321.

Le Bihan, Denis and E Breton. 1985. “Imagerie de diffusion in-vivo par résonance magnétique nucléaire”. *Comptes-Rendus de l’Académie des Sciences*, vol. 93, n° 5, p. 27–34.

Le Bihan, Denis, Eric Breton, Denis Lallemand, Philippe Grenier, Emmanuel Cabanis, and Maurice Laval-Jeantet. 1986. “MR imaging of intravoxel incoherent motions: application to diffusion and perfusion in neurologic disorders.”. *Radiology*, vol. 161, n° 2, p. 401–407.

Le Bihan, Denis, Peter Jezzard, James Haxby, Norihiro Sadato, Linda Rueckert, and Venkata Mattay. 1995. “Functional magnetic resonance imaging of the brain”. *Annals of Internal Medicine*, vol. 122, n° 4, p. 296–303.

Lee, Kangjoo, Jean-Marc Lina, Jean Gotman, and Christophe Grova. 2016a. “SPARK: Sparsity-based analysis of reliable k-hubness and overlapping network structure in brain functional connectivity”. *NeuroImage*, vol. 134, p. 434–449.

Lee, Megan H, Christopher D Smyser, and Joshua S Shimony. 2013. “Resting-state fMRI: a review of methods and clinical applications”. *American Journal of Neuroradiology*, vol. 34, n° 10, p. 1866–1872.

Lee, Young-Beom, Jeonghyeon Lee, Sungho Tak, Kangjoo Lee, Duk L Na, Sang Won Seo, Yong Jeong, Jong Chul Ye, Alzheimer’s Disease Neuroimaging Initiative, et al. 2016b. “Sparse SPM: Group Sparse-dictionary learning in SPM framework for resting-state functional connectivity MRI analysis”. *NeuroImage*, vol. 125, p. 1032–1045.

Leonardi, Nora, William R Shirer, Michael D Greicius, and Dimitri Van De Ville. 2014. “Disentangling dynamic networks: Separated and joint expressions of functional connectivity patterns in time”. *Human brain mapping*.

Levandowsky, Michael and David Winter. 1971. “Distance between sets”. *Nature*, vol. 234, n° 5323, p. 34–35.

Li, Hai, Zhong Xue, Lei Guo, Tianming Liu, Jill Hunter, and Stephen TC Wong. 2010. “A hybrid approach to automatic clustering of white matter fibers”. *NeuroImage*, vol. 49, n° 2, p. 1249–1258.

Li, Kaiming, Lei Guo, Jingxin Nie, Gang Li, and Tianming Liu. 2009. “Review of methods for functional brain connectivity detection using fMRI”. *Computerized Medical Imaging and Graphics*, vol. 33, n° 2, p. 131–139.

Li, Xin-Wei, Qiong-Ling Li, Shu-Yu Li, and De-Yu Li. 2015. “Local manifold learning for multiatlas segmentation: application to hippocampal segmentation in healthy population and Alzheimer’s disease”. *CNS neuroscience & therapeutics*, vol. 21, n° 10, p. 826–836.

Liang, Xuwei and Jie Wang. 2013. “Using Fourier Descriptor Features in the Classification of White Matter Fiber Tracts in DTI”. In *Computational and Information Sciences (ICCIS), 2013 Fifth International Conference on*. p. 694–697. IEEE.

Liang, Xuwei, Qi Zhuang, Ning Cao, and Changjiang Zhang. 2009. “Shape modeling and clustering of white matter fiber tracts using fourier descriptors”. In *Computational Intelligence in Bioinformatics and Computational Biology, 2009. CIBCB’09. IEEE Symposium on*. p. 292–297. IEEE.

Liem, Franziskus, Gaël Varoquaux, Jana Kynast, Frauke Beyer, Shahrzad Kharabian Masouleh, Julia M Huntenburg, Leonie Lampe, Mehdi Rahim, Alexandre Abraham, R Cameron Craddock, et al. 2017. “Predicting brain-age from multimodal imaging data captures cognitive impairment”. *NeuroImage*, vol. 148, p. 179–188.

Liu, Jin, Xuhong Liao, Mingrui Xia, and Yong He. 2018. “Chronnectome fingerprinting: Identifying individuals and predicting higher cognitive functions using dynamic brain connectivity patterns”. *Human brain mapping*, vol. 39, n° 2, p. 902–915.

Liu, Sidong, Weidong Cai, Siqi Liu, Fan Zhang, Michael Fulham, Dagan Feng, Sonia Pujol, and Ron Kikinis. 2015a. “Multimodal neuroimaging computing: a review of the applications in neuropsychiatric disorders”. *Brain informatics*, vol. 2, n° 3, p. 167–180.

Liu, Sidong, Weidong Cai, Siqi Liu, Fan Zhang, Michael Fulham, Dagan Feng, Sonia Pujol, and Ron Kikinis. 2015b. “Multimodal neuroimaging computing: the workflows, methods, and platforms”. *Brain informatics*, vol. 2, n° 3, p. 181–195.

Liu, Thomas T. 2012. “The development of event-related fMRI designs”. *Neuroimage*, vol. 62, n° 2, p. 1157–1162.

Logothetis, Nikos K. 2008. “What we can do and what we cannot do with fMRI”. *Nature*, vol. 453, n° 7197, p. 869.

Lowe, David G. 1999. “Object recognition from local scale-invariant features”. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*. p. 1150–1157. Ieee.

Lowe, David G. 2004. “Distinctive image features from scale-invariant keypoints”. *International journal of computer vision*, vol. 60, n° 2, p. 91–110.

Lustig, Michael, David L Donoho, Juan M Santos, and John M Pauly. 2008. “Compressed sensing MRI”. *IEEE Signal Processing Magazine*, vol. 25, n° 2, p. 72–82.

Maddah, Mahnaz, Andrea UJ Mewes, Steven Haker, W Eric L Grimson, and Simon K Warfield. 2005. Automated atlas-based clustering of white matter fiber tracts from dtmri. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2005*, p. 188–195. Springer.

Maddah, Mahnaz, WEL Crimson, and Simon K Warfield. 2006. “ Statistical modeling and EM clustering of white matter fiber tracts ”. In *Biomedical Imaging: Nano to Macro, 2006. 3rd IEEE International Symposium on*. p. 53–56. IEEE.

Maddah, Mahnaz, William M Wells III, Simon K Warfield, Carl-Fredrik Westin, and W Eric L Grimson. 2007. “ Probabilistic clustering and quantitative analysis of white matter fiber tracts ”. In *Information processing in medical imaging*. p. 372–383. Springer.

Maddah, Mahnaz, W Eric L Grimson, Simon K Warfield, and William M Wells. 2008. “ A unified framework for clustering and quantitative analysis of white matter fiber tracts ”. *Medical image analysis*, vol. 12, n° 2, p. 191–202.

Maier-Hein, Klaus H, Peter F Neher, Jean-Christophe Houde, Marc-Alexandre Côté, Eleftherios Garyfallidis, Jidan Zhong, Maxime Chamberland, Fang-Cheng Yeh, Ying-Chia Lin, Qing Ji, et al. 2017. “ The challenge of mapping the human connectome based on diffusion tractography ”. *Nature communications*, vol. 8, n° 1, p. 1349.

Mairal, Julien, Francis Bach, Jean Ponce, and Guillermo Sapiro. 2009. “ Online dictionary learning for sparse coding ”. In *Proceedings of the 26th Annual International Conference on Machine Learning*. p. 689–696. ACM.

Malcolm, James G, Martha E Shenton, and Yogesh Rathi. 2009. “ Neural tractography using an unscented Kalman filter ”. In *Information Processing in Medical Imaging*. p. 126–138. Springer.

Mangin, J-F, D Riviere, A Cachia, E Duchesnay, Y Cointepas, D Papadopoulos-Orfanos, Paola Scifo, T Ochiai, F Brunelle, and J Regis. 2004. “ A framework to study the cortical folding patterns ”. *Neuroimage*, vol. 23, p. S129–S138.

Marek, Kenneth, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Chris Coffey, Karl Kieburtz, Emily Flagg, Sohini Chowdhury, et al. 2011. “ The parkinson progression marker initiative (PPMI) ”. *Progress in neurobiology*, vol. 95, n° 4, p. 629–635.

McGill, Robert, John W Tukey, and Wayne A Larsen. 1978. “ Variations of box plots ”. *The American Statistician*, vol. 32, n° 1, p. 12–16.

Merlet, Sylvain, Emmanuel Caruyer, Aurobrata Ghosh, and Rachid Deriche. 2013. “ A computational diffusion MRI and parametric dictionary learning framework for modeling the diffusion signal and its features ”. *Medical Image Analysis*, vol. 17, n° 7, p. 830–843.

Mikolajczyk, Krystian and Cordelia Schmid. 2005. “ A performance evaluation of local descriptors ”. *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, n° 10, p. 1615–1630.

Miranda-Dominguez, Oscar, Brian D Mills, Samuel D Carpenter, Kathleen A Grant, Christopher D Kroenke, Joel T Nigg, and Damien A Fair. 2014. “ Connectotyping: model based fingerprinting of the functional connectome ”. *PloS one*, vol. 9, n° 11, p. e111048.

Miranda-Domínguez, Óscar, Eric Feczko, David S Grayson, Hasse Walum, Joel T Nigg, and Damien A Fair. 2017. “Heritability of the human connectome: a connectotyping study”. *Network Neuroscience*, , p. 1–48.

Mišić, Bratislav and Olaf Sporns. 2016. “From regions to connections and networks: new bridges between brain and behavior”. *Current opinion in neurobiology*, vol. 40, p. 1–7.

Moberts, Bart, Anna Vilanova, and Jarke J van Wijk. 2005. “Evaluation of fiber clustering methods for diffusion tensor imaging”. In *Visualization, 2005. VIS 05. IEEE.* p. 65–72. IEEE.

MomayyezSiahkal, Parya and Kaleem Siddiqi. 2013. “3D stochastic completion fields for mapping connectivity in diffusion MRI”. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, n° 4, p. 983–995.

Moreno, Gali Zimmerman, Guy Alexandroni, Nir Sochen, and Hayit Greenspan. 2016. “Sparse Representation for White Matter Fiber Compression and Calculation of Inter-Fiber Similarity”. In *International Conference on Medical Image Computing and Computer-Assisted Intervention.* p. 133–143. Springer.

Mori, Susumu and Peter van Zijl. 2002. “Fiber tracking: principles and strategies–a technical review”. *NMR in Biomedicine*, vol. 15, n° 7-8, p. 468–480.

Mori, Susumu, Barbara J Crain, VP Chacko, and Peter Van Zijl. 1999. “Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging”. *Annals of neurology*, vol. 45, n° 2, p. 265–269.

Mueller, Sophia, Danhong Wang, Michael D Fox, BT Thomas Yeo, Jorge Sepulcre, Mert R Sabuncu, Rebecca Shafee, Jie Lu, and Hesheng Liu. 2013. “Individual variability in functional connectivity architecture of the human brain”. *Neuron*, vol. 77, n° 3, p. 586–595.

Mueller, Susanne G, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. 2005. “The Alzheimer’s disease neuroimaging initiative”. *Neuroimaging Clinics*, vol. 15, n° 4, p. 869–877.

Muja, Marius and David G Lowe. 2009. “Fast approximate nearest neighbors with automatic algorithm configuration.”. *VISAPP (1)*, vol. 2, n° 331-340, p. 2.

Müller, Henning, Alba Garcia Seco de Herrera, Jayashree Kalpathy-Cramer, Dina Demner-Fushman, Sameer K Antani, and Ivan Eggel. 2012. “Overview of the ImageCLEF 2012 Medical Image Retrieval and Classification Tasks.”. In *CLEF (online working notes/labs/workshop).* p. 1–16.

Neher, Peter F, Maxime Descoteaux, Jean-Christophe Houde, Bram Stieltjes, and Klaus H Maier-Hein. 2015. “Strengths and weaknesses of state of the art fiber tractography pipelines–a comprehensive in-vivo and phantom evaluation study using tractometer”. *Medical image analysis*, vol. 26, n° 1, p. 287–305.

Nesterov, Yurii et al. 2007. “Gradient methods for minimizing composite objective function”.

Ng, Bernard, Matthew Toews, Stanley Durrleman, and Yonggang Shi. 2014. Shape analysis for brain structures. *Shape Analysis in Medical Image Analysis*, p. 3–49. Springer.

Nguyen, HV, Vishal M Patel, Nasser M Nasrabadi, and Rama Chellappa. 2012. “Kernel dictionary learning”. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. p. 2021–2024. IEEE.

Ning, Lipeng, Frederik Laun, Yaniv Gur, Edward VR DiBella, Samuel Deslauriers-Gauthier, Thinhinane Megherbi, Aurobrata Ghosh, Mauro Zucchelli, Gloria Menegaz, Rutger Fick, et al. 2015. “Sparse Reconstruction Challenge for diffusion MRI: Validation on a physical phantom to determine which acquisition scheme and analysis method to use?”. *Medical image analysis*, vol. 26, n° 1, p. 316–331.

Nir, Talia M, Neda Jahanshad, Julio E Villalon-Reina, Arthur W Toga, Clifford R Jack, Michael W Weiner, Paul M Thompson, Alzheimer’s Disease Neuroimaging Initiative (ADNI, et al. 2013. “Effectiveness of regional DTI measures in distinguishing Alzheimer’s disease, MCI, and normal aging”. *NeuroImage: clinical*, vol. 3, p. 180–195.

Noble, Stephanie, Dustin Scheinost, Emily S Finn, Xilin Shen, Xenophon Papademetris, Sarah C McEwen, Carrie E Bearden, Jean Addington, Bradley Goodyear, Kristin S Cadenhead, et al. 2017a. “Multisite reliability of MR-based functional connectivity”. *Neuroimage*, vol. 146, p. 959–970.

Noble, Stephanie, Marisa N Spann, Fuyuze Tokoglu, Xilin Shen, R Todd Constable, and Dustin Scheinost. 2017b. “Influences on the test–retest reliability of functional connectivity MRI and its relationship with behavioral utility”. *Cerebral Cortex*, vol. 27, n° 11, p. 5415–5429.

O’Donnell, Lauren and Carl-Fredrik Westin. 2005. White matter tract clustering and correspondence in populations. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2005*, p. 140–147. Springer.

O’Donnell, Lauren and Carl-Fredrik Westin. 2006. “High-dimensional white matter atlas generation and group analysis”. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. p. 243–251. Springer.

O’Donnell, Lauren J and C-F Westin. 2007a. “Automatic tractography segmentation using a high-dimensional white matter atlas”. *Medical Imaging, IEEE Transactions on*, vol. 26, n° 11, p. 1562–1575.

O’Donnell, Lauren J. and Carl-Fredrik Westin. 2007b. “Automatic tractography segmentation using a highdimensional white matter atlas”. *IEEE Trans. Med. Imag*, p. 1562–1575.

O’Donnell, Lauren J, Laura Rigolo, Isaiah Norton, William M Wells III, Carl-Fredrik Westin, and Alexandra J Golby. 2012. “fMRI-DTI modeling via landmark distance atlases for prediction and detection of fiber tracts”. *Neuroimage*, vol. 60, n° 1, p. 456–470.

O'Donnell, Lauren J, William M Wells III, Alexandra J Golby, and Carl-Fredrik Westin. 2012. " Unbiased groupwise registration of white matter tractography ". In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. p. 123–130. Springer.

O'Donnell, Lauren J, Alexandra J Golby, and Carl-Fredrik Westin. 2013. " Fiber clustering versus the parcellation-based connectome ". *Neuroimage*, vol. 80, p. 283–289.

O'Donnell, Lauren J, Yannick Suter, Laura Rigolo, Pegah Kahali, Fan Zhang, Isaiah Norton, Angela Albi, Olutayo Olubiyi, Antonio Meola, Walid I Essayed, et al. 2017. " Automated white matter fiber tract identification in patients with brain tumors ". *NeuroImage: Clinical*, vol. 13, p. 138–153.

O'Donnell, LJ, Marek Kubicki, Martha E Shenton, Mark H Dreusicke, W Eric L Grimson, and Carl-Fredrik Westin. 2006. " A method for clustering white matter fiber tracts ". *American Journal of Neuroradiology*, vol. 27, n° 5, p. 1032–1036.

Ogawa, Seiji, Tso-Ming Lee, Alan R Kay, and David W Tank. 1990. " Brain magnetic resonance imaging with contrast dependent on blood oxygenation ". *Proceedings of the National Academy of Sciences*, vol. 87, n° 24, p. 9868–9872.

Oishi, Kenichi, Kazi Akhter, Michelle Mielke, Can Ceritoglu, Jiangyang Zhang, Hangyi Jiang, Xin Li, Laurent Younes, Michael I Miller, Peter van Zijl, et al. 2011. " Multi-modal MRI analysis with disease-specific spatial filtering: initial testing to predict mild cognitive impairment patients who convert to Alzheimer's disease ". *Frontiers in neurology*, vol. 2, p. 54.

Olivetti, Emanuele, Giulia Berto, Pietro Gori, Nusrat Sharmin, and Paolo Avesani. 2017. " Comparison of distances for supervised segmentation of white matter tractography ". In *Pattern Recognition in Neuroimaging (PRNI), 2017 International Workshop on*. p. 1–4. IEEE.

Papalexakis, Evangelos E, Nicholas D Sidiropoulos, and Rasmus Bro. 2013. " From K-means to higher-way co-clustering: multilinear decomposition with sparse latent factors ". *Signal Processing, IEEE Transactions on*, vol. 61, n° 2, p. 493–506.

Pedersen, Mangor, Amir H Omidvarnia, Jennifer M Walz, and Graeme D Jackson. 2015. " Increased segregation of brain networks in focal epilepsy: an fMRI graph theory finding ". *NeuroImage: Clinical*, vol. 8, p. 536–542.

Pekalska, Elzbieta, Pavel Paclik, and Robert PW Duin. 2002. " A generalized kernel approach to dissimilarity-based classification ". *The Journal of Machine Learning Research*, vol. 2, p. 175–211.

Pekalska, Elzbieta, Artsiom Harol, Robert PW Duin, Barbara Spillmann, and Horst Bunke. 2006. Non-euclidean or non-metric measures can be informative. *Structural, Syntactic, and Statistical Pattern Recognition*, p. 871–880. Springer.

Pepe, Antonietta, Ivo Dinov, and Jussi Tohka. 2014. “ An automatic framework for quantitative validation of voxel based morphometry measures of anatomical brain asymmetry ”. *NeuroImage*, vol. 100, p. 444–459.

Peper, Jiska S, Rachel M Brouwer, Dorret I Boomsma, René S Kahn, Hulshoff Pol, and E Hilleke. 2007. “ Genetic influences on human brain structure: a review of brain imaging studies in twins ”. *Human brain mapping*, vol. 28, n° 6, p. 464–473.

Perrin, Muriel, Cyril Poupon, Yann Cointepas, Bernard Rieul, Narly Golestani, Christophe Pallier, Denis Rivière, André Constantinesco, Denis Le Bihan, and J-F Mangin. 2005. “ Fiber tracking in q-ball fields using regularized particle trajectories ”. In *Information Processing in Medical Imaging*. p. 52–63. Springer.

Pierpaoli, Carlo and Peter J Basser. 1996. “ Toward a quantitative assessment of diffusion anisotropy ”. *Magnetic resonance in Medicine*, vol. 36, n° 6, p. 893–906.

Pike, G Bruce. 2012. “ Quantitative functional MRI: concepts, issues and future challenges ”. *Neuroimage*, vol. 62, n° 2, p. 1234–1240.

Polderman, Tinca JC, Beben Benyamin, Christiaan A De Leeuw, Patrick F Sullivan, Arjen Van Bochoven, Peter M Visscher, and Danielle Posthuma. 2015. “ Meta-analysis of the heritability of human traits based on fifty years of twin studies ”. *Nature genetics*, vol. 47, n° 7, p. 702–709.

Poldrack, Russell A. 2012. “ The future of fMRI in cognitive neuroscience ”. *Neuroimage*, vol. 62, n° 2, p. 1216–1220.

Poldrack, Russell A, Timothy O Laumann, Oluwasanmi Koyejo, Brenda Gregory, Ashleigh Hover, Mei-Yen Chen, Krzysztof J Gorgolewski, Jeffrey Luci, Sung Jun Joo, Ryan L Boyd, et al. 2015. “ Long-term neural and physiological phenotyping of a single human ”. *Nature communications*, vol. 6, p. 8885.

Pooley, Robert A. 2005. “ Fundamental physics of MR imaging ”. *Radiographics*, vol. 25, n° 4, p. 1087–1099.

Powell, Michael A, Javier O Garcia, Fang-Cheng Yeh, Jean M Vettel, and Timothy Verstynen. 2018. “ Local connectome phenotypes predict social, health, and cognitive factors ”. *Network Neuroscience*, vol. 2, n° 1, p. 86–105.

Prasad, Gautam, Shantanu H Joshi, Neda Jahanshad, Julio Villalon-Reina, Iman Aganj, Christophe Lenglet, Guillermo Sapiro, Katie L McMahon, Greig I de Zubicaray, Nicholas G Martin, et al. 2014. “ Automatic clustering and population analysis of white matter tracts using maximum density paths ”. *Neuroimage*, vol. 97, p. 284–295.

Purcell, Edward M, H Co Torrey, and Robert V Pound. 1946. “ Resonance absorption by nuclear magnetic moments in a solid ”. *Physical review*, vol. 69, n° 1-2, p. 37.

Rademacher, J, U Bürgel, Stefan Geyer, T Schormann, A Schleicher, H-J Freund, and Karl Zilles. 2001. “Variability and asymmetry in the human precentral motor system”. *Brain*, vol. 124, n° 11, p. 2232–2258.

Ratnarajah, Nagulan, Andy Simmons, Alan Colchester, and Ali Hojjatoleslami. 2011. “Resolving complex fibre configurations using two-tensor random-walk stochastic algorithms”. In *SPIE Medical Imaging*. p. 79620R–79620R. International Society for Optics and Photonics.

Reitz, Christiane. 2016. “Toward precision medicine in Alzheimer’s disease”. *Annals of translational medicine*, vol. 4, n° 6.

Reuter, Martin, Franz-Erich Wolter, and Niklas Peinecke. 2006. “Laplace–Beltrami spectra as ‘Shape-DNA’ of surfaces and solids”. *Computer-Aided Design*, vol. 38, n° 4, p. 342–366.

Robinson, Emma C, Saad Jbabdi, Matthew F Glasser, Jesper Andersson, Gregory C Burgess, Michael P Harms, Stephen M Smith, David C Van Essen, and Mark Jenkinson. 2014. “MSM: a new flexible framework for Multimodal Surface Matching”. *Neuroimage*, vol. 100, p. 414–426.

Rogers, Baxter P, Victoria L Morgan, Allen T Newton, and John C Gore. 2007. “Assessing functional connectivity in the human brain by fMRI”. *Magnetic resonance imaging*, vol. 25, n° 10, p. 1347–1357.

Ros, Christian, Daniel Güllmar, Martin Stenzel, Hans-Joachim Mentzel, and Jürgen Rainer Reichenbach. 2013. “Atlas-guided cluster analysis of large tractography datasets”. *PloS one*, vol. 8, n° 12, p. e83847.

Rousseeuw, Peter J. 1987. “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”. *Journal of computational and applied mathematics*, vol. 20, p. 53–65.

Roweis, Sam T and Lawrence K Saul. 2000. “Nonlinear dimensionality reduction by locally linear embedding”. *science*, vol. 290, n° 5500, p. 2323–2326.

Rubinstein, R, M Zibulevsky, and M Elad. 2010. “Double sparsity: Learning sparse dictionaries for sparse signal approximation”. *Signal Processing, IEEE Transactions on*, vol. 58, n° 3, p. 1553–1564.

Ruiz-Blondet, Maria V, Zhanpeng Jin, and Sarah Laszlo. 2016. “CEREBRE: A Novel Method for Very High Accuracy Event-Related Potential Biometric Identification”. *IEEE Transactions on Information Forensics and Security*, vol. 11, n° 7, p. 1618–1629.

Rypma, Bart and Mark D’Esposito. 1999. “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*, vol. 96, n° 11, p. 6558–6563.

Schmitt, J Eric, Lisa T Eyler, Jay N Giedd, William S Kremen, Kenneth S Kendler, and Michael C Neale. 2007. “Review of twin and family studies on neuroanatomic phenotypes and typical neurodevelopment”. *Twin Research and Human Genetics*, vol. 10, n° 5, p. 683–694.

Schultz, Thomas and H-P Seidel. 2008. “Estimating crossing fibers: A tensor decomposition approach”. *Visualization and Computer Graphics, IEEE Transactions on*, vol. 14, n° 6, p. 1635–1642.

Seung, Sebastian, 2012. *Connectome: how the brain’s wiring makes us who we are*. Houghton Mifflin Harcourt.

Sharma, Sushil. 2017. “Translational Multimodality Neuroimaging”. *Current Drug Targets*, vol. 18, n° 9, p. 1039–1050.

Shen, Kai-Kai, Stephen Rose, Jurgen Fripp, Katie L McMahon, Greig I de Zubicaray, Nicholas G Martin, Paul M Thompson, Margaret J Wright, and Olivier Salvado. 2014. “Investigating brain connectivity heritability in a twin study using diffusion imaging data”. *NeuroImage*, vol. 100, p. 628–641.

Shen, Xilin, Emily S Finn, Dustin Scheinost, Monica D Rosenberg, Marvin M Chun, Xenophon Papademetris, and R Todd Constable. 2017. “Using connectome-based predictive modeling to predict individual behavior from brain connectivity”. *nature protocols*, vol. 12, n° 3, p. 506.

Siless, V, S Medina, G Varoquaux, and B Thirion. 2013. “A comparison of metrics and algorithms for fiber clustering”. In *Pattern Recognition in Neuroimaging (PRNI), 2013 International Workshop on*. p. 190–193. IEEE.

Siless, Viviana, Ken Chang, Bruce Fischl, and Anastasia Yendiki. 2018. “anatomicuts: Hierarchical clustering of tractography streamlines based on anatomical similarity”. *NeuroImage*, vol. 166, p. 32–45.

Smith, Stephen M, Mark Jenkinson, Heidi Johansen-Berg, Daniel Rueckert, Thomas E Nichols, Clare E Mackay, Kate E Watkins, Olga Ciccarelli, M Zaheer Cader, Paul M Matthews, et al. 2006. “Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data”. *Neuroimage*, vol. 31, n° 4, p. 1487–1505.

Smith, Stephen M, Karla L Miller, Steen Moeller, Junqian Xu, Edward J Auerbach, Mark W Woolrich, Christian F Beckmann, Mark Jenkinson, Jesper Andersson, Matthew F Glasser, et al. 2012. “Temporally-independent functional modes of spontaneous brain activity”. *Proceedings of the National Academy of Sciences*, vol. 109, n° 8, p. 3131–3136.

Smith, Stephen M, Christian F Beckmann, Jesper Andersson, Edward J Auerbach, Janine Bijsterbosch, Gwenaëlle Douaud, Eugene Duff, David A Feinberg, Ludovica Griffanti, Michael P Harms, et al. 2013a. “Resting-state fMRI in the human connectome project”. *Neuroimage*, vol. 80, p. 144–168.

Smith, Stephen M, Diego Vidaurre, Christian F Beckmann, Matthew F Glasser, Mark Jenkinson, Karla L Miller, Thomas E Nichols, Emma C Robinson, Gholamreza Salimi-Khorshidi, Mark W Woolrich, et al. 2013b. “Functional connectomics from resting-state fMRI”. *Trends in cognitive sciences*, vol. 17, n° 12, p. 666–682.

Smith, Stephen M, Thomas E Nichols, Diego Vidaurre, Anderson M Winkler, Timothy EJ Behrens, Matthew F Glasser, Kamil Ugurbil, Deanna M Barch, David C Van Essen, and Karla L Miller. 2015. “A positive-negative mode of population covariation links brain connectivity, demographics and behavior”. *Nature neuroscience*, vol. 18, n° 11, p. 1565.

Sotiropoulos, Stamatios N and Andrew Zalesky. 2017. “Building connectomes using diffusion MRI: Why, how and but”. *NMR in Biomedicine*.

Sotiropoulos, Stamatios N, Li Bai, Paul S Morgan, Cris S Constantinescu, and Christopher R Tench. 2010. “Brain tractography using Q-ball imaging and graph theory: Improved connectivities through fibre crossings via a model-based approach”. *NeuroImage*, vol. 49, n° 3, p. 2444–2456.

Sotiropoulos, Stamatios N, Saad Jbabdi, Junqian Xu, Jesper L Andersson, Steen Moeller, Edward J Auerbach, Matthew F Glasser, Moises Hernandez, Guillermo Sapiro, Mark Jenkinson, et al. 2013. “Advances in diffusion MRI acquisition and processing in the Human Connectome Project”. *Neuroimage*, vol. 80, p. 125–143.

Sprechmann, Pablo and Guillermo Sapiro. 2010. “Dictionary learning and sparse coding for unsupervised clustering”. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. p. 2042–2045. IEEE.

Standring, Susan, 2015. *Gray’s anatomy e-book: the anatomical basis of clinical practice*. Elsevier Health Sciences.

Sudlow, Cathie, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. 2015. “UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age”. *PLoS medicine*, vol. 12, n° 3, p. e1001779.

Sui, Jing, Tülay Adali, Qingbao Yu, Jiayu Chen, and Vince D Calhoun. 2012. “A review of multivariate methods for multimodal fusion of brain imaging data”. *Journal of neuroscience methods*, vol. 204, n° 1, p. 68–81.

Sui, Jing, Eduardo Castro, Hao He, David Bridwell, Yuhui Du, Godfrey D Pearlson, Tianzi Jiang, and Vince D Calhoun. 2014. “Combination of FMRI-SMRI-EEG data improves discrimination of schizophrenia patients by ensemble feature selection”. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*. p. 3889–3892. IEEE.

Sui, Jing, Godfrey D. Pearlson, Yuhui Du, Qingbao Yu, Thomas R. Jones, Jiayu Chen, Tianzi Jiang, Juan Bustillo, and Vince D. Calhoun. 2015. “In Search of Multimodal Neuroimaging Biomarkers of Cognitive Deficits in Schizophrenia”. *Biological Psychiatry*, vol. 78, n° 11, p. 794 - 804.

Sun, Jian, Huibin Li, Zongben Xu, et al. 2016. “Deep ADMM-Net for compressive sensing MRI”. In *Advances in Neural Information Processing Systems*. p. 10–18.

Symms, M, HR Jäger, K Schmierer, and TA Yousry. 2004. “A review of structural magnetic resonance neuroimaging”. *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 75, n° 9, p. 1235–1244.

Tenenbaum, Joshua B, Vin De Silva, and John C Langford. 2000. “A global geometric framework for nonlinear dimensionality reduction”. *science*, vol. 290, n° 5500, p. 2319–2323.

Thomas, Cibu, Q Ye Frank, M Okan Irfanoglu, Pooja Modi, Kad harbatcha S Saleem, David A Leopold, and Carlo Pierpaoli. 2014. “Anatomical accuracy of brain connections derived from diffusion MRI tractography is inherently limited”. *Proceedings of the National Academy of Sciences*, vol. 111, n° 46, p. 16574–16579.

Thomason, Moriah E and Paul M Thompson. 2011. “Diffusion imaging, white matter, and psychopathology”. *Clinical Psychology*, vol. 7, n° 1, p. 63.

Thompson, Paul M, Tyrone D Cannon, Katherine L Narr, Theo Van Erp, Veli-Pekka Poutanen, Matti Huttunen, Jouko Lönnqvist, Carl-Gustaf Standertskjöld-Nordenstam, Jaakko Kaprio, Mohammad Khaledy, et al. 2001. “Genetic influences on brain structure”. *Nature neuroscience*, vol. 4, n° 12, p. 1253–1258.

Thompson, Paul M, Tian Ge, David C Glahn, Neda Jahanshad, and Thomas E Nichols. 2013. “Genetics of the connectome”. *Neuroimage*, vol. 80, p. 475–488.

Thompson, Paul M, Jason L Stein, Sarah E Medland, Derrek P Hibar, Alejandro Arias Vasquez, Miguel E Renteria, Roberto Toro, Neda Jahanshad, Gunter Schumann, Barbara Franke, et al. 2014. “The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data”. *Brain imaging and behavior*, vol. 8, n° 2, p. 153–182.

Toews, M. and W. M. Wells. April 2016. “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)*. p. 847-850.

Toews, Matthew and William M Wells. 2013. “Efficient and robust model-to-image alignment using 3D scale-invariant features”. *Medical image analysis*, vol. 17, n° 3, p. 271–282.

Toews, Matthew, William Wells, D Louis Collins, and Tal Arbel. 2010. “Feature-based morphometry: Discovering group-related anatomical patterns”. *NeuroImage*, vol. 49, n° 3, p. 2318–2327.

Toews, Matthew, William M Wells, and Lilla Zöllei. 2012. “A feature-based developmental model of the infant brain in structural MRI”. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. p. 204–211. Springer.

Toews, Matthew, Christian Wachinger, Raul San Jose Estepar, and William M Wells III. 2015. “ A Feature-based Approach to Big Data Analysis of Medical Images ”. In *International Conference on Information Processing in Medical Imaging*. p. 339–350. Springer.

Toga, Arthur W and Paul M Thompson. 2003. “ Mapping brain asymmetry ”. *Nature reviews. Neuroscience*, vol. 4, n° 1, p. 37.

Tong, Tong, Robin Wolz, Pierrick Coupé, Joseph V Hajnal, Daniel Rueckert, Alzheimer’s Disease Neuroimaging Initiative, et al. 2013. “ Segmentation of MR images via discriminative dictionary learning and sparse coding: Application to hippocampus labeling ”. *NeuroImage*, vol. 76, p. 11–23.

Tsai, Andy, C-F Westin, Alfred O Hero, and Alan S Willsky. 2007. “ Fiber tract clustering on manifolds with dual rooted-graphs ”. In *Computer Vision and Pattern Recognition, 2007. CVPR’07. IEEE Conference on*. p. 1–6. IEEE.

Tsai, Chih-Fong. 2012. “ Bag-of-words representation in image annotation: A review ”. *ISRN Artificial Intelligence*, vol. 2012.

Tuch, David S. 2004. “ Q-ball imaging ”. *Magnetic resonance in medicine*, vol. 52, n° 6, p. 1358–1372.

Turpin, Andrew and Falk Scholer. 2006. “ User performance versus precision measures for simple search tasks ”. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. p. 11–18. ACM.

Tuytelaars, Tinne, Krystian Mikolajczyk, et al. 2008. “ Local invariant feature detectors: a survey ”. *Foundations and trends® in computer graphics and vision*, vol. 3, n° 3, p. 177–280.

Uğurbil, Kâmil. 2012. “ The road to functional imaging and ultrahigh fields ”. *Neuroimage*, vol. 62, n° 2, p. 726–735.

Uğurbil, Kamil, Junqian Xu, Edward J Auerbach, Steen Moeller, An T Vu, Julio M Duarte-Carvajalino, Christophe Lenglet, Xiaoping Wu, Sebastian Schmitter, Pierre Francois Van de Moortele, et al. 2013. “ Pushing spatial and temporal resolution for functional and diffusion MRI in the Human Connectome Project ”. *Neuroimage*, vol. 80, p. 80–104.

van den Heuvel, Martijn P, Cornelis J Stam, René S Kahn, and Hilleke E Hulshoff Pol. 2009. “ Efficiency of functional brain networks and intellectual performance ”. *Journal of Neuroscience*, vol. 29, n° 23, p. 7619–7624.

Van Essen, David C and Kamil Ugurbil. 2012. “ The future of the human connectome ”. *Neuroimage*, vol. 62, n° 2, p. 1299–1310.

Van Essen, David C, Kamil Ugurbil, E Auerbach, D Barch, TEJ Behrens, R Bucholz, A Chang, Liyong Chen, Maurizio Corbetta, Sandra W Curtiss, et al. 2012. “ The Human Connectome Project: a data acquisition perspective ”. *Neuroimage*, vol. 62, n° 4, p. 2222–2231.

Van Essen, David C, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, WU-Minn HCP Consortium, et al. 2013. “The WU-Minn human connectome project: an overview”. *Neuroimage*, vol. 80, p. 62–79.

Vanderwal, Tamara, Jeffrey Eilbott, Emily S Finn, R Cameron Craddock, Adam Turnbull, and F Xavier Castellanos. 2017. “Individual differences in functional connectivity during naturalistic viewing conditions”. *NeuroImage*, vol. 157, p. 521–530.

Varol, Erdem, Aristeidis Sotiras, Christos Davatzikos, Alzheimer’s Disease Neuroimaging Initiative, et al. 2017. “HYDRA: Revealing heterogeneity of imaging and genetic patterns through a multiple max-margin discriminative analysis framework”. *NeuroImage*, vol. 145, p. 346–364.

Vidal-Naquet, Michel and Shimon Ullman. 2003. “Object Recognition with Informative Features and Linear Classification.”. In *ICCV*. p. 281.

Visser, Eelke, Emil HJ Nijhuis, Jan K Buitelaar, and Marcel P Zwiers. 2011. “Partition-based mass clustering of tractography streamlines”. *Neuroimage*, vol. 54, n° 1, p. 303–312.

Vogel, Edward K, Andrew W McCollough, and Maro G Machizawa. 2005. “Neural measures reveal individual differences in controlling access to working memory”. *Nature*, vol. 438, n° 7067, p. 500.

Von Luxburg, Ulrike. 2007. “A tutorial on spectral clustering”. *Statistics and computing*, vol. 17, n° 4, p. 395–416.

Wachinger, Christian, Polina Golland, William Kremen, Bruce Fischl, Martin Reuter, Alzheimer’s Disease Neuroimaging Initiative, et al. 2015a. “BrainPrint: a discriminative characterization of brain morphology”. *NeuroImage*, vol. 109, p. 232–248.

Wachinger, Christian, Matthew Toews, Georg Langs, W Wells, and Polina Golland. 2015b. “Keypoint transfer segmentation”. In *International Conference on Information Processing in Medical Imaging*. p. 233–245. Springer.

Wachinger, Christian, David H Salat, Michael Weiner, Martin Reuter, and Alzheimer’s Disease Neuroimaging Initiative. 2016. “Whole-brain analysis reveals increased neuroanatomical asymmetries in dementia for hippocampus and amygdala”. *Brain*, vol. 139, n° 12, p. 3253–3266.

Waller, Lea, Henrik Walter, Johann D. Kruschwitz, Lucia Reuter, Sabine Müller, Susanne Erk, and Ilya M. Veer. 2017a. “Evaluating the replicability, specificity, and generalizability of connectome fingerprints”. *NeuroImage*, vol. 158, p. 371 - 377.

Waller, Lea, Henrik Walter, Johann D Kruschwitz, Lucia Reuter, Sabine Müller, Susanne Erk, and Ilya M Veer. 2017b. “Evaluating the replicability, specificity, and generalizability of connectome fingerprints”. *NeuroImage*, vol. 158, p. 371–377.

Wang, Danhong, Randy L Buckner, Michael D Fox, Daphne J Holt, Avram J Holmes, Sophia Stoecklein, Georg Langs, Ruiqi Pan, Tianyi Qian, Kuncheng Li, et al. 2015. “ Parcellating cortical functional networks in individuals ”. *Nature neuroscience*, vol. 18, n° 12, p. 1853.

Wang, Hua, Feiping Nie, Heng Huang, and Chris Ding. 2011a. “ Nonnegative matrix tri-factorization based high-order co-clustering and its fast implementation ”. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*. p. 774–783. IEEE.

Wang, Hua, Feiping Nie, Heng Huang, and Fillia Makedon. 2011b. “ Fast nonnegative matrix tri-factorization for large-scale data co-clustering ”. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Two*. p. 1553–1558. AAAI Press.

Wang, Qi, Mengying Sun, Liang Zhan, Paul Thompson, Shuiwang Ji, and Jiayu Zhou. 2017. “ Multi-Modality Disease Modeling via Collective Deep Matrix Factorization ”. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. p. 1155–1164. ACM.

Wang, Qian, Pew-Thian Yap, Guorong Wu, and Dinggang Shen. 2013. “ Application of neuroanatomical features to tractography clustering ”. *Human brain mapping*, vol. 34, n° 9, p. 2089–2102.

Wang, Xiaogang, W Eric L Grimson, and Carl-Fredrik Westin. 2011c. “ Tractography segmentation using a hierarchical Dirichlet processes mixture model ”. *NeuroImage*, vol. 54, n° 1, p. 290–302.

Wassermann, Demian, Maxime Descoteaux, and Rachid Deriche. 2008. “ Diffusion maps clustering for magnetic resonance q-ball imaging segmentation ”. *Journal of Biomedical Imaging*, vol. 2008, p. 5.

Wassermann, Demian, Luke Bloy, Efstathios Kanterakis, Ragini Verma, and Rachid Deriche. 2010. “ Unsupervised white matter fiber clustering and tract probability map generation: Applications of a Gaussian process framework for white matter fibers ”. *NeuroImage*, vol. 51, n° 1, p. 228–241.

Wassermann, Demian, Nikos Makris, Yogesh Rathi, Martha Shenton, Ron Kikinis, Marek Kubicki, and Carl-Fredrik Westin. 2016. “ The white matter query language: a novel approach for describing human white matter anatomy ”. *Brain Structure and Function*, vol. 221, n° 9, p. 4705–4721.

Wedeen, Van J, Patric Hagmann, Wen-Yih Isaac Tseng, Timothy G Reese, and Robert M Weisskoff. 2005. “ Mapping complex tissue architecture with diffusion spectrum magnetic resonance imaging ”. *Magnetic resonance in medicine*, vol. 54, n° 6, p. 1377–1386.

Wikibooks. 2017. “ Basic Physics of Nuclear Medicine MRI & Nuclear Medicine — Wikibooks, The Free Textbook Project ”. <https://en.wikibooks.org/w/index.php?title=

Basic_Physics_of_Nuclear_Medicine/MRI_26_Nuclear_Medicine&oldid=3225239>. [Online; accessed 9-April-2018].

Wikipedia. 2014. “Physics of Magnetic Resonance Imaging”. <http://en.wikipedia.org/wiki/Physics_of_magnetic_resonance_imaging>.

Wright, John, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma. 2009. “Robust face recognition via sparse representation”. *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, n° 2, p. 210–227.

Wright, John, Yi Ma, Julien Mairal, Guillermo Sapiro, Thomas S Huang, and Shuicheng Yan. 2010. “Sparse representation for computer vision and pattern recognition”. *Proceedings of the IEEE*, vol. 98, n° 6, p. 1031–1044.

Xia, Yan, U Turken, Susan L Whitfield-Gabrieli, and John D Gabrieli. 2005. Knowledge-based classification of neuronal fibers in entire brain. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2005*, p. 205–212. Springer.

Xu, Ting, Alexander Opitz, R Cameron Craddock, Margaret J Wright, Xi-Nian Zuo, and Michael P Milham. 2016. “Assessing variations in areal organization for the intrinsic brain: from fingerprints to reliability”. *Cerebral Cortex*, vol. 26, n° 11, p. 4192–4211.

Yan, Chaogan, Gaolang Gong, Jinhui Wang, Deyi Wang, Dongqiang Liu, Chaozhe Zhu, Zhang J Chen, Alan Evans, Yufeng Zang, and Yong He. 2011. “Sex-and brain size–related small-world structural cortical networks in young adults: a DTI tractography study”. *Cerebral cortex*, vol. 21, n° 2, p. 449–458.

Yang, Jianchao, Kai Yu, Yihong Gong, and Thomas Huang. 2009. “Linear spatial pyramid matching using sparse coding for image classification”. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. p. 1794–1801. IEEE.

Yang, Jun, Yu-Gang Jiang, Alexander G Hauptmann, and Chong-Wah Ngo. 2007. “Evaluating bag-of-visual-words representations in scene classification”. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*. p. 197–206. ACM.

Yarkoni, Tal. 2015. “Neurobiological substrates of personality: A critical overview”. *APA handbook of personality and social psychology*, vol. 4, p. 61–83.

Ye, Dong Hye, Jihun Hamm, Dongjin Kwon, Christos Davatzikos, and Kilian M Pohl. 2012. “Regional manifold learning for deformable registration of brain MR images”. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. p. 131–138. Springer.

Ye, Dong Hye, Benoit Desjardins, Jihun Hamm, Harold Litt, and Kilian M Pohl. 2014. “Regional manifold learning for disease classification”. *IEEE transactions on medical imaging*, vol. 33, n° 6, p. 1236–1247.

Yeatman, Jason D, Robert F Dougherty, Nathaniel J Myall, Brian A Wandell, and Heidi M Feldman. 2012. “Tract profiles of white matter properties: automating fiber-tract quantification”. *PloS one*, vol. 7, n° 11, p. e49790.

Yeh, Fang-Cheng and Wen-Yih Isaac Tseng. 2011. “NTU-90: a high angular resolution brain atlas constructed by q-space diffeomorphic reconstruction”. *Neuroimage*, vol. 58, n° 1, p. 91–99.

Yeh, Fang-Cheng, Van Jay Wedeen, and Wen-Yih Isaac Tseng. 2010. “Generalized-sampling imaging”. *IEEE transactions on medical imaging*, vol. 29, n° 9, p. 1626–1635.

Yeh, Fang-Cheng, Timothy D Verstynen, Yibao Wang, Juan C Fernández-Miranda, and Wen-Yih Isaac Tseng. 2013. “Deterministic diffusion fiber tracking improved by quantitative anisotropy”. *PloS one*, vol. 8, n° 11, p. e80713.

Yeh, Fang-Cheng, David Badre, and Timothy Verstynen. 2016a. “Connectometry: A statistical approach harnessing the analytical potential of the local connectome”. *NeuroImage*, vol. 125, p. 162–171.

Yeh, Fang-Cheng, Jean Vettel, Aarti Singh, Barnabas Poczos, Scott Grafton, Kirk Erickson, Wen-Yih Tseng, and Timothy Verstynen. 2016b. “Local Connectome Fingerprinting Reveals the Uniqueness of Individual White Matter Architecture”. *bioRxiv*, p. 043778.

Yeh, Fang-Cheng, Jean M Vettel, Aarti Singh, Barnabas Poczos, Scott T Grafton, Kirk I Erickson, Wen-Yih I Tseng, and Timothy D Verstynen. 2016c. “Quantifying Differences and Similarities in Whole-brain White Matter Architecture Using Local Connectome Fingerprints”. *PLOS Computational Biology*, vol. 12, n° 11, p. e1005203.

Zhang, Fan, Peter Savadjiev, Weidong Cai, Yang Song, Yogesh Rathi, Birkan Tunç, Drew Parker, Tina Kapur, Robert T Schultz, Nikos Makris, et al. 2017. “Whole brain white matter connectivity analysis using machine learning: an application to autism”. *NeuroImage*.

Zhang, Jun, Hao Ji, Ning Kang, and Ning Cao. 2005. “Fiber tractography in diffusion tensor magnetic resonance imaging: A survey and beyond”. In *Proceedings of the 2005 International Symposium on Medical Imaging and Computing*. p. 39–56.

Zhang, Wenlu, Rongjian Li, Houtao Deng, Li Wang, Weili Lin, Shuiwang Ji, and Dinggang Shen. 2015. “Deep convolutional neural networks for multi-modality isointense infant brain image segmentation”. *NeuroImage*, vol. 108, p. 214–224.

Zhang, Zhengwu, Maxime Descoteaux, Jingwen Zhang, Gabriel Girard, Maxime Chamberland, David Dunson, Anuj Srivastava, and Hongtu Zhu. 2018. “Mapping population-based structural connectomes”. *NeuroImage*, vol. 172, p. 130–145.

Zhu, Dajiang, Tuo Zhang, Xi Jiang, Xintao Hu, Hanbo Chen, Ning Yang, Jinglei Lv, Junwei Han, Lei Guo, and Tianming Liu. 2013. “Fusing DTI and fMRI data: A survey of methods and applications”. *NeuroImage*.

Zhu, Dajiang, Liang Zhan, Joshua Faskowitz, Madelaine Daianu, Neda Jahanshad, Greig I de Zubicaray, Katie L McMahon, Nicholas G Martin, Margaret J Wright, and Paul M Thompson. 2015. “Genetic analysis of structural brain connectivity using DICCCOL models of diffusion MRI in 522 twins”. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. p. 1167–1171. IEEE.

Zhu, Dajiang, Neda Jahanshad, Brandalyn C Riedel, Liang Zhan, Joshua Faskowitz, Gautam Prasad, and Paul M Thompson. 2016. “Population learning of structural connectivity by white matter encoding and decoding”. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. p. 554–558. IEEE.

Zilles, Karl and Katrin Amunts. 2013. “Individual variability is not noise”. *Trends in cognitive sciences*, vol. 17, n° 4, p. 153–155.

Zou, Zhang-Yu, Chang-Yun Liu, Chun-Hui Che, and Hua-Pin Huang. 2016. “Toward precision medicine in amyotrophic lateral sclerosis”. *Annals of translational medicine*, vol. 4, n° 2.

Zvitia, Orly, Arnaldo Mayer, Ran Shadmi, Shmuel Miron, and Hayit K Greenspan. 2010. “Co-registration of white matter tractographies by adaptive-mean-shift and Gaussian mixture modeling”. *Medical Imaging, IEEE Transactions on*, vol. 29, n° 1, p. 132–145.