UC Santa Barbara

UC Santa Barbara Electronic Theses and Dissertations

Title

Towards Interpretable Models of Health

Permalink

https://escholarship.org/uc/item/7z03m0f5

Author

Hallgrimsson, Harldur Tomas

Publication Date

2020

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

University of California
Santa Barbara

Towards Interpretable Models of Health

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Computer Science

by

Haraldur Tómas Hallgrímsson

Committee in charge:

Professor Ambuj K. Singh, Co-chair
Professor Scott T. Grafton, Co-char
Professor Subhash Suri
Professor Xifeng Yan

September 2020

The Dissertation of Haraldur Tómas Hallgrímsson is approved.

Professor Subhash Suri

Professor Xifeng Yan

Professor Scott T. Grafton, Committee Co-chair

Professor Ambuj K. Singh, Committee Co-Chair

September 2020

Towards Interpretable Models of Health

Copyright © 2020

by

Haraldur Tómas Hallgrímsson

For my parents Hallgrímur and Anna,
your love and support knows no bounds

Acknowledgements

My PhD would not have been possible without the help and support from many people. I would first like to thank my research advisor, Professor Ambuj Singh, for his guidance in asking the right questions. You have been been a great mentor, your feedback has sharpened my thinking, and you have afforded me incredible opportunities to grow as a researcher and person. I would also like to thank Professors Scott Grafton, Subhash Suri, and Xifeng Yan, who have served on my PhD committee and who have offered invaluable advice.

I would like to acknowledge the members of my research lab, the Dynamo Lab, for always providing a wall to bounce my ideas off of. They have been my support network, my collaborators, and—first and foremost—my friends. Thank you Arlei, Sourav, Wei, Omid, Zexi, Sean, Alex, Furkan, Mert, Sikun, Hongyuan, Aneesha, Koa, Yuning, Nikhil, Ashwini, Richika, Chandana, Victor, Xuan-Hong, Minh, Roman, Meredith, Daniel, Jason, Bo, and Petko. I have also been fortunate to have been affiliated with the Network Science IGERT lab, where there was always an opportunity for wide ranging conversations on all aspects of research and, more generally, life.

I am grateful to the great people I have collaborated with throughout my PhD. I thank Luca Foschini, who has been my mentor, colleague, and friend. I am grateful for his perspective on important research questions, unfailing attention to detail to truly ask the right questions, and the importance of enjoying life throughout it all. I had the privilege of working alongside him and the fantastic data scientists at Evidation Health during two summer internships. I would also like to thank Matt Cieslak, whose patience, kindness, and insights were invaluable when I first started at UC Santa Barbara. My gratitude also extends to my internship advisor Henry Rowley for valuable support and guidance.

More than anyone else, I would like to thank my family. Their love and support has never faltered. Thank you Hallgrímur, Anna, my sister Erla, and my brother Ragnar. I especially wish to thank my grandparents, Haraldur and Bubba, for always fiercely believing in me and supporting me; I will be forever grateful to them for teaching me to never stop.

Curriculum Vitæ

Haraldur Tómas Hallgrímsson

Education

2020	Ph.D. in Computer Science, University of California, Santa Barbara
2014	B.S. in Computer Science, University of Iceland
2013	B.S. in Electrical and Computer Engineering, University of Iceland

Professional Experience

06/2014-09/2020	Graduate Student Researcher, UC Santa Barbara
Summer 2018	Data Science intern, Evidation Health, Santa Barbara, CA
Summer 2017	Software Engineering intern, Google Research, Mountain View, CA
Summer 2016	Software Engineering intern, YouTube, Mountain View, CA
Summer 2015	Data Science intern, Evidation Health, Santa Barbara, CA
06/2013-06/2014	Research Engineer, Nox Medical, Reykjavík, Iceland

Publications

1. **Haraldur Hallgrímsson***, Richika Sharan*, Scott Grafton, Ambuj Singh, *"Estimating localized complexity of white-matter wiring with GANs"*, In Proc. of Medical Imgaging meets NeurIPS (December, 2019)
2. **Haraldur Hallgrímsson**, Filip Jankovic, Tim Althoff, Luca Foschini, *"Learning Individualized Cardiovascular Responses from Large-scale Wearable Sensors Data"*, In Proc. of Machine Learning for Health workshop at NeurIPS (December, 2018)
3. **Haraldur Hallgrímsson**, Matthew Cieslak, Luca Foschini, Scott T Grafton, Ambuj K Singh, *"Spatial coherence of oriented white matter microstructure: Applications to white matter regions associated with genetic similarity"*, NeuroImage (2018)
4. David Stück, **Haraldur Hallgrímsson**, Greg Ver Steeg, Alessandro Epasto, Luca Foschini, *"The spread of physical activity through social networks"*, In Proc. of 26th International Conference on World Wide Web (2017)

Working papers

5. **Haraldur Hallgrímsson**, Scott Grafton, Ambuj Singh, *"A generative model of brain structure quantifies informativeness of individual characteristics"* (2020)

6. **Haraldur Hallgrímsson**, Scott Grafton, Ambuj Singh, *"Discovering predictive regions of white matter with a learned population-level saliency map"* (2020)

Professional Service

Served as reviewer for a number of computer science conferences and journals, including: AAAI'15'18-20, WWW'16-19, KDD'17-20, IJCAI'19,20, ICDM'16'18'20, SDM'17, and TKDD

Teaching Experience

Teaching assistant for following courses:

- Data Structures and Algorithms, UCSB — Spring 2018
- Automata and Formal Languages, UCSB — Spring 2017
- Mathematics and Scientific Computing, University of Iceland — Spring 2014
- Formal Languages and Computability, University of Iceland — Fall 2013
- Mathematical Analysis 1, University of Iceland — Fall 2013
- Digital Circuit Design and Analysis, University of Iceland — Spring 2013
- Computer Science 1a (introduction to Matlab), University of Iceland — Fall 2012
- Computer Organization, University of Iceland — Fall 2012
- Electrical Measurements and Instruments, University of Iceland — Fall 2012
- Digital Circuit Design and Analysis, University of Iceland — Spring 2012

Abstract

Towards Interpretable Models of Health

by

Haraldur Tómas Hallgrímsson

We have witnessed massive advances in predictive modeling within the past decade, with machine learning models achieving superhuman performance in a variety of tasks. However, the notion that such models are a 'black-box' with little to no explanatory power has limited their impact on fields where erroneous data-driven decisions can have severe consequences, such as to do with our health. Health data in particular has the potential for transformational impact to our lives with the improved efficiency machine learning models have brought to other domains. There is a need for new computational approaches that can derive insight from health data while addressing these concerns.

In this dissertation, we describe novel computational methods on health data that do not just achieve high performance on singular performance metrics but chiefly to characterise the data. The methods combine insights from statistical and machine learning, network science, and Bayesian uncertainty quantification to improve our understanding of human health data through the lens of two main sources of data: the human brain as imaged by diffusion MRI, and human physiology as measured by wearable sensors. (1) We show how to find brain regions that are more cohesive within a population of interest, discovering that nearly 4% of white matter is associated with genetic similarity. (2) We quantify how informative an individual's attributes are for generative regions of white matter, and (3) also the dual problem of measuring how predictive a region of white matter is of an attribute. (4) Lastly, we demonstrate how to measure a cardiovascular transfer function from digital activity traces by learning from 'natural experiments'

during daily living conditions, and show that these transfer functions are predictive of variables associated with cardiovascular health.

0.1 Permissions and Attributions

The majority of the materials described in this dissertation have either been published by the author of the dissertation or are currently in the process of submission. The author has made principal contributions to all stages of the published works as described below.

1. The contents of Chapter 2, Appendix A and Appendix B have been previously published as:

   H. T. Hallgrímsson, M. Cieslak, L. Foschini, S. T. Grafton, and A. K. Singh, *Spatial coherence of oriented white matter microstructure: Applications to white matter regions associated with genetic similarity*, *NeuroImage* **172** (2018) 390–403

2. Portions of the contents of chapter 3 have been previously published as:

   H. T. Hallgrimsson, R. Sharan, S. T. Grafton, and A. K. Singh, *Estimating localized complexity of white-matter wiring with gans*, *Medical Imaging meets NeurIPS (MED-NeurIPS)* (2019)

   with large portions currently in submission.

3. The contents of chapter 4 are currently in submission.

4. The contents of chapter 5 have been previously published as:

   H. T. Hallgrímsson, F. Jankovic, T. Althoff, and L. Foschini, *Learning individualized cardiovascular responses from large-scale wearable sensors data*, *Machine Learning for Health workshop at NeurIPS* (2018)

Contents

Curriculum Vitae **vii**

Abstract **ix**
0.1 Permissions and Attributions xi

List of Figures **xiv**

List of Tables **xviii**

1 Introduction **1**

2 Discovering Regions of White Matter Associated with a Population **5**
2.1 Introduction 5
2.2 Methods 12
2.3 Results 22
2.4 Discussion 30
2.5 Conclusion 34

3 Characterizing White Matter with Generative Models **35**
3.1 Introduction 35
3.2 Imaging data and preprocessing 38
3.3 Learning localized wiring patterns 42
3.4 Model architecture 46
3.5 Results 48
3.6 Discussion 55
3.7 Conclusion 57

4 Characterizing White Matter with Predictive Models **59**
4.1 Introduction 59
4.2 Imaging data and preprocessing 62
4.3 Discovering population-wide salient regions 65
4.4 Results 68

4.5 Discussion 70
4.6 Conclusion 71

5 Learning Cardiovascular Health Signatures **73**
5.1 Introduction 73
5.2 Data 75
5.3 Cardiovascular Signature Network 76
5.4 Experimental results 78
5.5 Discussion 80

6 Conclusion **81**

A Voxel-wise Population Differences **84**

B Controlling for Morphological Similarity **89**

List of Figures

2.1 Two fascicles from two hypothetical groups of individuals (top row). These fascicles would generate very similar anisotropy images and tractograms (middle row). Coherent regions can be identified that agree across groups (bottom left, gray outlined) and that are dissimilar across groups (red outline in center). A sample of the MDA vectors of each population from the dissimilar region is shown on the bottom right. 9

2.2 Measuring coherence across voxels within a single subject. (a) A two-dimensional slice of the measured Orientation Distribution Function (ODF) from a single subject measuring the Brownian motion of water that is constrained by oriented white matter microstructure. (b) The multidirectional anisotrophy (MDA) values extracted from the local peaks of the ODFs (from pink box in *a*). (c) Measuring the coherence of neighboring voxels with respect to their ODFs by overlaying the extracted MDA vectors from the center voxel (highlighted in red in *b*) onto all spatially adjacent white matter voxels in this 2D slice. 10

2.3 Distribution of dissimilarity, or incoherence, between all adjacent white matter voxels within a single subject. Incoherence is mostly small but a long tail of dissimilar neighboring voxels exists. 11

2.4 Example distributions of similarities as computed from Eq. 2.5 between all twins and all strangers from two dyads, (2.4a) from a dyad that is significantly more similar within twins than strangers and (2.4b) from a dyad where no significant differences exist. For clarity, a non-parametric kernel density estimation has been overlaid. 19

2.5 Distribution of p-values for each dyad of neighboring white matter voxels assuming the null hypothesis in Eq. 2.8, that monozygotic and dizygotic twins are not more similar than strangers. 24

2.6 Axial slices of all 35.1k voxels (blue) and 19.3k (red) that were part of a neighboring voxel dyad found to be significantly more similar ($p < 10^{-4}$, FDR = 1.5%) ($p < 10^{-4}$, FDR = 3.2%) among monozygotic and dizygotic twins as compared to a control population of strangers using Eq. 2.5 with 1 and 2 peaks, respectively. Purple voxels are those that feature in the intersection of both, and form the vast majority of the extent of otherwise red voxels. Generalized Fractional Anisotrophy (GFA) template as background. Image created using ITK-SNAP (4). 25
2.7 The twenty-two largest white matter regions in which monozygotic and dizygotic twins are more similar than a control population of strangers, as visualized on a transparent background of a T1w volume. Image captured using Slicer 4 (5). 26
2.8 Distributions of region dissimilarities $d_{\mathcal{R}}(X, Y)$ per discovered white matter region of the monozygotic and dizygotic twin (left halves) and stranger populations (right halves). Quartiles of each distribution are shown as dashed lines. 27
2.9 Pearson correlation coefficients r of region dissimilarities $d_{\mathcal{R}}(X, Y)$ of each pair of monozygotic twins, dizygotic twins, siblings, and strangers, for each pair of regions. The percentage correlation is reported as a whole number (i.e., $100r$), with proportional shading added for clarity. 29
2.10 The distribution of subject pair dissimilarities $d(X, Y)$ as computed by Eq. 2.11. Black lines indicate each individual pair of subjects. 30

3.1 Three examples of synthesized brain structure. Two dimensional slices of white matter regions from real diffusion data (left column) and corresponding generated white matter microstructure (right), where the region being generated is delimited by the red square. The surrounding white matter outside the box is the contextual information provided in synthesizing the brain region. The red overlay displays the average aleatoric uncertainty within each generated voxel, where darker red corresponds to more uncertainty. The magnitude of the third dimension of each vector is shown as their blue-to-yellow color. 39
3.2 The complexity atlas of white matter structure aligns well with measures of anisotrophic signals within white matter. For dMRI scans, this measures the baseline difficulty of generating the white matter microstructure across the brain given near maximal contextual information. 49
3.3 Between subject variance of generative uncertainty is highest near junctions of large white matter bundles (as seen in left most figure) and near where white matter fans out towards cortical boundary. 51

3.4 The inherent complexity we measure of white matter voxels is positively correlated with the number of differentially trackable fiber bundles within them (Spearman's $\rho = 0.616$, $p < 10^{-10}$. The boxplot shows the quartiles of each distribution, with the box itself bounding the 25 and 75 percentiles and the midline at the median. The whiskers extend to three standard deviations around the mean. 53
3.5 Comparison with results from Volz et al. (6), segmenting our result by the number of trackable directions they define. Though overall there is significant agreement in results, we see considerable variance within voxels containing the same number of differentially directed fiber bundles. . . . 54

4.1 Simultaneously learning a patch-based classifier as well as the relative spatial saliency of each patch. During optimization, random patches of the MRI scan are sampled (left, within white boxes) and provided to the single patch-based classifier (middle). As not all possible patches are informative for the classification task, the relative performance of classifying different patches varies depending on a patch's spatial location. We propose learning a separate saliency function $s(i, j, k)$ which learns the relative informativeness of spatial locations (right), emphasising informative regions for both inference as well as backpropagation of gradients, while enabling an informative saliency map for the predictive task. 62
4.2 The learned saliency maps obtained from two independently trained models. A classification model considering only 40^3mm patches centered on these portions of the brain scan achieved 85.1% (top) and 78.7% (bottom) accuracy in predicting the biological sex of the 94 subjects in the held-out test set. 69

5.1 Example physical activity and sleep (upper row) and heart rate (bottom row) sensor data from three individuals, demonstrating how heart rate responds to onset of exercise (left column) and sleep (middle column). Changes in heart rate do not always occur due to physical activity (right column), with onset of anxiety or stress being potential unmeasured confounders. As expected, applying a signature from a different person (demonstrated in orange) results in increased reconstruction error. 75
5.2 Diagram of proposed model architecture. The signature encoder predicts a cardiovascular signature from measured sensor data (top dashed box), and the signature decoder uses that same signature as well as physical activity data to predict the heart rate (bottom dotted box). 77

A.1 Axial slices of voxels identified as significantly more similar within twins than strangers (yellow), and the subset of those voxels found to be significantly more similar within monozygotic compared to dizygotic twins (teal) and in siblings compared to strangers (orange), as computed using Eq. 2.3. Background image is of a population averaged Generalized Fractional Anisotrophy (GFA), where lighter regions indicate higher GFA values. Image created partly using ITK-SNAP (4). 87
A.2 Axial slices of voxels identified as significantly more similar within monozygotic and dizygotic twins than strangers as computed using Eq. A.1 (blue) and Eq. 2.5 (red). Background image is of a population averaged Generalized Fractional Anisotrophy (GFA), where lighter regions indicate higher GFA values. Image created partly using ITK-SNAP (4). 88

B.1 Axial slices of voxels in red whose log-jacobian values from the registration process are found to be significantly more similar within monozygotic and dizygotic twins than strangers, suggesting possible morphological similarity. Background image is a population averaged T1 weighted MRI image. Image created partly using ITK-SNAP (4). 91

List of Tables

2.1 Age and gender demographics of each pair of twins and siblings in the study population. 16
2.2 White matter regions discovered that are significantly more similar in monozygotic and dizygotic twins than in strangers. Effect size is Cohen's d as compared to the control population of strangers, or difference in means standardized by pooled standard deviations. 28

4.1 Specification of patch classifier model. 67
4.2 Specification of saliency model. 67

5.1 Experimental results. The trained proposed model was validated on the 2017 data, and also using 2017 signatures applied to 2018 data. While varying signature sizes (results shown in left column) the full training set was used, and when varying training set size (results shown in right column) a size-32 signature was used. 78
5.2 The performance of the baseline models trained on 2017 data and validated on 2018 data. 79

Chapter 1

Introduction

Tölva: Icelandic word for computer. Portmanteau of *tala* (number) and *völva* (prophetess): she who predicts numbers

---

Data is transforming science. The release of large, high-quality, and publicly available datasets have galvanized many fields of study, enabling insights that otherwise would not have been discovered which ultimately benefit our daily lives. Health data is no exception, with advances in our understanding there having the promise of transformational impact on human health and well-being.

Due to such data, massive advances have been made in predictive modeling. Superhuman performance has been achieved for many machine learning tasks, ranging from winning versus the best human players at the game of Go to diagnosing from medical imaging data with performance greater than that of teams of trained professionals. However, especially in fields such as health where erroneous decisions can have vast negative consequences, there has been trepidation in adopting these latest advances in machine

learning. Commonly, such models are rightfully criticised for not offering any insight into why they make one particular decision over another. This becomes problematic if, for instance, they are applied to new data which may not match the distribution they were trained on, and for which they may provide overly confidant predictions with no indication of trouble. These are exciting challenges, and call for clever methodologies in adapting to them.

In this dissertation, we present new computational methods to characterize health datasets. The aim is not to improve on singular prediction metrics but to gain actionable insight from the data, enabling better understanding of human health. To that end, we examine health data from the two extremes along the axis of acquisition cost: brain MRI scans and physical activity traces. Massive improvements in measurement fidelity within recent decades have led to much improved studies of the human brain in vivo, allowing researchers to further probe at our seat of consciousness. However, such data come with a great challenge: sample points are acquired in much smaller quantity relative to other data, and are of exceptionally high dimension due to the aforementioned improvements in sensory equipment and acquisition methods. At the other extreme are physical activity traces, acquired by ubiquitous digital devices including mobile phones and wearable devices. These can be collected continuously at massive scale, but with the associated challenge of dealing with vast heterogeneity between subjects and settings.

To effectively leverage these data, we have developed novel statistical and machine learning methods which deliver new insight as part of their design. Each of the following chapters present new methodology to understand either the physical connectivity within the human brain or digital traces of physical activity. In particular, this dissertation is organized as follows.

- In Chapter 2, we seek to discover regions of structural connectivity within the brain

that are preserved in a population of interest, as compared to some control population. We propose a novel measure of *coherence* between voxels within a brain scan, and develop a statistical framework which considers dyads of neighboring voxels across subjects to discover large connected components of the brain in which the oriented microstructure of the brain is more similar for the population of interest. We apply this methodology to discover nearly 4% of white matter is associated with genetic similarity based on a study of 109 twins, and show that these regions are preferentially located within deep white matter.

- In Chapter 3, we measure how individual traits of the subject affect local regions of white matter. We develop a generative framework based on Generative Adversarial Networks (GANs) which directly measures the uncertainty associated with generating regions of the brain given contextual information.

- In Chapter 4, we consider the dual problem of Chapter 3, namely discovering how predictive regions of white matter are of individual characteristics. We develop a method which decomposes a Convolutional Neural Network (CNN) classifier of a brain scan into two components: a classifier which predicts the characteristic under consideration on the basis of small patches of the brain scan, and a saliency map which denotes how informative each brain region is for the prediction. We learn this saliency map across a population, which offers an interpretable atlas of which brain regions are most correlated with which characteristics.

- In Chapter 5, we turn our attention to physical activity traces. Using fine-grained continuous measurements from a cohort of eighty thousand individuals over a span of a month, we develop methodology to infer a person's cardiovascular health. We leverage the long time-window of data to preferentially learn from spontaneously occurring 'natural experiments' where their cardiovascular response to activity can

be learned, and show that a descriptor of that response is predictive of variables associated with cardiovascular function, such as age and body mass index (BMI).

In each of these chapters we build upon advances in statistical learning, data and network science, and machine learning to characterize aspects of the datasets. The methods we propose are data-driven, and particularly adapted to meeting the variety of challenges and limitations inherent to the sensor data we apply them to. The constraints imposed by the data we consider lead to interesting design decisions for how best to model each problem, especially with the goal of deriving explanatory power from the models and frameworks we propose.

Chapter 2

Discovering Regions of White Matter Associated with a Population

> If a machine is expected to be infallible, it cannot also be intelligent.
>
> Alan Turing

2.1 Introduction

Structural connectomics of the human brain is increasingly recognized as an essential complement to functional imaging. Imaging the physical connectivity in the brain is primarily based on diffusion-weighted MRI (dMRI). While this imaging continues to improve in angular resolution of diffusion signals, there remain significant challenges in image reconstruction, representation of diffusion features, and statistical analysis of white matter structures across populations of subjects.

There is an abundance of methods for analyzing dMRI, many of which show promise in

diagnosing brain abnormalities such as strokes (7) and discovering correlates of many cognitive processes including metacognition (8). Current approaches generally fall into one of two categories: *Brain Graph* methods (9) use dMRI to estimate "connection strength" between pairs of cortical regions while *Scalar-based* methods calculate a single value at each voxel that is interpreted as reflecting "white matter integrity" (10).

Brain Graphs succinctly represent long-range connectivity between non-overlapping parcels of gray matter. The analyst chooses a gray matter parcellation, then uses a tractography algorithm to trace paths across white matter voxels. There are many approaches to tractography, but they all utilize diffusion orientation information to grow streamlines through space. Tractography results therefore depend on the accuracy of the voxel-wise estimates of white matter orientation, which can be complicated in structures such as crossing fibers (11) with (12) suggesting these tractography methods are readily dominated by false positive streamlines. Brain Graphs represent cortical regions as nodes and use a property of streamlines (such as their count) to weight edges, resulting in a connectivity matrix. These connectivity matrices are the basis for numerous network-based approaches (13) that have shown promise in understanding the development of large-scale brain connectivity (14) and processes such as aging, disease, and cognition (15; 16).

Scalar-based methods typically reduce the 6-dimensional dMRI data, a 3D oriented diffusion field measured in a 3D space, into a 3D volume. These scalar-valued volumes can easily be spatially normalized and statistically compared across individuals. The most common example of this is the analysis of Fractional Anisotropy (FA) derived from diffusion tensor imaging (DTI). FA is a function of the eigenvalues of a fitted diffusion tensor, with higher values reflecting a large degree of diffusion along a single orientation while lower values can reflect white matter damage (17; 18; 19; 20; 21; 22) or the presence of fiber populations projecting in multiple orientations (10; 23). The inability of tensors to represent multiple directions has been addressed by methods that use higher angular-

resolution dMRI to calculate an orientation distribution function (ODF) in each voxel where multiple fiber populations appear as "lobes" (24), as seen in Fig. 2.2a. Although ODFs can represent multiple fiber orientations, popular ODF-based scalars such as generalized fractional anisotropy (GFA) (25) and multidimensional anisotropy (MDA) (26) are still heavily reduced in voxels with fiber crossings (23). A major benefit to scalar-based techniques is that 3D interpolation can be performed accurately during spatial normalization, whereas interpolating 6D ODFs has been shown to systemically affect tractography (27); this normalization is important to be able to compare the same spatial regions of the brain between subjects that, in general, have different brain shapes and sizes.

Although the resampling of entire ODFs after applying a spatially-normalizing displacement field can produce undesirable results (28), 3D vector fields are generally well-behaved when spatially warped. We can take advantage of this by extracting directional maxima from each ODF and treating them as vectors. One vector is produced from each lobe of each ODF and warped to a group template where they can be compared across subjects. We calculate a similarity measure between each voxel and its neighbors instead of performing tractography on this spatially-normalized vector field. Where tractography seeks to determine whether axons *project into* a neighboring voxel, similarity scores reflect whether two voxels *are part of the same white matter structure*; this can be considered a generalization of tractography, capturing both the projections and cross-sections of a single white matter structure.

Fig. 2.1 shows how this approach compares to other current methods. Consider two fascicles in the brain that have been spatially normalized to overlap in space (top row). Two groups have different projections even though scalar based measures and tractography (middle row) would look identical. The bottom row shows the fascicles from both groups superimposed on one another. Distance measures between neighboring

voxels would reveal four areas that are *coherent* both between and across groups. In contrast, the vectors in the center crossing region are coherent within each group but differ across groups. The output of this pipeline is a set of regions like the red outlined area of crossings in Fig. 2.1, where directed ODF maxima are similar within groups but differ across groups.

Directional ODF maxima tend to vary smoothly in space albeit with large discontinuities around anatomical features, as seen in Fig. 2.2. We can measure the similarity of neighboring voxels by defining a distance between two ODFs that takes into account both magnitude and direction of each peak. Fig. 2.3 shows an example of *incoherence*, or dissimilarity, between ODFs from all dyads of neighboring white matter voxels within a single subject. Most dyads exhibit very low dissimilarity, with a long tail of voxel dyads with large dissimilarities. *Dyadic distances form the basis for the method proposed here. These distances are used to build a lattice network, which expands the comparison from neighboring voxels to large white matter regions. Region-based distances are then used to compare between groups.*

To demonstrate the validity and usefulness of the dyad approach, we consider the problem of finding spatially contiguous regions of white matter that are associated with a population of interest as compared to some control population. This problem mirrors the approach taken to subdivide the gray matter of the brain into functional regions (29; 30). We develop a non-parametric method for discovering arbitrarily shaped white matter regions that are significantly more similar with the population of interest, not on the basis of their connectivity to gray matter regions but instead on a group-wise local consistency in oriented white matter microstructure. This is accomplished by discovering spatially contiguous white matter voxels that are significantly more coherent within the population than would be expected from a matched control group. (Alternatively, especially in the context of neurological disorders and/or injuries, one could additionally search for regions

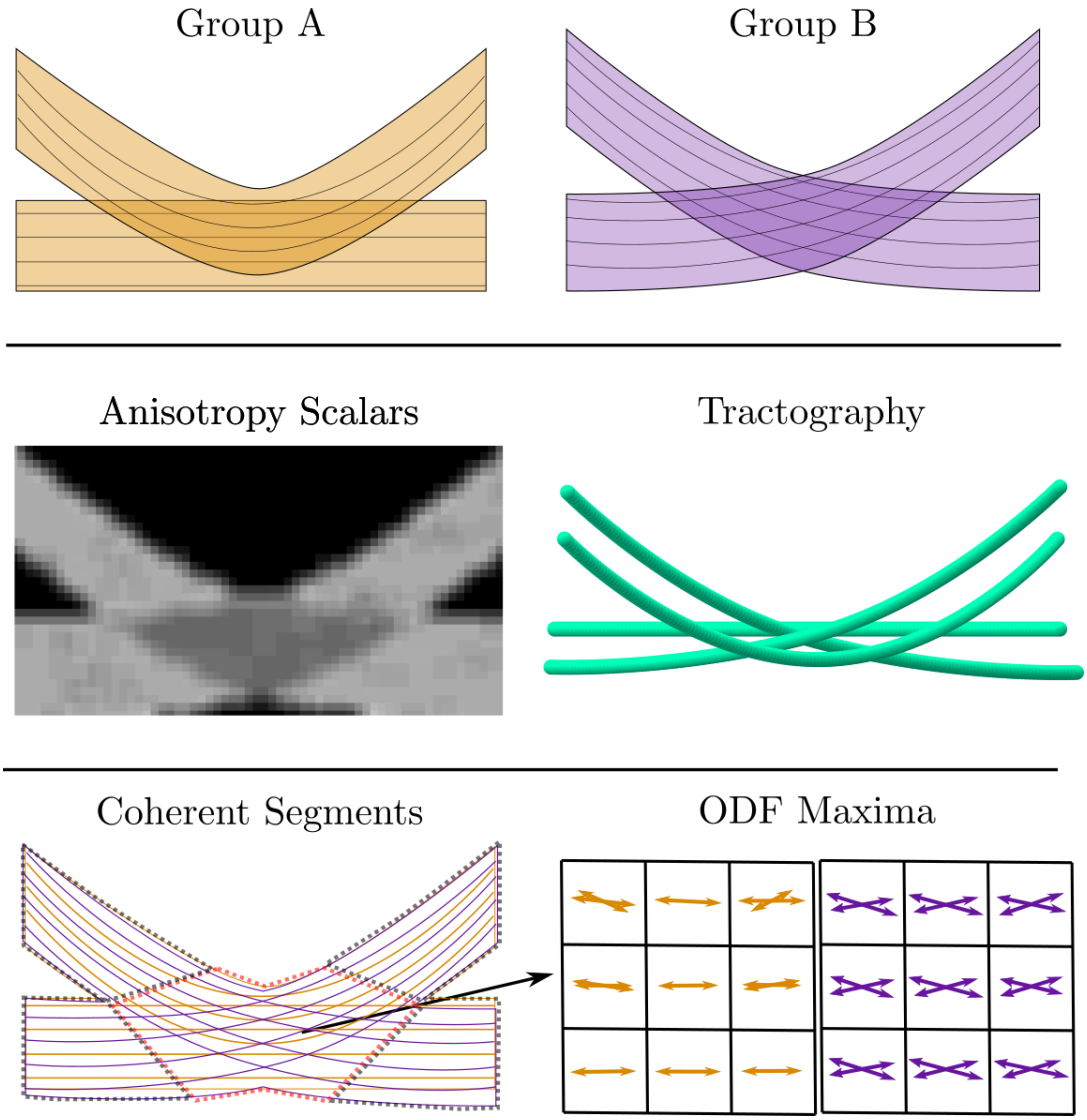


Figure 2.1: Two fascicles from two hypothetical groups of individuals (top row). These fascicles would generate very similar anisotropy images and tractograms (middle row). Coherent regions can be identified that agree across groups (bottom left, gray outlined) and that are dissimilar across groups (red outline in center). A sample of the MDA vectors of each population from the dissimilar region is shown on the bottom right.

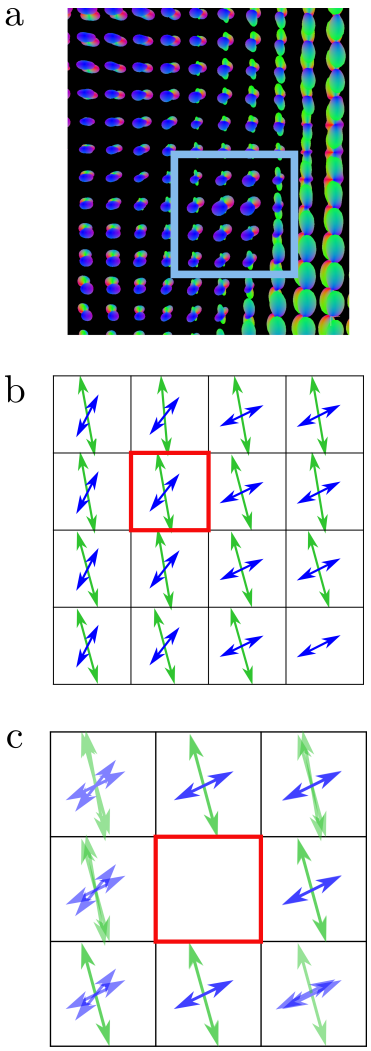


Figure 2.2: Measuring coherence across voxels within a single subject. (a) A two-dimensional slice of the measured Orientation Distribution Function (ODF) from a single subject measuring the Brownian motion of water that is constrained by oriented white matter microstructure. (b) The multidirectional anisotrophy (MDA) values extracted from the local peaks of the ODFs (from pink box in a). (c) Measuring the coherence of neighboring voxels with respect to their ODFs by overlaying the extracted MDA vectors from the center voxel (highlighted in red in b) onto all spatially adjacent white matter voxels in this 2D slice.

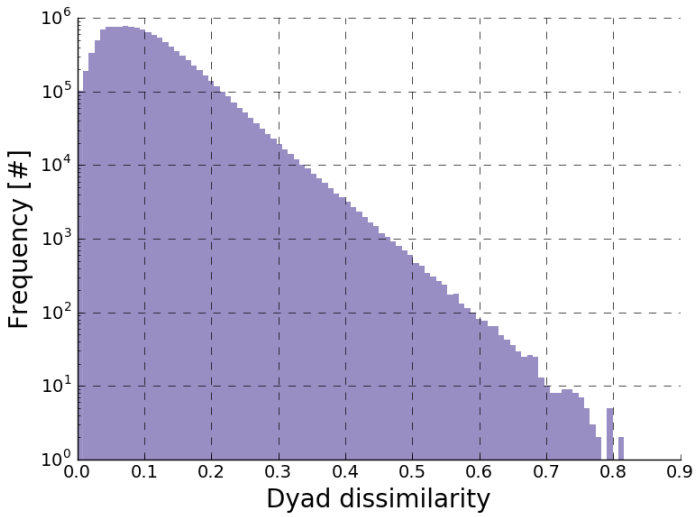


Figure 2.3: Distribution of dissimilarity, or incoherence, between all adjacent white matter voxels within a single subject. Incoherence is mostly small but a long tail of dissimilar neighboring voxels exists.

that are *less* coherent in the population of interest.) In contrast to previous studies (31), we measure coherence simultaneously across both subjects *and* neighboring voxels.

We apply this method to diffusion scans from the Human Connectome Project on a population of monozygotic (MZ) and dizygotic (DZ) twins to discover white matter regions that are associated with genetic similarity and/or a common upbringing. We hypothesize that in this situation, there should be significantly more coherence in the MZ twins than DZ twins, and DZ twins than unrelated individuals. The discovered regions are more similar within MZ and DZ twins than as compared to a control population of strangers. We also test the robustness of the discovered areas by generalizing them to a previously unseen population of non-twin siblings that display as much similarity in the white matter regions as the DZ twins.

Previous work has identified genetic influences of various quantitative measures of the brain, including total brain volume (32), the volume of gray and white matter regions (33; 34), peaks of fiber orientation functions and tracts derived therefrom (35),

brain asymmetries (36; 37), as well as other aspects of white matter (38; 39; 40). A simpler representation of the ODFs, FA, which measures how anisotropic a ODF is, has previously been shown to measure putative heritable influences (41). Our work extends that work by considering a richer representation of the ODFs. It is important to note that in our work we are interested in similarity associated with oriented white matter microstructure and not gross anatomical morphology, which we control for by excluding those white matter voxels whose log Jacobian determinant obtained during spatial normalization are more similar in MZ and DZ twins as opposed to strangers, see B.

If the population of interest does not correspond to any significant or strong signal of similarity within the oriented white matter microstructure, our method would identify only sparse and spatially distributed portions of white matter with an associated high false discovery rate. We apply our method on a population of MZ and DZ twins as we expect, and demonstrate, this population of interest as having a very strong signal to validate our method. That the results we present cluster spatially to a very high extent, are associated with a low false discovery rate and large effect sizes, and generalize to a previously unseen portion of the population serves as a strong validation of the method.

2.2 Methods

2.2.1 Imaging data and preprocessing

The preprocessing pipeline used for this study was identical to that used in (23), but is reported here as well for completeness. These data were collected as part of the Washington University-Minnesota Consortium Human Connectome Project (42; 43; 44). Participants were recruited from Washington University (St. Louis, MO) and surrounding area. All participants gave informed consent. The data is derived from 630 participants

(358 female, 272 male).

The structural and diffusion data were collected on 3T Connectome Skyra system (Siemens, Erlangen, Germany). The diffusion volumes were collected with a spatial resolution of 1.25x1.25x1.25 mm^3, using three shells at $b = 1000$, 2000, and 3000 s/mm^2 with 90 diffusion directions per shell and 10 additional b0s per shell. Spatial distortion and eddy currents were corrected using information from acquisitions in opposite phase-encoding directions, as well as head motion (45). The high-resolution structural T1 weighted and T2 weighted volumes were acquired on the same scanner at 0.7mm isotropic resolution. Minimally preprocessed images were reconstructed in DSI Studio (http://dsi-studio.labsolver.org) using Generalized Q-Sampling Imaging (46).

Skull stripped, aligned, and distortion corrected T1w and T2w volumes (45) were rigidly registered to the subject's GFA volume. The symmetric group wise normalization (SyGN) method implemented in Advanced Normalization Tools (ANTs, http://stnava.github.io/ANTs/) was used to construct a custom multimodal brain template using the data of 38 HCP subjects (47) that included proportions of racial, gender, and handedness that chosen through stratified random sampling according to these features. Of those 38, seven are monozygotic twins and nine are dizygotic twins that are a part of the population of interest for this study, and a further four are part of the non-twin siblings set. Each subject's GFA, T1w, and T2w volumes were used during template creation with weighting factors of 0.5 (GFA) $\times$ 1 (T1w) $\times$ 1 (T2w). Templates were created after 5 iterations. Templates from the 4th and 5th iterations of multi-modal template construction were inspected to check that the templates had stabilized. All individual datasets were ultimately normalized to this template using all 3 modalities and symmetric diffeomorphic normalization (SyN) as implemented in ANTs (48).

2.2.2 Extracting MDA Vectors

Each ODF $\Psi(\theta)$ was calculated with GQI on a set of 642 approximately-evenly spaced directions $\theta \in \Theta$ on a tesselated icosahedron. ODF magnitudes were rescaled so that the sum of each ODF is $\sum_{\theta\in\Theta} \Psi(\theta) = 1$. We then calculated the multi-directional anisotropy (MDA) value for each direction θ as

$$\mathrm{MDA}(\theta) = \frac{1-\mu}{\sqrt{1+2\mu^2}} \tag{2.1}$$

where

$$\mu = \left(\frac{\Psi(\theta)}{\Psi(\theta_{min})}\right)^{2/3} \tag{2.2}$$

and θ_{min} is the direction with the smallest ODF magnitude.

MDA values were calculated for the four largest local maxima in every ODF, resulting in values denoted MDA0, MDA1, MDA2, and MDA3 which are ordered by decreasing size. The four corresponding directions $\theta_0, \theta_1, \theta_2, \theta_3$ were also extracted and saved as 3D vector fields for each of $\theta_0, \ldots \theta_3$. In a separate study of the same data we found that ODF peaks become very noisy after the 4th direction (23). Vector fields corresponding to the local maxima were warped into the group template using ANTs. 3D volumes containing MDA0-3 were also warped to the group template and used to scale the normalized vectors. White matter voxels were determined by segmenting the weighted average template of T1w, T2w, and GFA volumes in FreeSurfer (49).

2.2.3 Estimating voxel expansion due to normalization

The 3D warps generated by ANTs were used to calculate the Jacobian matrix at each voxel. The log of the determinant of this matrix is an indication of whether the tissue in that voxel expanded or contracted in size in order to match the template images (50).

These values are commonly used to test for morphological differences between groups. We use these log-Jacobian values to dismiss any systematic morphological or misregistration effects that might affect this study (see B).

2.2.4 Subjects

The Human Connectome Project includes 109 pairs of twins, of which 57 are monozygotic (MZ) and 52 are dizygotic (DZ), and a further 47 pairs of non-twin siblings which are disjoint from the population of twins. The analysis in the rest of this paper is focused on the subjects in these three populations. Table 2.1 details the demographic information of these subjects.

The control groups for the twin and sibling populations are obtained by sampling with replacement an equal number of pairs of non-related subjects from the same population. We control for gender- and age-related confounders by matching gender and age-ranges such that the control populations have the same demographic distribution as detailed in Table 2.1. We refer to these control groups of non-related pairs of individuals as *strangers*.

2.2.5 Defining voxel-wise similarity

We seek to identify regions of white matter that contain significantly more similarly oriented white matter structures within a population of interest when compared to a control population. We first present a similarity metric between subjects defined with respect to a single voxel. We then extend that metric to measure similarity, or coherence, across *dyads of neighboring voxels* between pairs of subjects. Connected components of dyads that are significantly similar within the population of interest form the arbitrarily shaped regions of white matter associated with that population.

Table 2.1: Age and gender demographics of each pair of twins and siblings in the study population.

Monozygotic twins	22-25 22-25	22-25 26-30	22-25 31-35	26-30 26-30	26-30 31-35	31-35 31-35
Both female				24		19
Both male	3			7		4
Dizygotic twins						
Both female	1			17		13
Both male	5			9		7
Siblings						
Both female	1	3		1	2	2
Mixed gender	6	7		3	7	2
Both male	1	4	1	2	3	2

We define a similarity metric between a pair of subjects on the basis of their 6D ODFs within a voxel. We reduce the distributions down to the most probable underlying oriented white matter microstructure, which we describe as an ordered series of vectors, by extracting local peaks of the ODFs as MDA vectors. We order MDA vectors from the same voxel by their magnitude normalized by the isotrophic component of the distribution. We extract the four largest peaks in any given MDA distribution, as detailed in Section 2.2.2.

We measure how similar a pair of individuals are with respect to their oriented white matter microstructure within a voxel by comparing their extracted MDA vectors. The similarity should take into account similarity in both direction and magnitude. Common methods to compare vector similarity that incorporate both magnitude and direction include the dot product as well as various p-norms of the vector differences. We use the L^2 norm, or Euclidean distance, to compare individual vectors. This choice of metric corresponds well to the geometric space the measured microstructure exists in, as well as

is robust to noisy MDA vectors which manifest as having small magnitudes.

Let X_v^i be the i-th three-dimensional vector in voxel v for subject X. This is a directed vector representation of an ODF, which fundamentally is not directed (as seen in Fig. 2.2b). As such, when computing a dissimilarity between ODFs we consider the minimum distance between the vector representation of an ODF *or its reflection around the origin* to another such vector representation *or that vector's reflection around the origin*, where the origin is the center of a given ODF. We compute subject X's dissimilarity to subject Y in voxel v as

$$d(X, Y, v) = \sum_{i=1}^{k} \min\left(\left\|X_v^i - Y_v^i\right\|, \left\|X_v^i + Y_v^i\right\|\right), \tag{2.3}$$

where we have k MDA vectors and use the L^2 norm in $d = 3$ dimensions,

$$\|x\| = \sqrt{\sum_{k=1}^{d} |x_k|^2}. \tag{2.4}$$

2.2.6 Defining voxel dyad similarity

As we report in A, Eq. 2.3 is a suitable method to discover white matter voxels that are significantly more similar in a population of interest when compared to a control population, and that these voxels spatially cluster into white matter regions. However, we derive a related method that directly encodes the notion of spatially adjacent voxels and serves as a more natural way to discover white matter regions.

Individually significant voxel dyads can be aggregated to form large arbitrarly shaped white matter regions. To this end, we define a lattice network over the white matter voxels and search for subnetworks, or white matter regions, that exhibit significant similarity. Each white matter voxel serves as a node in this network, and we consider dyads of neighboring voxels (those that share a common face, edge, or corner, i.e. each voxel may

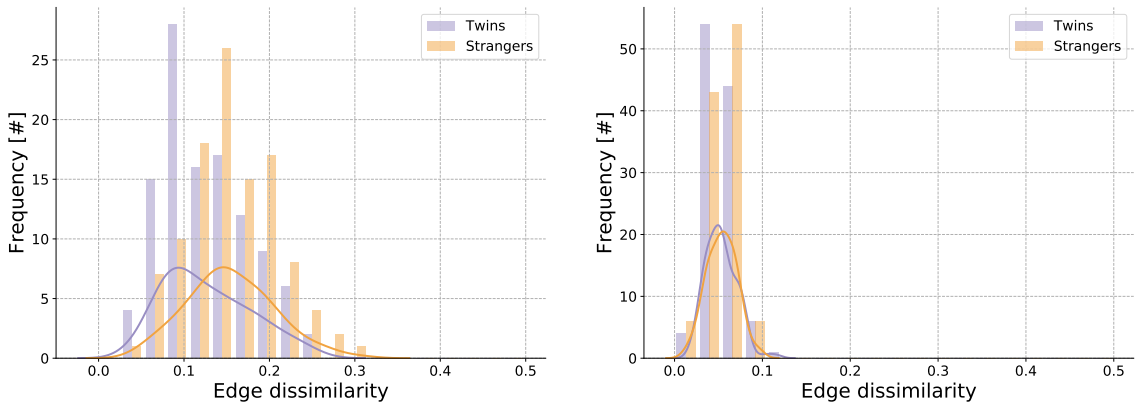
have up to 26 spatially adjacent white matter voxels that form a cube surrounding the center voxel). This forms a lattice network of the white matter voxels connecting the nearly one million white matter voxels together as 12.2 million voxel dyads. We again exclude those voxels which we have evidence for being more similarly registered in the population of interest, as reported in B.

Using this network approach, the random variable of interest no longer corresponds to a single voxel but instead to dyads of neighboring voxels. We modify Eq. 2.3 such that subjects X and Y's dissimilarity with respect to the undirected voxel dyad (u, v) is

$$\begin{aligned} d(X, Y, u, v) = \frac{1}{2} \sum_{i=1}^{k} \Big(& \min \left(\left\| X_u^i - Y_v^i \right\|, \left\| X_u^i + Y_v^i \right\| \right) \\ & + \min \left(\left\| X_v^i - Y_u^i \right\|, \left\| X_v^i + Y_u^i \right\| \right) \Big), \end{aligned} \tag{2.5}$$

where we take the arithmetic mean between the directed pairs (u, v) and (v, u) such that the dissimilarity is symmetric between subjects (and reordering of the dyad) as is Eq. 2.3.

This dissimilarity can be considered as the *incoherence* between neighboring voxels in a pair of subjects. A low dissimilarity implies that the fiber populations in the two voxels across subjects contain similarly oriented white matter structures, though not necessarily that there exists a fiber population that travels between the two voxels. A high dissimilarity could correspond to perpendicular fiber populations, or large deviations in the measured magnitude of the MDA peaks. An example of this dissimilarity, or incoherence, can be seen in Fig. 2.2(c).



(a) Voxel dyad in which twins are significantly more similar than strangers. Kernel density estimates overlaid as solid lines.

(b) Voxel dyad in which twins are *not* significantly more similar than strangers. Kernel density estimates overlaid as solid lines.

Figure 2.4: Example distributions of similarities as computed from Eq. 2.5 between all twins and all strangers from two dyads, (2.4a) from a dyad that is significantly more similar within twins than strangers and (2.4b) from a dyad where no significant differences exist. For clarity, a non-parametric kernel density estimation has been overlaid.

2.2.7 Population differences and significance

For each dyad $e = (u, v)$ of spatially adjacent voxels u and v in the white matter lattice network we obtain a sample of the distribution of dissimilarities in the population under consideration, $e_1^T, e_2^T, \ldots, e_n^T$, and for the control population, $e_1^S, e_2^S, \ldots, e_n^S$, empirically:

$$e_i^T = d(T_{i,1}, T_{i,2}, u, v), \tag{2.6}$$

$$e_i^S = d(S_{i,1}, S_{i,2}, u, v). \tag{2.7}$$

Where T_i and S_i, $1 \leq i \leq n$, are the i-th pair of subjects from the population under consideration and control population, respectively.

We seek to identify those dyads in which the pairs of subjects from the population of interest are significantly more similar than that in the control population. As the

distributions of dissimilarities e^T and e^S are non-normal, instead of a t-test we employ a Mann-Whitney U test (51) to test for differences in the two distributions. Having visually verified that the same shape assumption holds, the Mann-Whitney U test is a non-parametric rank test of the null hypothesis that the two samples of dissimilarities from edge e are equally likely to be as large,

$$P(v^T < v^S) = P(v^T > v^S), \tag{2.8}$$

against the one-sided alternative hypothesis that the population of interest tends to have lower dissimilarities,

$$P(v^T < v^S) > P(v^T > v^S); \tag{2.9}$$

that is that the population of interest tends to be more similar. For the dyads in which we reject the null hypothesis, we have evidence that the population of interest is more coherent, or similar. Fig. 2.4 shows an example distribution of coherence from dyads that are and are not significantly similar among a population of interest (twins) as compared to a control group (strangers).

Alternatively, for populations of interest for which the analyst hypothesizes should have common *less coherent* regions—such as populations of subjects with neurological disorders or injuries—the analyst might instead test against a one-sided alternative hypothesis that the population of interest tends to have higher dissimilarities. In either case, care must be taken to not aggregate together dyads using a two-sided hypothesis, or dyads from different one-sided hypotheses, as a region formed by such aggregated dyads does not form a single unified region of interest.

To account for multiple hypothesis across all neighboring voxels in white matter, we estimate the false-discovery rate given a particular p-value threshold (52; 53). Fig. 2.5

shows a flat baseline of p-values for this null hypothesis with a sharp peak as p approaches zero indicating high statistical power of the test being employed.

2.2.8 Defining regions and region-wise similarity

Of the set of neighboring white matter voxels for which we reject the null hypothesis and that have been corrected for multiple hypothesis, we further prune unlikely spatially isolated dyads. This is accomplished by aggregating together dyads that form connected components in the white matter lattice network and keeping only the largest such components. These form arbitrarily shaped disjoint *regions* of white matter that can each be considered as single units of interest for further analysis.

For a pair of subjects X and Y and a white matter region $\mathcal{R}$ (a set of white matter voxel dyads) we defined a dissimilarity between the subjects with respect to $\mathcal{R}$ as the median dissimilarity of all dyads in $\mathcal{R}$ using Eq. 2.5,

$$d_{\mathcal{R}}(X, Y) = \text{median}\left(\{d(X, Y, u, v), \forall (u, v) \in \mathcal{R}\}\right). \tag{2.10}$$

2.2.9 Defining between-subject similarity

We define a single dissimilarity measure between a pair of subjects X and Y on the basis of multiple white matter regions as the mean region similarity across each of the white matter regions $\mathcal{R}_0, \mathcal{R}_1, \ldots, \mathcal{R}_{R-1}$,

$$d(X, Y) = \frac{1}{R} \sum_{i=0}^{R-1} d_{\mathcal{R}_i}(X, Y). \tag{2.11}$$

2.3 Results

For each dyad of neighboring white matter voxels, we computed the incoherence using Eq. 2.5 with $k = 1$ MDA peak across each pair of subjects in the monozygotic (MZ) twin, dizygotic (DZ) twin, and matched stranger populations. We found those white matter dyads for which we have enough evidence to rule out the null hypothesis described by Eq. 2.8 in favor of the alternative hypothesis given by Eq. 2.9, where we consider the population of interest all pairs of MZ and DZ twins (and *not* the non-twin sibling pairs). We then examined the largest connected subnetworks and their properties.

We control for similarity due to the morphology of the brain that would otherwise confound this analysis by excluding voxels which can be shown to have been similarly morphed into the normal space in twins but not in strangers (see B).

2.3.1 Twin-similar white-matter regions

Of the 12.2 million dyads of spatially adjacent white matter voxels, we identified 71,857 as significantly more similar within MZ and DZ twins as compared to a matched control group of strangers ($p < 10^{-4}$, false discovery rate 1.5%), see Fig. 2.5. These dyads contained 35,119 unique white matter voxels, as seen in Fig. 2.6. The dyads form 3,145 connected components in the white matter lattice network, of which 1,791—more than half—were trivial subnetworks of a single dyad of two voxels. More interestingly, twenty-nine subnetworks connected more than one hundred voxels, or a volume of white matter that is approximately 200mm^3 in normalized template space. We selected the twenty-two largest subnetworks as units for further analysis as these comprise 75% of all significant voxel dyads. See table 2.2 for relevant statistics of these twenty-two largest white matter regions, and Fig. 2.7 for a visualization of these twenty-two white matter regions.

Increasing the number of peaks from $k = 1$ in Eq. 2.5 steadily decreases the size and significance of the results. With $k = 2$ peaks, we observe less than half the significant voxel dyads or 35,268 dyads containing 19,253 unique voxels ($p < 10^{-4}$, FDR 3.2%) and which can be seen to be nearly fully encompassed by the results with $k = 1$ in Fig. 2.6. The largest cluster of voxels identified using two peaks and not one is in the centrum semiovale, which has previously been identified as an area containing multiple crossing fibers (23). Further increasing to $k = 3$ peaks reduces the significant dyads to 30,134 containing 15,623 unique voxels ($p < 10^{-4}$, FDR 4.0%). This decrease in the size and significance of the results can be attributed to the noise associated with the higher MDA peaks; 32.9% of voxels are known to be singly connected such that in these the local MDA peaks past the first contain no true signal (23). Unless otherwise stated, all further results are presented with $k = 1$ MDA peaks.

Furthermore, significant voxel dyads are biased towards non-diagonal connections between voxels. Forming a lattice network only between voxels which share either a face or edge, allowing up to eighteen neighbors per voxel, and retesting the null hypothesis we discover 58,159 voxel dyads containing 32,986 unique voxels to be significant ($p < 10^{-4}$, FDR 1.2%). As this reduced lattice network has only 69% (roughly 18/26) of the dyads present as compared to the original, if no bias towards non-diagonal connections existed we would expect 69% of the original dyads, or 49.6k, to remain significant. This is much less than what we observe. To some degree, this is not surprising as the diagonal dyads should be expected to be less spatially coherent due to the greater distance between their centers and the smaller effect of spatial smoothing inherent to the imaging process. The aggregate white matter regions remain fairly invariant to the choice of lattice network, with the total number of regions decreasing from 5,699 of which 29 contain at least 100 voxels to 5,310 of which at least 27 contain at least 100 voxels when going from a 26- to 18-connected lattice.

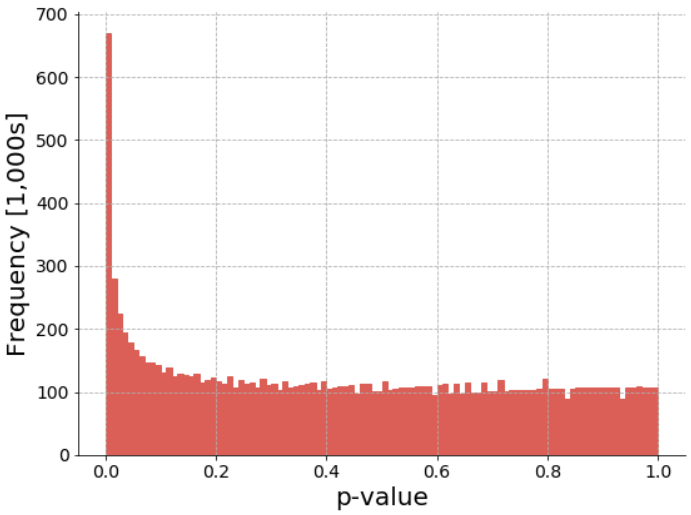
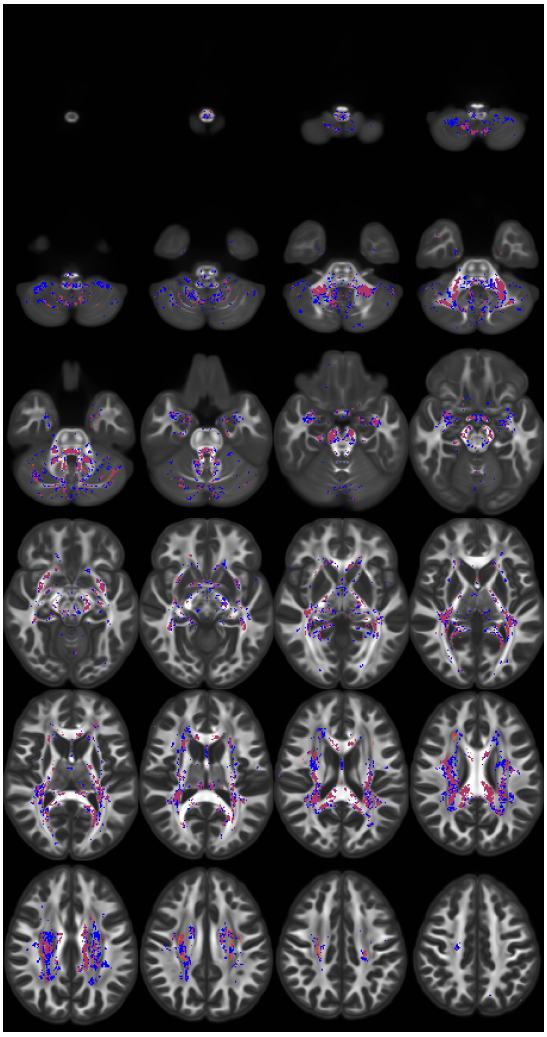


Figure 2.5: Distribution of p-values for each dyad of neighboring white matter voxels assuming the null hypothesis in Eq. 2.8, that monozygotic and dizygotic twins are not more similar than strangers.

2.3.2 Effect size of white matter regions

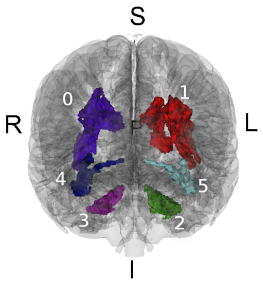
Having aggregated together individually significant voxel dyads to form large arbitrarily shaped white matter regions, we measure similarity between pairs of subjects on the basis of a single region using Eq. 2.10. We show that this region-wise similarity measure corresponds to a large effect size when comparing pairs of MZ and DZ twins to a control group of strangers, and that this measure generalizes to a previously unseen group of siblings. The sibling data was not considered previous to this point with one exception: four siblings were among the 38 subjects sampled from all HCP scans to define a normal template.

In each of the twenty-two largest discovered regions, the distribution of such region dissimilarity in the twin population and the stranger population (as seen in Fig. 2.8) is, although overlapping, reasonably well separable. Table 2.2 shows the aggregate statistics of the regions. Larger regions correlate with larger effect sizes (as measured by the Cohen's d statistic), which is to be expected as they are composed of a greater amount of

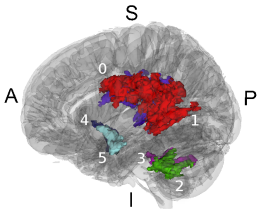


q

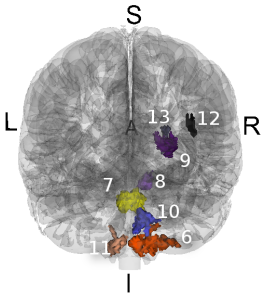
Figure 2.6: Axial slices of all 35.1k voxels (blue) and 19.3k (red) that were part of a neighboring voxel dyad found to be significantly more similar ($p < 10^{-4}$, FDR = 1.5%) ($p < 10^{-4}$, FDR = 3.2%) among monozygotic and dizygotic twins as compared to a control population of strangers using Eq. 2.5 with 1 and 2 peaks, respectively. Purple voxels are those that feature in the intersection of both, and form the vast majority of the extent of otherwise red voxels. Generalized Fractional Anisotrophy (GFA) template as background. Image created using ITK-SNAP (4).



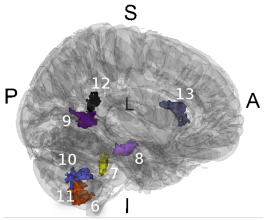
(a) Anterior view of the six largest regions.



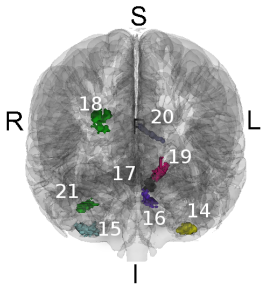
(b) Left view of the six largest regions.



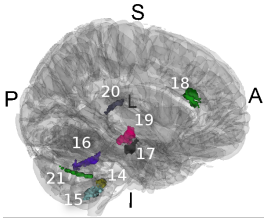
(c) Posterior view of the seventh to 14th largest regions.



(d) Right view of the seventh to 14th largest regions.



(e) Anterior view of the 15th to 22nd largest regions.



(f) Right view of the 15th to 22nd largest regions.

Figure 2.7: The twenty-two largest white matter regions in which monozygotic and dizygotic twins are more similar than a control population of strangers, as visualized on a transparent background of a T1w volume. Image captured using Slicer 4 (5).

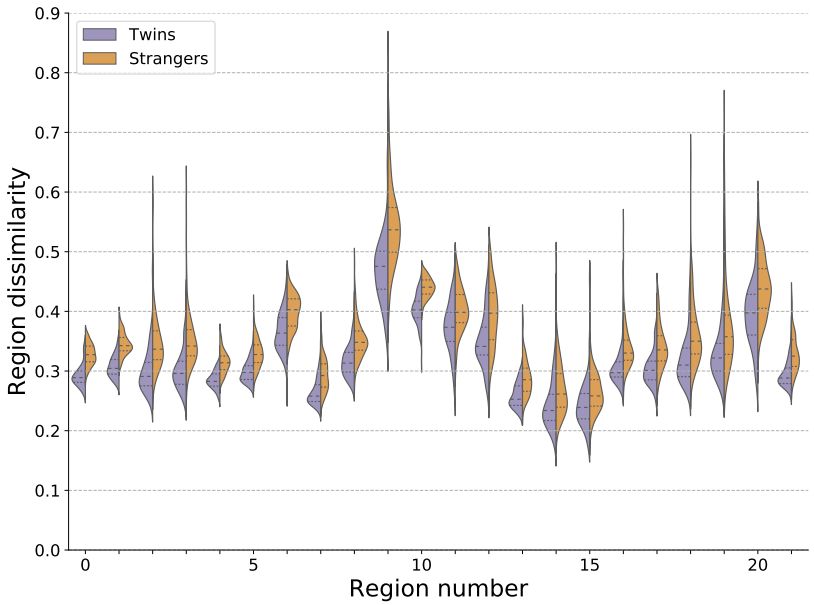


Figure 2.8: Distributions of region dissimilarities $d_{\mathcal{R}}(X, Y)$ per discovered white matter region of the monozygotic and dizygotic twin (left halves) and stranger populations (right halves). Quartiles of each distribution are shown as dashed lines.

significant voxel dyads, though all regions are associated with very large effect sizes in the range of 0.6 to 2.1 pooled standard deviations. We note that the effect sizes generalize to a population of siblings, which were not used in identifying the regions, showing that this measure of similarity between MZ and DZ twins generalizes to similarity among siblings though with medium effect sizes.

Though these white matter regions all exhibit a large difference in the distributions of MZ and DZ twins and strangers, and to a lesser extent also between siblings and strangers, they do so somewhat independently. We compute the Pearson correlation coefficient between each pair of regions with respect to the measured region similarity for each pair of monozygotic and dizygotic twins as well as strangers, as seen in Fig. 2.9. We see that larger regions tend to correlate more, which is unsurprising as they discriminate between the groups better. Similarly, regions 14 and 15—which have the lowest effect sizes—correlate with each other to a greater extent than with all other regions. There is

Table 2.2: White matter regions discovered that are significantly more similar in monozygotic and dizygotic twins than in strangers. Effect size is Cohen's d as compared to the control population of strangers, or difference in means standardized by pooled standard deviations.

Region number	**Number of dyads**	**Number of voxels**	**Twin effect size [STD]**	**Sibling effect size [STD]**
0	17,336	5,699	1.94	0.84
1	14,382	5,369	2.06	0.92
2	5,458	1,311	1.15	0.33
3	3,793	1,146	1.20	0.34
4	1,689	745	1.74	0.82
5	1,374	599	1.46	0.77
6	1,322	594	1.09	0.35
7	1,218	431	1.13	0.66
8	896	306	1.24	0.55
9	874	308	1.11	0.55
10	750	408	1.78	0.78
11	579	222	0.77	0.25
12	576	142	0.86	0.32
13	516	255	1.16	0.57
14	476	156	0.58	0.39
15	420	220	0.60	-0.13
16	368	173	1.02	0.29
17	340	121	1.11	0.56
18	334	198	0.85	0.02
19	310	184	0.63	0.29
20	308	110	0.88	0.45
21	302	146	1.27	0.43

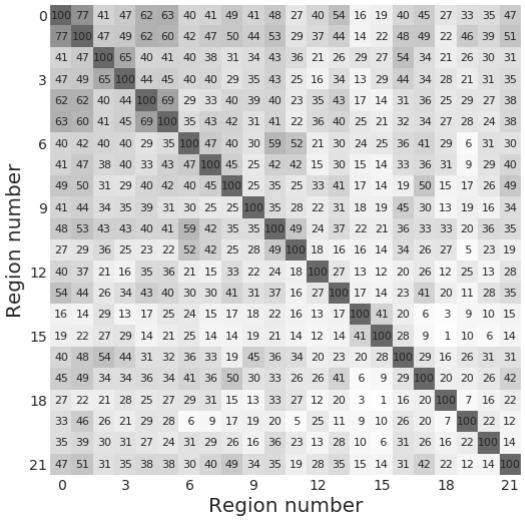


Figure 2.9: Pearson correlation coefficients r of region dissimilarities $d_{\mathcal{R}}(X, Y)$ of each pair of monozygotic twins, dizygotic twins, siblings, and strangers, for each pair of regions. The percentage correlation is reported as a whole number (i.e., $100r$), with proportional shading added for clarity.

weak evidence for other such clusters of white matter regions that predict similarly for each pair of subjects. In particular, region pairs that appear mirrored across hemispheres correlate with each other more than they do other regions, such as 4 and 5 or 14 and 15.

2.3.3 Subject similarity

Using Eq. 2.11 we measure a single dissimilarity between each pair of MZ twins, DZ twins, non-twin siblings, and strangers on the basis of the $R = 22$ discovered white matter regions. The distribution of dissimilarities for each of these groups can be seen in Fig. 2.10. The modes of the distributions define a clear order from least to most genetic similarity with strangers having high dissimilarities and MZ twins very low, with DZ twins situated in-between. We see this measure generalizes to the population of non-twin siblings, whose data were not used in obtaining the regions, and which have comparable dissimilarities to DZ twins though with a longer tail of high dissimilarities. The majority

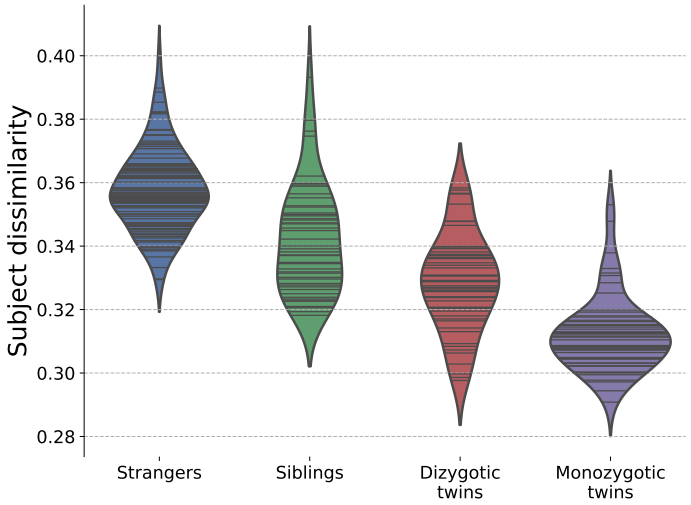


Figure 2.10: The distribution of subject pair dissimilarities $d(X, Y)$ as computed by Eq. 2.11. Black lines indicate each individual pair of subjects.

of this long tail corresponds to mixed-gender sibling pairs, of which none exist in the DZ twin population. Aside from several of the mixed-gender sibling pairs in the long-tail, the majority of mixed-gender sibling pairs are also well separable from the stranger population.

2.4 Discussion

We have identified a large fraction of deep white matter as being associated with genetic similarity. Areas of white matter with genetic similarity include the superior longitudinal fasciculus, the optic radiations, the middle cerebellar peduncle (particularly near the cerebellar nuclei), the corticospinal tract (through the posterior limb of the internal capsule and cerebral peduncle), and within the anterior temporal lobe adjacent to the amygdalae. These regions encapsulate nearly all of deep white matter. The large spatial extent of similarity among twins may reflect how fascicles are spatially arranged during neonatal development. Indeed, similarity of deep white matter organization may

be partially responsible for similarity in gray matter thickness and curvature if Van Essen's tension hypothesis is true (54).

Previous voxel-based studies of white matter associated with genetic similarity have identified overlap with our results. For example, (37) report that voxels in the temporal and frontal lobe subregions as showing the highest genetic influence, and additionally they identify significant subregions underlying posterior cortex. (41) report large portions of white matter as being affected by genetic control and the interaction of that with age, sex, socioeconomic status, and intelligence quotient. Both of these studies focused on per-voxel fractional anisotrophy scalars. In contrast, (41) consider orientation in a limited manner and report heritable effects in deep white matter based on the amplitudes of the first and second peaks of the per-voxel fiber orientation distribution.

We see a greater effect of white matter similarity with increasing genetic similarity. Monozygotic (MZ) twins are consistently more similar as compared to dizygotic (DZ) twins, which also display greater variance in their similarity. Non-twin siblings likewise are shown to be as similar in terms of these twenty-two white matter regions as DZ twins, though a longer tail of less similar pairs exists that is on par with strangers. This long tail is nearly entirely composed of mixed-gender siblings, however. No mixed-gender sibling exists in either of the MZ or DZ twin groups.

That these results generalize to non-twin siblings serves as an important validation of our method and results. These non-twin siblings were not used in the derivation of the twenty-two white matter regions except for being including as a larger group of Human Connectome Subjects to obtain a normalized template space from which the diffusion voxels were compared in.

The voxel dyads which we discover to be associated with genetic similarity are ones in which there exists sufficient individual variability for there to exist a group-wise difference, and which display a strong enough similarity across most twins as compared to the non-

related strangers. There is a strong assumption made that all pairs of interest are similar in the same way, i.e. that a single model of genetic similarity in white matter is sufficient to describe differences between all related siblings and non-related strangers. A promising future direction for this work is to consider multiple effects of similarity to exist in the population under consideration, such as looking for (potentially non-disjoint) partitions of the population such that the each has a strong similarity within only a single or small number of regions.

A potential limitation of the spatial coherence approach is in areas of high curvature. Such areas exhibit rapid changes in orientation between adjacent voxels and as such naturally have a lower baseline of spatial coherence. Though it is still possible to expect a population of interest to have a greater coherence in such areas, it might be harder to pick out the effect statistically which in turns biases results away from such regions.

In our results, we see that the effect of increasing the number of MDA peaks k above one decreases the extent and significance of the voxels identified. It is reasonable to assume this is because the additional orientations are largely associated with noise as they are in white matter regions where it is unlikely there are crossing fibers (per (23), a study on the same dataset, about 32.9% of white matter voxels are singly connected). Indeed, in Fig. 2.6, we see all major areas identified as significantly similar among twins with $k = 2$ peaks as being encompassed by the result with $k = 1$ peak with an exception of the centrum semiovale which has previously been identified as an area containing multiple crossing fibers. As we do not observe new large areas outside of the intersection of these two results, we conclude that either the similarity in the first peak is sufficient to detect regions associated with this population or that we are not adequately able to incorporate multiple orientations in our method in a manner that is robust enough against noise.

These twenty-two white matter regions, due to being associated with genetic background and/or upbringing common to the twins and siblings, could serve as a first ap-

proximation for a basis of defining white matter fingerprints that could be used to identify an individual over the course of their lifetime (55). However, this analysis might overlook regions that would be better suited to fingerprint individuals that are not preserved between pairs of twins or siblings, i.e. are not associated with genetic similarity but instead some broader concept of individual variance. A promising future direction for this method is in the application to a population of pairs of scans obtained from the same individual for a set of subjects, especially over a period of years, so as to understand what white matter regions contribute to such individual variance and how that changes over the course of our lifetimes.

A crucial component to the presented method is that it considers *pairs* of subjects as the fundamental unit of analysis, and not a single subject. This has two immediate consequences. The first is that it narrows down the scope of the analysis to those regions of white matter which display high similarity, or a small distance for some measure of distance, which is amenable to statistical analysis. The second is that this expands the size of the population under consideration from N individual scans to the order of N^2, given that sufficient care is taken during the analysis in sampling pairs and interpreting results, and given that every subject can be expected to have a high pair-wise similarity to the rest of the population of interest (which is not the case for the twin and sibling populations considered in the results). This aspect of considering pairs of input data is akin to Siamese Neural Networks (56), which have achieved state-of-the-art performance for learning models with very limited data and which has previously been applied to clinical diagnosis from functional MRI data (57).

2.5 Conclusion

In this Chapter we presented a method for identifying spatially contiguous but otherwise arbitrarily shaped white matter regions that are associated with a population of interest. This is a bottom-up approach which builds on the simplest possible building block, or neighboring white matter voxel dyads. We defined a similarity metric on such dyads and find a subset which are significantly more similar within the population of interest as compared to a control population, and control for multiple hypothesis testing. The largest such regions, composed of maximally sized overlapping dyads, are used for further analysis: a region-wise similarity is defined and is shown to have a large effect size between the two populations and generalizes to a previously unseen portion of the population of interest. Finally, a single similarity between a pair of subjects is defined on the basis of a set of such regions and is shown to separate the populations well.

This method is demonstrated on a population of monozygotic (MZ) and dizygotic (DZ) twins, with a control group composed of the same individuals with their pairings scrambled such as to keep the same demographic profiles but otherwise form unrelated strangers. The method discovers 3.7% of all white matter voxels to be associated with genetic similarity (35.1k voxels, $p < 10^{-4}$, false discovery rate 1.5%), 75% of which form twenty-two contiguous white matter regions. These white matter regions generalize to a population of non-twin siblings and are shown to be a good indicator of genetic similarity there as well, as compared to a population of strangers. The regions encapsulate nearly all of deep white matter.

Chapter 3

Characterizing White Matter with Generative Models

> ...when the brain is released from the constraints of reality, it can generate any sound, image, or smell in its repertoire, sometimes in complex and "impossible" combinations.
>
> Oliver Sacks, Hallucinations

3.1 Introduction

Recent advances in our understanding of the human brain have been enabled by improvements in magnetic imaging and the release of large, high quality, and publicly available data (45; 43; 58). In particular, diffusion-weighted magnetic resonance imaging (dMRI) reveals the organization of the brain's structural connectivity by mapping white

matter fiber tracts connecting disparate brain regions (59; 60). In seeking to understand the complex patterns present within the white matter microstructure—and what factors influence those patterns—a particularly powerful approach is to use a family of tools collectively known as generative modeling (61; 62). These approaches seek to understand the data using a predictive model, and in particular to understand what information is relevant for synthesizing such data.

In this chapter, we introduce a generative model for individual regions of white matter by predicting plausible and realistic directed white matter microstructure. In particular, our central concern is in understanding, and quantifying, to what degree contextual information is relevant to these predictions. We define contextual information as subject-specific characteristics that can be hypothesized to affect the local structure of the brain. For instance, how important is knowing an individual's handedness in predicting a given region, and how does that differ between regions? Contextual information can also include nearby brain structure, as the degree that a subject's latent characteristics, such as handedness, affects neighboring brain matter is relevant for the prediction of the region under consideration.

Towards the goal of quantifying the informativeness of such contextual information, we build upon recent advances in deep learning using Generative Adversarial Networks (GAN) (63; 64) and Bayesian uncertainty quantification (65). The GAN framework enables learning and sampling from a distribution given only samples from it, for example generating realistic but artificial photos of faces given only a data set of faces (66). This is accomplished by optimizing in tandem two models, namely a generator and critic model. The generator generates candidate samples of the distribution, while the critic evaluates if a given sample is from that of the true data distribution or from the generator. During alternative training steps, the generator seeks an optimal solution that leads the critic model to misclassify its outputs as real, while the critic learns to distinguish between

the distribution of generated versus real samples. These adversarial goals converge to a stable equilibrium when the distribution of generated data is indistinguishable from that of real.

Deep learning models including GANs have achieved significant success in mapping from high dimensional data to typically much simpler outputs. However, especially when the decisions made by such models are highly consequential—such as for health decisions, approval of bank loans, or navigating self-driving cars—it is difficult to place blind trust in their mappings (67). Uncertainty quantification of deep learning models seeks to measure when their decisions are likely to be erroneous and thus to signal when less confidence should be placed in their output (65). This is especially important in light of the potential propensity of neural network models to overfit on training data (68), a problem exacerbated in domains where training data is scarce or expensive to acquire. Generally, it is desirable to disentangle this predictive variance by attributing it to one of two major types of uncertainty: *epistemic* uncertainty, due to insufficient training samples, or *aleatoric*, due to noise inherent to the data or measurement thereof. In the following, we leverage these distinctions between different sources of uncertainty to develop two complementary metrics of white matter complexity.

The problem we consider is reconstructing the structural connectivity of an MR-imaged region of white matter given contextual information of the subject. In this initial work towards that goal, we consider a setting where the contextual information given is maximized: in predicting a region of brain matter, we provide as input a larger region which contains a masked version of the desired region, and aim to measure how difficult it is to *inpaint*. In this setting, we consider that the contextual information of adjacent voxels' diffusion data implicitly contains all the subject's latent characteristics relevant for the prediction: if a subject's handedness is informative for predicting the region than that information is present in the adjacent white matter. This provides a useful baseline

measurement of the relative inherent difficulty in generating different regions of white matter. An example of such generated regions can be seen in Fig. 3.1. The general task of predicting a masked region of an image is known an image inpainting. The GAN framework in particular has achieved success in inpainting, and has been successfully applied to MR images including to ameliorate the effects of localized perturbations (69) and to synthesize lesions of the liver for data augmentation (70).

A similar line of inquiry to ours was proposed by Tanno et al. (71), in which they showed the usefulness of uncertainty estimation to enhance the resolution of acquired MRI images. In their results, they mapped how uncertainty varied across the brain, particularly around tumours in individual scans. Their result however did not correct for higher uncertainty in regions of greater white matter intensity. Kwon et al. (72) showed how uncertainty quantification enhanced stroke segmentation in individual brain scans. More generally, generative approaches have been used with diffusion imaging to model the network that is the human connectome, including to probe at the mechanisms behind its growth and controllability (73; 61; 62) and the influence of the spatial embedding of the brain itself to the connectome (74). We build upon our previous work (2) which first proposed the in-painting methods, and introduce more refined methods and analysis of results that include the use of both spatial and non-spatial features.

3.2 Imaging data and preprocessing

The data was processed identically to Volz et al. (6), but is reported here as well for completeness. We build upon the work of of the Washington University-Minnesota Consortium Human Connectome Project (44; 43) which recruited participants from Washington University (St. Louis, MO, USA) and surrounding area. All participants gave informed consent. The data is derived from 630 participants (358 female, 272 male). For

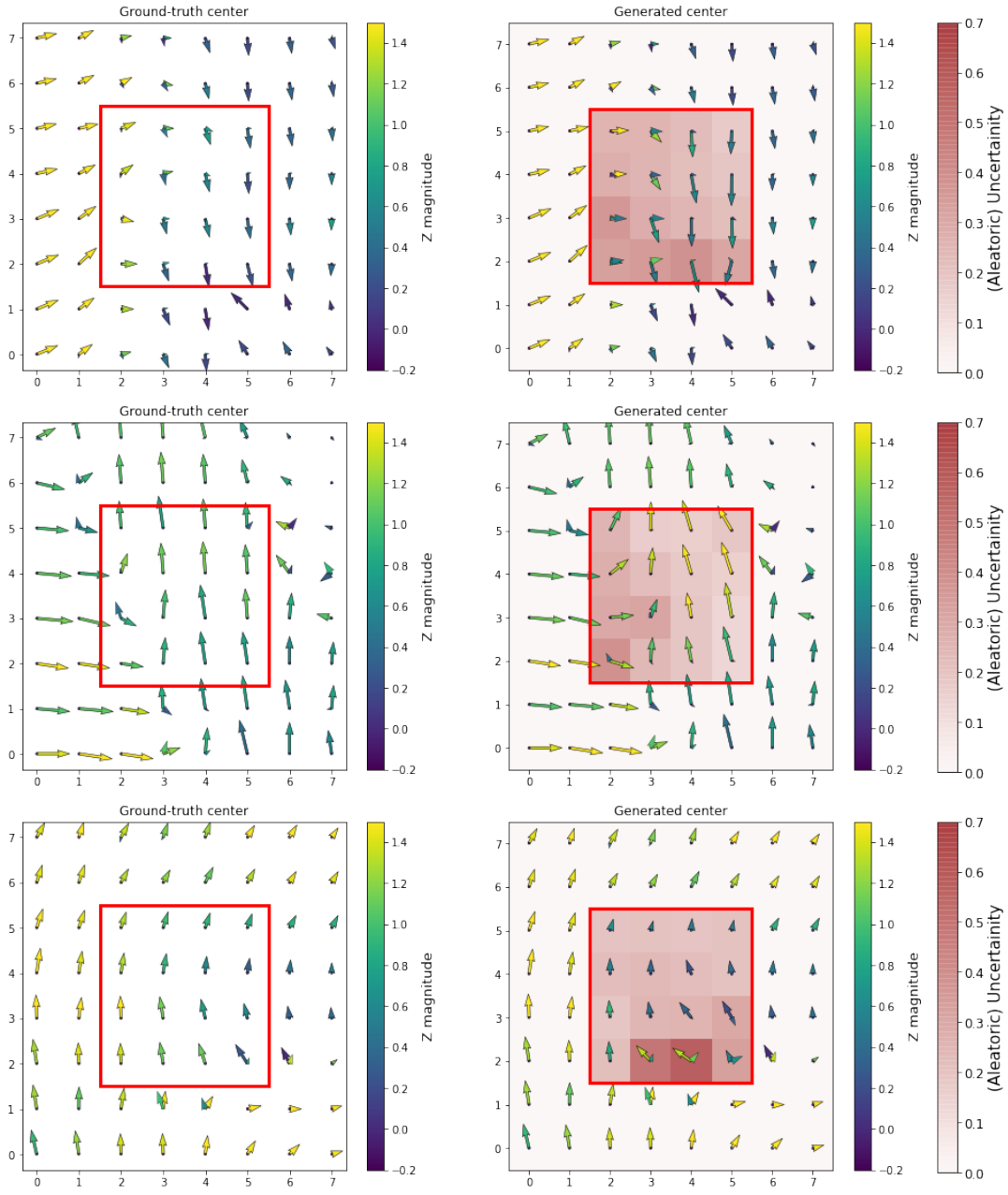


Figure 3.1: Three examples of synthesized brain structure. Two dimensional slices of white matter regions from real diffusion data (left column) and corresponding generated white matter microstructure (right), where the region being generated is delimited by the red square. The surrounding white matter outside the box is the contextual information provided in synthesizing the brain region. The red overlay displays the average aleatoric uncertainty within each generated voxel, where darker red corresponds to more uncertainty. The magnitude of the third dimension of each vector is shown as their blue-to-yellow color.

this study, the participants were randomly assigned to one of three sets: the models were fit on a training set ($n = 442$), the model hyerparameters were tuned based on their performance on a validation set ($n = 94$), and all results presented are from a test set ($n = 94$).

The structural and diffusion data were collected on 3T Connectome Skyra system (Siemens, Erlangen, Germany). The diffusion volumes were collected with a spatial resolution of 1.25x1.25x1.25 mm^3, using three shells at $b = 1000$, 2000, and 3000 s/mm^2 with 90 diffusion directions per shell and 10 additional b0s per shell. Spatial distortion and eddy currents were corrected using information from acquisitions in opposite phase-encoding directions, as well as head motion (45). The high-resolution structural T1 weighted and T2 weighted volumes were acquired on the same scanner at 0.7mm isotropic resolution. Minimally preprocessed images were reconstructed in DSI Studio (http://dsi-studio.labsolver.org) using Generalized Q-Sampling Imaging (GQI) (46).

Skull stripped, aligned, and distortion corrected T1w and T2w volumes (45) were rigidly registered to the subject's GFA volume. The symmetric group wise normalization (SyGN) method implemented in Advanced Normalization Tools (ANTs, http://stnava.github.io/ANTs/) was used to construct a custom multimodal brain template using the data of 38 HCP subjects (47) that included proportions of racial, gender, and handedness that chosen through stratified random sampling according to these features. Each subject's GFA, T1w, and T2w volumes were used during template creation with weighting factors of 0.5 (GFA) $\times$ 1 (T1w) $\times$ 1 (T2w). Templates were created after 5 iterations. Templates from the 4th and 5th iterations of multi-modal template construction were inspected to check that the templates had stabilized. All individual datasets were ultimately normalized to this template using all 3 modalities and symmetric diffeomorphic normalization (SyN) as implemented in ANTs (48).

3.2.1 Anisotrophy indices

Anisotrophy indices were estimated using the following formulae (26) based on orientation distribution functions (ODFs) obtained by GQI reconstruction, and by fitting the standard tensor model used in diffusion tensor imaging (DTI) (60), as implemented in DSI studio. FA was calculated based on tensor fits

$$\mathrm{FA} = \sqrt{\frac{3}{2}}\sqrt{\frac{\sum_{i=1}^{3}(D_i - \tilde{D})^2}{\sum_{i=1}^{3} D_i^2}} \tag{3.1}$$

with D_i denoting the directional diffusivity corresponding to the ith eigenvector of the diffusion tensor. ODFs reconstructed using GQI were used to calculate GFA and MDA. GFA was computed as

$$\mathrm{GFA} = \sqrt{\frac{n\sum_{i=1}^{n}(\Psi_{u_i} - \tilde{\Psi})^2)}{(n-1)\sum_{i=1}^{n}\Psi_{u_i}^2}} \tag{3.2}$$

with Ψ representing the ODF and u_i denoting the u_i-th direction of the ODF. In contrast to the FA, GFA incorporates diffusion coefficients from the whole set of discrete directions included in the reconstructed ODF instead of only the directions corresponding to eigenvectors of a diffusion tensor fit (25). This is also true for MDA which was estimated as

$$\mathrm{MDA} = \frac{1-\mu}{\sqrt{1+2\mu^2}}, \text{ where } \mu = \left(\frac{\Psi_{\min}}{\Psi_{\max}}\right)^{2/3} \tag{3.3}$$

with $\Psi_{\min}$ and $\Psi_{\max}$ representing the smallest and largest directions sampled in the ODF.

3.2.2 Extracting MDA vectors

Each ODF $\Psi(\theta)$ was calculated with GQI on a set of 642 approximately-evenly spaced directions $\theta \in \Theta$ on a tesselated icosahedron. ODF magnitudes were rescaled so that the sum of each ODF is $\sum_{\theta \in \Theta} \Psi(\theta) = 1$. We then calculated the multi-directional anisotropy (MDA) value for each direction θ using Eq. 3.3, but with

$$\mu(\theta) = \left(\frac{\Psi(\theta)}{\Psi(\theta_{min})} \right)^{2/3} \tag{3.4}$$

MDA values and their corresponding directions were calculated for the four largest local maxima in every ODF, which are ordered by decreasing size. A separate study of the same data found that ODF peaks become very noisy after the 4th direction (6). The vector fields corresponding to the local maxima were warped into the group template using ANTs. 3D volumes containing the MDA magnitudes were also warped to the group template and used to scale the normalized vectors. White matter voxels were determined by segmenting the weighted average template of T1w, T2w, and GFA volumes in FreeSurfer (49). In this chapter we consider generative models for the largest two MDA vectors within each voxel.

3.3 Learning localized wiring patterns

3.3.1 Problem formulation

We consider the problem of generating the structural connectivity of a white matter region R given contextual information relevant to that prediction. In particular, we desire that the predictions form realistic wiring patterns and we aim to quantify *how* relevant the contextual information is to the prediction.

In this chapter, we consider the region R to be a 3D cube with k voxels along each

side. The contextual information needed to predict the voxels within R should be at minimum enough to locate R, such as a description of which region it corresponds to or its spatial coordinates. It may be also desirable to provide subject-specific information, for instance their handedness, biological sex, and/or presence of neurological disorders. In this initial work towards that goal we seek to understand the difficulty associated with generating a region in the limit of maximal contextual information: In predicting the white matter microstructure of a region, we provide as context C the adjacent white matter microstructure of R consisting of a cube with $2k$ voxels to a side, with the center voxels corresponding to R masked. We denote this as maximal contextual information as any latent characteristics of the subject that affects the brain structure within the region would affect neighboring brain matter.

We ensure that the predicted $\hat{R}$ forms a realistic MR imaged brain region with a GAN approach, in which alongside the generator a critic model is trained to tell apart synthetic examples from real. Both the generator and critic are alternatively updated to adversarially outperform the other, converging to an equilibrium in which synthetic and real examples appear indistinguishable to the critic. In particular, we adopt the Wasserstein-GAN formulation (64) which leads to improved stability in the convergence of the models by providing a smother loss surface for the critic.

Our objective is not solely to learn a generative model of structural connectivity but to quantify how informative the contextual information provided is. To that end we build upon work by Kendall and Gal which incorporated the two major types of uncertainty with a deep learning framework (65): epistemic uncertainty, which is due to insufficient data to accurately infer model parameters, and aleatoric uncertainty, due to noise inherent to the data for instance because of sensory limitations. Following Kendall and Gal, epistemic uncertainty is implemented as dropout variational inference, where dropout is applied to layers of the model during *both* training and inference to allow for stochastic

draws of model parameters. Aleatoric variance is obtained by the model predicting both a typical maximum likelihood point estimate as well as the variance of that estimate. This is accomplished by having the predicted variance serve as a loss attenuation during optimization, that is point estimates with large variance are penalized less than those with greater associated confidence, while high variance is separately penalized.

We seek to minimize the following objective function for the generative model of brain structure,

$$\mathcal{L}_G = -\mathrm{D}(\hat{R}|C) + \frac{1}{|R|}\sum_{v\in R}\left(\frac{1}{2\hat{\sigma}_v^2}\mathrm{d}(R_v, \hat{R}_v) + \frac{1}{2}\log\hat{\sigma_v^2}\right) \tag{3.5}$$

,

where v indexes individual vector components of the predicted region R. $D(\hat{R}|C)$ corresponds to the critic model (64) trained to discriminate between generated $\hat{R}$ and ground-truth samples given contextual information C. It outputs an unconstrained scalar value whose sign indicates the critic's decision of real versus fake. The model weights are sampled by activating the dropout layers during both training and inference, with both the predicted mean $\hat{R}$ and variance $\hat{\sigma^2}$ being average across multiple draws from the model weights. The aleatoric variance $\hat{\sigma}^2$ is estimated by the generative model itself, and in practice can be interpreted as a learned loss attenuation. We define the $\mathrm{d}(\cdot,\cdot)$ as the minimum L2 distance between the MDA vectors and their reflections (1),

$$d(x,y) = ||\min(x-y, x+y, -x-y, -x+y)||_2 \tag{3.6}$$

as the ODFs they are derived from are fundamentally undirected whereas the vector representations are necessarily directed.

The first term of Eq. 3.5 is the adversarial loss of the Wasserstein GAN formulation (64), and encourages that the generated samples of white matter be anatomically

correct. The critic outputs an unconstrained scalar whose sign indicates if an input R is real or synthetic, and its magnitude to how confident the critic model is in that decision. Any difference in the distributions of real versus generated data would be exploted by the critic model. As such, with a sufficiently capable critic when both generator and critic have converged to the game theoretic equilibrium of the generated distribution matching that of the real data, with the critic unable to perform better than a coin toss (63).

The second term of Eq. 3.5 is the reconstruction error of the generated sample compared to the ground-truth, where predicted components with large aleatoric variance contribute less loss. However, the third term penalizes large aleatoric variance $\hat{\sigma^2}$, providing an opposing force to the second term to encourage high $\hat{\sigma^2}$ only where there is insufficient contextual information to accurately generate a sample, and prevents trivial solutions of high variance throughout every prediction.

We implement both the generator and discriminator as convolutional neural networks with the size of the brain region to be predicted set at $k = 4$, balancing a more holistic notion of brain structure with the ability to localize our results sharply within the neuroanatomy of the brain. This corresponds to a brain region of size $5 \times 5 \times 5$ mm^3 due to the 1.25mm spatial resolution of the diffusion volumes. For further details of the models including their architecture and optimization, we refer the reader to 3.4.

3.3.2 Quantifying informativeness of contextual information

Our central interest is in quantifying *how* challenging it is to predict a region of white matter if given contextual information. We define the aleatoric variance of brain region as the average coefficient of variation of the aleatoric variance across all predicted

components of that region,

$$\text{var}(R) = \frac{1}{M|R|} \sum_{v \in R} \frac{\sum_{m \in v} \sigma_{v_m}}{||v||_2} \tag{3.7}$$

where v_m is the mth component of voxel v's MDA vectors, with $1 \leq m \leq M$. This normalizes the aleatoric standard deviation by the vector's magnitude, allowing for meaningful comparison across brain regions with differing amount of white matter intensity in the diffusion scan. In the limit of maximal contextual information, we term this measure of a region's uncertainty as its complexity.

3.4 Model architecture

As described in Section 3.3, we implemented the generative model as a conditional GAN using a convolutional neural network architecture. The exact implementation of the model are presented as follows.

Generator The generator followed the coarse-to-fine architecture proposed by Yu et al. (75), in which the network outputs a 'coarse' prediction with an auxiliary loss of only reconstruction error, which is then refined to a final output which optimizes Eq. 3.5. In our implementation, these two stages of the generator have identical network architecture but with separate weights that are optimized end to end. Each receives as input a $(2k)^3$-sized cube of voxels, where the center k^3 voxels correspond to the region R being generated which is initially masked with zeros. The output of the first stage of the generator is overlaid over this center which serves as the input to the second stage. As described in Section 3.3, in this chapter we selected $K = 4$.

Each stage of the generator consisted of four 3D convolutional layers, the first two with filter size of 3^3 while the last two have a filter size of 1^3, with the number of

filters set at 128, 128, 64, and finally 6—corresponding to the two 3D MDA vectors being generated at every voxel. Each convolution was applied with no padding. Each had a leaky ReLU (76) activation with 0.3 slope for negative values, except the output layer which has a tanh activation. Following the activation we applied a 3D spatial dropout (77) with probability 0.3, which allowed for stochastic sampling to compute epistemic uncertainty of the generated samples by applying these dropout layers during inference. For the second stage of the generator a parallel last layer predicted the aleatoric variance $\log \hat{\sigma}^2 =$ (see Eq. 3.5, where the logarithm provides a more stable training regime avoding division by zero, as noted by Kendal and Gal (65)), and this layer was otherwise identical aside from having no activation function applied.

Critic The critic is a Wasserstein GAN (64), predicting an unconstrained value for each $(2k)^3$ input received. It is composed of four 3D convolutional layers followed by three dense layers. The first three convolutional layers have 128 filter sizes of $(2k)^3$, with the last having 64 1^3 filters. The dense layers receive their flattened output, and have filter sizes 128, 32, and 1. Excepting the final layer which has no nonlinearity, leaky ReLUs with negative slope 0.3 followed each layer with the convolutional layers additionally applying batch normalization (78). To ensure the Lipschitz continuity of the critic, its weights were constrained to the range $[-0.02, 0.02]$ by clipping. It should be noted that there exist better methods to ensure the Lipschitz constraint (79), but they did make a significant difference to the converged models and so the simpler procedure was kept.

Optimization procedure The weights of each layer for both models were initialized from a He normal distributions (77). During each epoch of training, the critic was first advanced through `n_critic` $= 5$ batches before the generator iterates once. Each batch consisted of 32 samples, each uniformly at random selected from white matter voxels,

with each epoch consisting of one sample from each of the scans within the training set. Training proceeded for 20 thousand epochs, elapsing approximately 12 hours on an NVIDIA GeForce RTX 2080 GPU. Both models were updated with the RMSprop optimizer, with learning rates set to 10^{-4} for both models. Hyperparameters were tuned based on their performance on the validation set, and all results are presented from the test set.

3.5 Results

We evaluated the trained model on every white matter voxel across the 94 subjects in the test set. We consider only the largest two MDA vectors in each voxel. The aleatoric variance of the model reflecting the noise inherent to the data was well calibrated, on average across all subjects and all regions the standard deviation of the aleatoric uncertainty correlates highly with the actual error between the ground-truth and generated regions (Pearson's $R = 0.872$, $p < 10^{-10}$). Examples of such generated regions are shown in Fig. 3.1.

We next compute the complexity atlas of white matter structure as the mean $\text{var}(R)$ across subjects using Eq. 3.7 for every region R in white matter. Each $\hat{R}$ was computed as the mean across $T = 30$ sampled outputs using variational dropout inference to capture epistemic uncertainty, i.e. due to variance in the posterior distribution of model parameters. This uncertainty was minimal in comparison to the aleatoric, or only proportionally 37.0% on average (variance 15.8%). We observe that this complexity map correlates highly with measures of voxel anisotrophy, explaining half of the variance of GFA (mean 50.1%, ranging 47.1 - 52.6%), MDA0 (mean 50.5%, ranging 46.5 - 53.0%), and FA (mean 46.4%, ranging 43.1 - 49.4%) within white matter of individual scans (these results can be downloaded behind the following link, TODO).

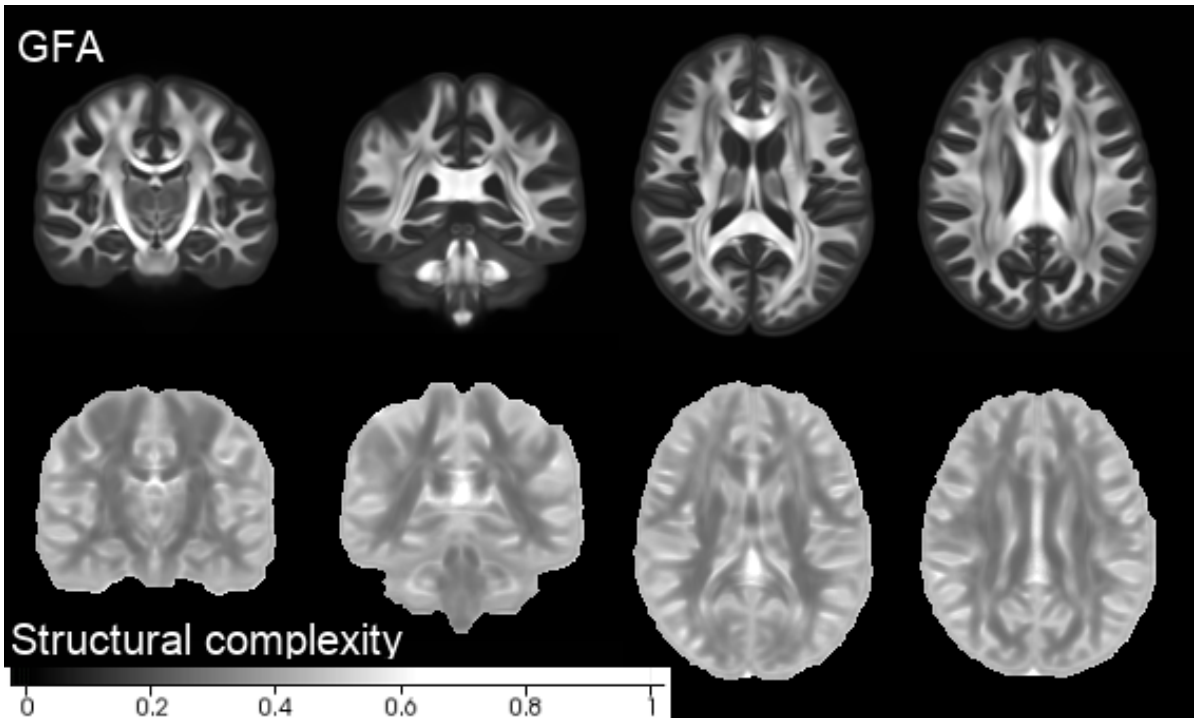


Figure 3.2: The complexity atlas of white matter structure aligns well with measures of anisotrophic signals within white matter. For dMRI scans, this measures the baseline difficulty of generating the white matter microstructure across the brain given near maximal contextual information.

As can be expected by the amount of variance explained in anisotrophic measures, the complexity map aligns well with those measures. Major fiber bundles are easily discernable in Fig 3.2 as regions of low relative complexity, with high values being observed closer to the cortical boundary where white matter connections disperse into gray matter. The largest departure from group isotrophic maps are in the splenium and genu of the corpus callosum, where large neural bundles both cross between hemisphere as well as cross the trunk of the corpus callosum. We observe that this is partly due to a limitation of representing a voxel's anisotrophy as a directed vectors sampled from a half plane, as there is a discontinuity in the vector representation near the boundary of the half plane from which the vectors range. Although accounting for this using Eq. 3.6, we observe high aleatoric uncertainty associated with vectors on either side of that boundary.

We further analyse the heterogeneity between subjects in deriving this complexity atlas. We define heterogeneity as the relative standard deviation of Eq. 3.7 across the population. As seen in Fig. 3.3, we observe greater variance between subjects along large white matter bundles fan out towards the cortical boundary. The greatest variance occurs in regions where the generative model is limited by the provided contextual information, namely the adjacent voxels' white matter microstructure. This also occurs at the base of the brain stem, where the context is abruptly cut off, and at the splenium of the corpus callosum where the MDA vectors are near a discontinuity in their representation as described above. Interestingly, we also observe relatively high heterogeneity between subjects in the same 'crossing pocket' region that Volz et al. observe (6), lending further evidence to this region being problematic for anisotrophy-based fiber tracking algorithms.

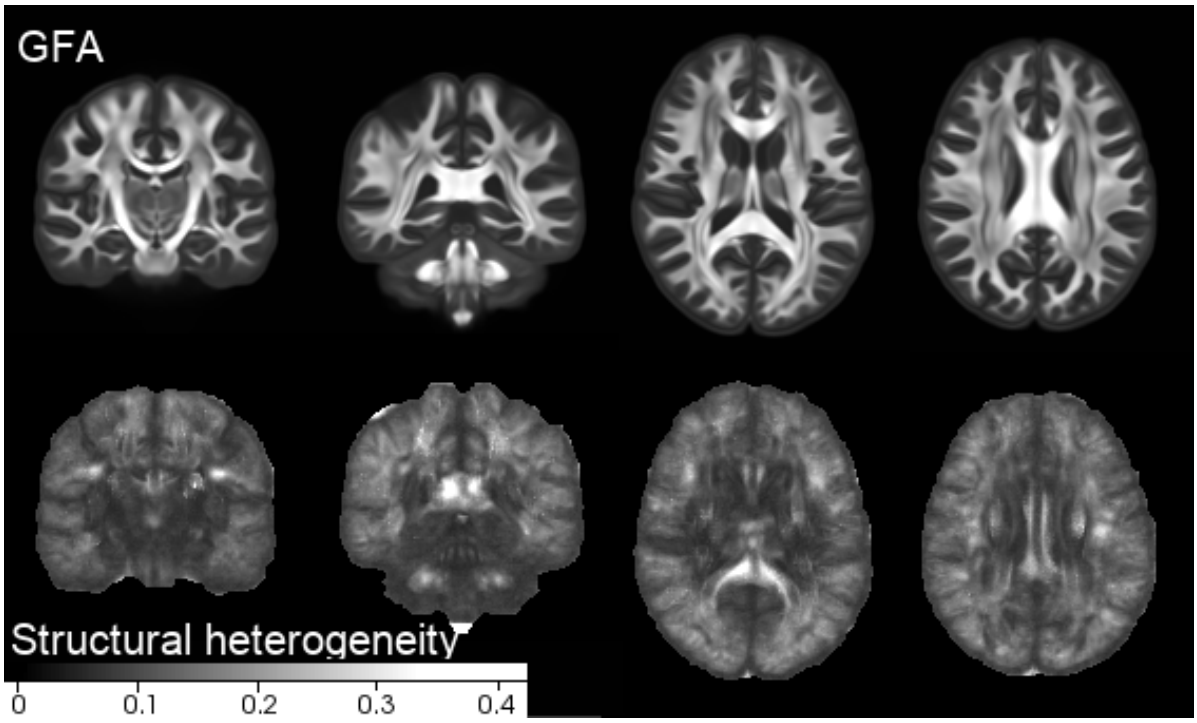


Figure 3.3: Between subject variance of generative uncertainty is highest near junctions of large white matter bundles (as seen in left most figure) and near where white matter fans out towards cortical boundary.

3.5.1 Quantifying confounders for voxel-based analysis

Commonly, the statistical analysis of diffusion data between and within individuals is on the basis of voxel-based measurements of anisotrophy, the degree to which the diffusion signal within a voxel is directed rather than isotropic. Clinically relevant differences in such anisotropic measures, which are interpreted as proxies for the integrity of white matter (10), have been discovered for multiple subject groups including patients with Alzheimer's disease (80), schizophrenia (81), or COVID-19 (82). However, care must be taken in interpreting these results as deficiencies of white matter integrity as the diffusion signal itself is confounded by partial volume effects (83; 84) and voxels containing multiple fiber bundles projecting in multiple directions.

Towards ameliorating the confounding effects of multiple directions of fiber tracts within voxels, Volz et al. proposed an probabilistic atlas of fiber counts (6). This atlas has a majority of voxels containing two differentially directed fiber bundles (44.7%), and they further showed that compartmentalizing voxels according to their number of fiber directions significantly reduces the variance in analysis of anisotropic measures. To a first approximation, regions of white matter containing such crossing or kissing fibers should be expected to be more difficult to generate synthetic examples of; indeed, our results of regional complexity of white matter is in significant agreement with this fiber atlas crossing atlas (Spearman's $\rho = 0.616$, $p < 10^{-10}$), as seen in Fig 3.4. However, as seen in Fig. 3.5, we see significant heterogeneity of regional complexity within the voxel groups they define. We believe our results provide an orthogonal measurement of confounding factors to diffusion analysis that they consider.

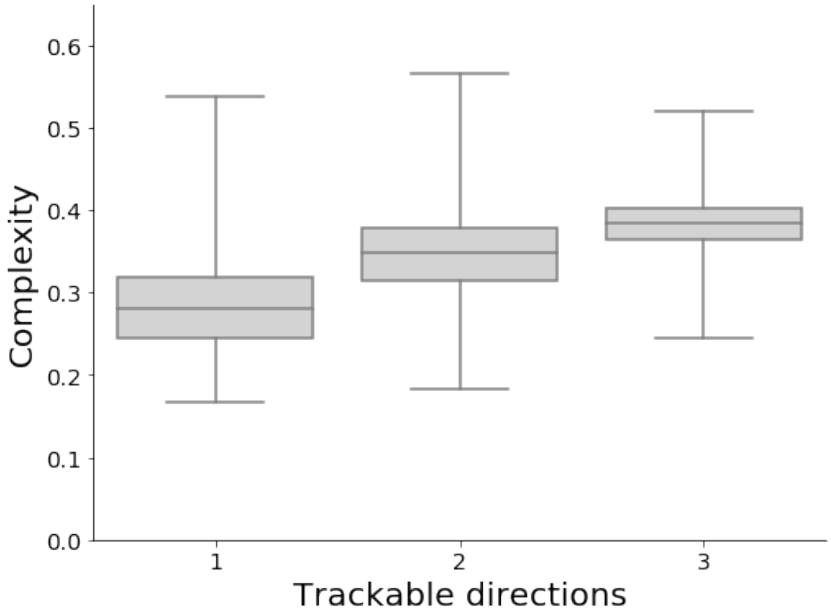


Figure 3.4: The inherent complexity we measure of white matter voxels is positively correlated with the number of differentially trackable fiber bundles within them (Spearman's $\rho = 0.616$, $p < 10^{-10}$. The boxplot shows the quartiles of each distribution, with the box itself bounding the 25 and 75 percentiles and the midline at the median. The whiskers extend to three standard deviations around the mean.

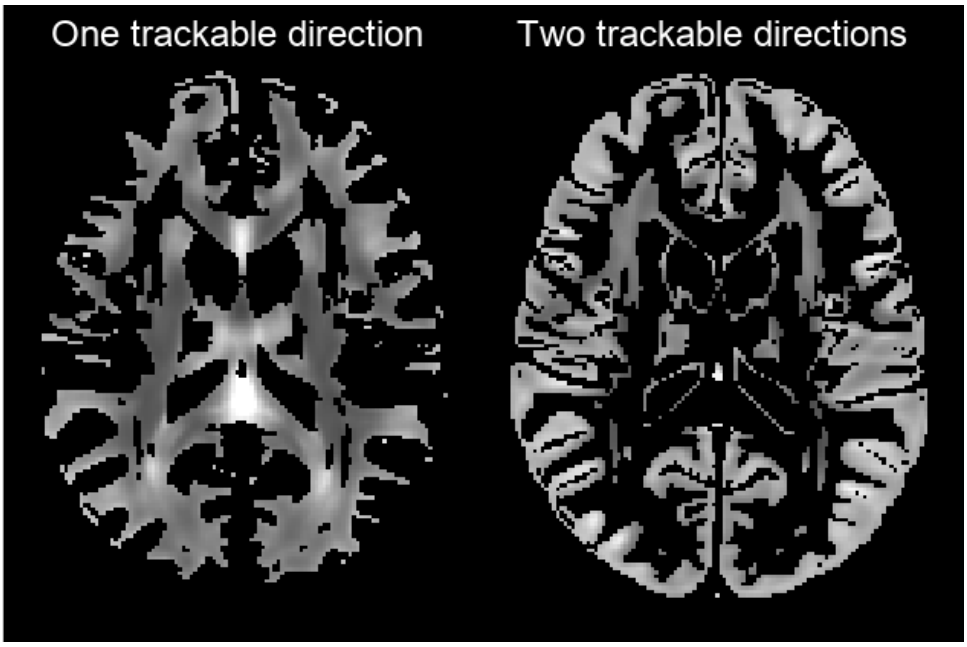


Figure 3.5: Comparison with results from Volz et al. (6), segmenting our result by the number of trackable directions they define. Though overall there is significant agreement in results, we see considerable variance within voxels containing the same number of differentially directed fiber bundles.

3.6 Discussion

These results measure the minimum difficulty in generating the white matter microstructure of a brain region given maximal contextual information from the subejct in doing so. We believe such generative models, and uncertainty quantification thereof, are a promising framework for better understanding the relationship between brain regions and various characteristics of the individual.

3.6.1 Local versus global generative models

The formulation we choose for this chapter was that of a general purpose generative model of brain structure, capable of synthesizing any brain region equally well. An alternative formulation worth considering is to learn a specialized 'local' generative model for a single brain region, that is trained on a subset of the data specific to only that region. Conceptually, the benefit of this alternative approach is that it more directly measures the information content of a region under consideration in isolation from other regions. In addition, the relative performance of such a local model versus the global model could disentangle some of the complexity due to the region itself versus that due the heterogeneity between subjects: an otherwise complicated wiring pattern that would be difficult for a global model to generate may be identical between all subjects and thus easily learned by local model. The practical downside of such local models is that the size of the training data becomes smaller—but we believe a feasible future path towards realizing such models is with transfer learning from a global model to a local one (85).

3.6.2 Possible extensions

This chapter studied the inherent difficulty in generating the directed microstructure of white matter regions under the limit of maximizing the contextual information given.

Other variations of this framework are also interesting to study, for instance a simplified setting of predicting directly anisotropic measurements of the voxels, or with more limited contextual information such as predicting the white matter region given their coordinates and limited information of the individual. This general framework of generative models has been considered for network models of structural connectivity derived from diffusion images (61; 73), or for other modalities including T1- and T1-weighted images (86) or computed tomography (87). However, incorporating a notion of how informative contextual information is to these generative models enables further characterization of the structure of the human brain.

3.6.3 Limitations

A limitation to our method of measuring the total uncertainty associated with a brain region with Eq. 3.7 is that this measure does not account for covariance between voxel components. For example, if the brain region contained a uniform vector field with one degree of freedom in the the generative uncertainty, Eq. 3.7 would overestimate the total variance by a factor equal to the number of voxels present. To more appropriately measure the total variance one should discount the covariance present between voxels. A possible step towards that direction is, if the variance is assumed to be normal, then the total differential entropy is porportional to the log determinant of the covariance matrix (88),

$$H(R) = \frac{k}{2}\ln(2\pi e) + \frac{1}{2}\ln(|\Sigma|), \tag{3.8}$$

where k is the dimensionality of the region R and $|\Sigma|$ the determinant of the covariance matrix.

There are two challenges in correcting for this limitation, namely estimating the

covariance matrix and numerical difficulties associated with measuring the volume of a high dimensional ellipse (89). The approach adopted in this chapter to measure the aleatoric variance of the prediction (65) only measures the diagonal of the covariance matrix. However, if the epistemic variance is assumed to be Guassian, then an empirical epistemic covariance matrix Σ_E can be computed from repeated samples from the model given the same input. If furthermore the covariational structure of the aleatoric and epistemic variance is assumed to be similar up to scaling factors, then the aleatoric covariance matrix can be estimated as

$$\Sigma_a \approx \text{diag}(\Sigma_a)^{\frac{1}{2}}\text{diag}(\Sigma_e)^{-\frac{1}{2}}\Sigma_e\text{diag}(\Sigma_e)^{-\frac{1}{2}}\text{diag}(\Sigma_a)^{\frac{1}{2}} \tag{3.9}$$

,

that is by scaling the epistemic covariance matrix such that its diagonal equals that of the aleatoric covariance. However, in practice, we observe that the second challenge of measuring the volume of the high-dimensional Guassian corresponding to this estimated aleatoric covariance matrix is not easily surmountable; there is significant covariance between neighboring voxels causing the total differential entropy to be computed as nearly zero.

3.7 Conclusion

We proposed a framework for measuring the correlation between brain regions with contextual information of an individual. Here, contextual information is defined as characteristics and/or attributes of the individual that may affect brain matter. To measure that relationship, we build upon Baysian uncertainty quantification to estimate the predictive confidence of a well-trained model in generating brain regions given such contex-

tual information.

We developed a generative model based on Generative Adversarial Networks (GANs), which have in recent years achieved superlative performance in sampling from high-dimensional and complex distributions, such as that of portraits of human faces. Our model is able to synthesize plausible and realistic regions oriented white matter microstructure as imaged by diffusion MRI.

We demonstrated this framework in the limit of maximum contextual information, in which we assume that any factors affecting a given brain region are also present in neighboring voxels. This allows for measuring a baseline of wiring complexity throughout white matter, which allow for quantifying the degree to which voxels contain ambiguous connectivity at current sensor resolutions. We observe significant agreement with our results and that of previous work in measuring the number of distinct fiber bundles per voxel (6).

Chapter 4

Characterizing White Matter with Predictive Models

> It's a useful habit never to believe more than half of what people tell you, and not to concern yourself with the rest. Rather keep your mind free and your path your own.
>
> Halldór Laxness,
> Independent People

4.1 Introduction

Convolutional neural networks (CNNs) have achieved breakthrough performance on most machine learning tasks, for instance achieving super-human performance in classifying natural images (90) and playing the game of Go (91). There has been much interest in applying such models on magnetic resonance imaging (MRI) brain data, with

success in predicting biological age (92; 93), classification of Alzheimer's disease (94) and autism (95), and detecting lacunes within brain scans (96). However, training such models on brain scan data is challenging as compared to that of natural images: brain scan data is gathered in much smaller quantities resulting in smaller datasets, and each individual scan is typically of much higher dimension than commonly studied computer vision datasets.

These challenges call for efficient use of the collected data by developing methods that are well adapted to the spatial structure of the brain. Standard lessons from computer vision do not always transfer well to this domain, for instance despite pioneering work in applying CNNs to brain image analysis in considering two-dimensional slices or patches (97; 98; 99), a recent review suggests that, despite the increased computational costs and memory requirements, three-dimensional approaches should ultimately outperform (100).

A defining feature of brain scans as compared to natural images is that brain data is spatially homogeneous: commonly, as part of the preprocessing, MRI brain data is spatially normalized to a shared template to allow for voxel-wise comparison of the same spatial region across a population. Indeed, incorporating within-brain spatial location has been shown to improve the performance of models (101). Recent work has incorporated this translation variance property of processed MRI data by learning a different set of convolutional filters at different locations of the brain (102; 92).

A critical concern, and common complaint, is that deep neural network models are considered to be opaque black-box models which may achieve high performance on singular metrics but yield little understanding of the dataset itself (67; 103). As such, developing interpretable models is critical for evidence-based decision making and a necessary step towards widespread use of deep learning models for patient care. Piercing the veil of machine learning models to realize this does not necessarily have to come at the

cost of performance, but often goes hand-in-hand with efficient domain-guided architectures. For instance, a recently proposed framework for Alzheimer's disease classification achieved diagnostic performance surpassing that of a team of 11 practicing neurologists with a CNN model that provided disease probability maps throughout the brain scan with every individual diagnosis (94).

We propose a CNN classifier for brain scans that incorporates within-brain spatial information to learn which regions of the brain are most informative for a prediction, across the population of interest. We hypothesize that for many common classification tasks, a sparse subset of the brain image is *sufficient* to learn a good classifier, enabling learning a model on otherwise relatively few high-dimensional data points while offering an interpretable explanation of the relative informativeness of brain regions. To that end, we propose decomposing a whole brain CNN into a patch-based classifier and a patch-saliency model that only considers the spatial location. During optimization, these two models are trained jointly by sampling patches and simultaneously learning the relative informativeness of the patches as well as a classification model of them, as seen in Fig. 4.1. By optimizing on only a sparse subset of an image at time, we ameliorate the memory and computational costs of three-dimensional convolutions and see a regularizing effect which prevents overfitting on training data.

We apply our method to learn salient regions of white matter, as imaged by diffusion weight MRI, for predicting the biological sex of individuals. Sex differences are known to be associated with white matter microstructure, likely mediated through sex hormones (104). Furthermore, understanding sex differences within the neural wiring of the brain is of societal interest due to their prominence in the behavior of humans (105).

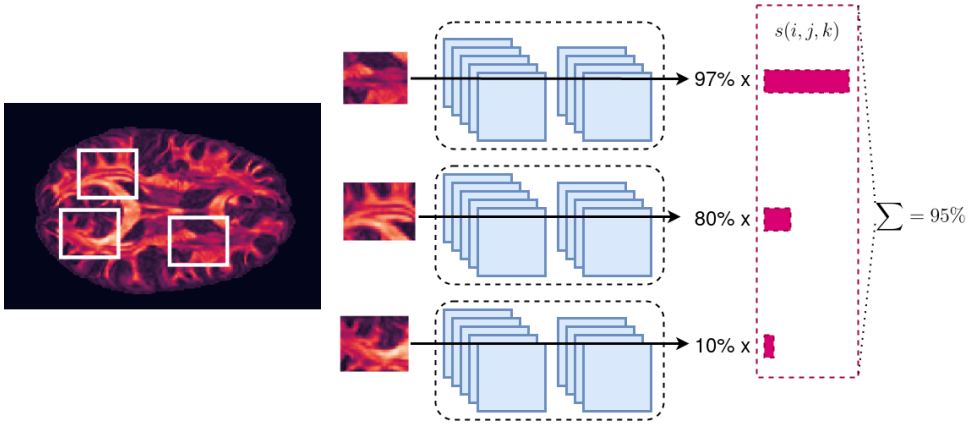


Figure 4.1: Simultaneously learning a patch-based classifier as well as the relative spatial saliency of each patch. During optimization, random patches of the MRI scan are sampled (left, within white boxes) and provided to the single patch-based classifier (middle). As not all possible patches are informative for the classification task, the relative performance of classifying different patches varies depending on a patch's spatial location. We propose learning a separate saliency function $s(i, j, k)$ which learns the relative informativeness of spatial locations (right), emphasising informative regions for both inference as well as backpropagation of gradients, while enabling an informative saliency map for the predictive task.

4.2 Imaging data and preprocessing

The data was processed identically to Volz et al. (6), but is reported here as well for completeness. We build upon the work of of the Washington University-Minnesota Consortium Human Connectome Project (44; 43) which recruited participants from Washington University (St. Louis, MO, USA) and surrounding area. All participants gave informed consent. The data is derived from 630 participants (358 female, 272 male). For this study, the participants were randomly assigned to one of three sets: the models were fit on a training set ($n = 442$), the model hyerparameters were tuned based on their performance on a validation set ($n = 94$), and all results presented are from a test set ($n = 94$).

The structural and diffusion data were collected on 3T Connectome Skyra system (Siemens, Erlangen, Germany). The diffusion volumes were collected with a spatial

resolution of 1.25x1.25x1.25 mm^3, using three shells at $b = 1000$, 2000, and 3000 s/mm^2 with 90 diffusion directions per shell and 10 additional b0s per shell. Spatial distortion and eddy currents were corrected using information from acquisitions in opposite phase-encoding directions, as well as head motion (45). The high-resolution structural T1 weighted and T2 weighted volumes were acquired on the same scanner at 0.7mm isotropic resolution. Minimally preprocessed images were reconstructed in DSI Studio (http://dsi-studio.labsolver.org) using Generalized Q-Sampling Imaging (GQI) (46).

Skull stripped, aligned, and distortion corrected T1w and T2w volumes (45) were rigidly registered to the subject's GFA volume. The symmetric group wise normalization (SyGN) method implemented in Advanced Normalization Tools (ANTs, http://stnava.github.io/ANTs/) was used to construct a custom multimodal brain template using the data of 38 HCP subjects (47) that included proportions of racial, gender, and handedness that chosen through stratified random sampling according to these features. Each subject's GFA, T1w, and T2w volumes were used during template creation with weighting factors of 0.5 (GFA) $\times$ 1 (T1w) $\times$ 1 (T2w). Templates were created after 5 iterations. Templates from the 4th and 5th iterations of multi-modal template construction were inspected to check that the templates had stabilized. All individual datasets were ultimately normalized to this template using all 3 modalities and symmetric diffeomorphic normalization (SyN) as implemented in ANTs (48).

4.2.1 Anisotrophy indices

with Ψ representing the ODF and u_i denoting the u_i-th direction of the ODF. In contrast to the FA, GFA incorporates diffusion coefficients from the whole set of discrete directions included in the reconstructed ODF instead of only the directions corresponding to eigenvectors of a diffusion tensor fit (25). This is also true for MDA which was

estimated as

$$\text{MDA} = \frac{1-\mu}{\sqrt{1+2\mu^2}}, \text{ where } \mu = \left(\frac{\Psi_{\min}}{\Psi_{\max}}\right)^{2/3} \tag{4.1}$$

with $\Psi_{\min}$ and $\Psi_{\max}$ representing the smallest and largest directions sampled in the ODF.

4.2.2 Extracting MDA vectors

Each ODF $\Psi(\theta)$ was calculated with GQI on a set of 642 approximately-evenly spaced directions $\theta \in \Theta$ on a tesselated icosahedron. ODF magnitudes were rescaled so that the sum of each ODF is $\sum_{\theta\in\Theta} \Psi(\theta) = 1$. We then calculated the multi-directional anisotropy (MDA) value for each direction θ using Eq. 4.1, but with

$$\mu(\theta) = \left(\frac{\Psi(\theta)}{\Psi(\theta_{min})}\right)^{2/3} \tag{4.2}$$

MDA values and their corresponding directions were calculated for the four largest local maxima in every ODF, which are ordered by decreasing size. A separate study of the same data found that ODF peaks become very noisy after the 4th direction (6). The vector fields corresponding to the local maxima were warped into the group template using ANTs. 3D volumes containing the MDA magnitudes were also warped to the group template and used to scale the normalized vectors. White matter voxels were determined by segmenting the weighted average template of T1w, T2w, and GFA volumes in FreeSurfer (49). In this work we consider generative models for the largest two MDA vectors within each voxel.

4.3 Discovering population-wide salient regions

We propose jointly learning two functions, a patch-based classifier $f : \mathbb{R}^{P\times P\times P\times C} \to \mathbb{R}$ and a saliency map $s : \mathbb{R}^3 \to \mathbb{R}$ which scores how informative each spatial region of the brain is relative to others. During training, we uniformly sample a set $\mathcal{P}$ of $|\mathcal{P}| = N$ patches from the brain scan and minimize the cross-entropy between the s-weighted average of the classification of $\mathcal{P}$,

$$\mathcal{L} = H\left(y, \sum_{p\in\mathcal{P}} s_p f(p)\right) - \lambda \sum_{p\in\mathcal{P}} s_p \log s_p \tag{4.3}$$

where $H(y, \hat{y}) = -(y \log \hat{y} + (1-y)\log(1-\hat{y}))$ is the cross-entropy of the ground-truth label y and predicted $\hat{y}$, and

$$s_p = \frac{\exp s(\text{coord}(p)}{\sum_{p\in\mathcal{P}} \exp s(\text{coord}(p))} \tag{4.4}$$

the saliency function evaluated at the coordinates of a patch p, with a softmax activation over the sampled patches $\mathcal{P}$ which normalizes the saliency per iteration while emphasising the largest value. The second term of Eq. 4.3 regularizes the entropy of the learned saliency map with a positive weighting constant λ to be selective between brain regions, which has been shown to effectively sharpen attention mechanisms in neural networks (106).

This formulation simultaneously learns an interpretable saliency map common to the population under consideration, while directing the gradient updates to only relatively informative regions of the brain. Additionally, assuming that the expected volume of the brain being considered at every iteration (that is, the total expected volume of N patches of volume P^3 each is smaller than the total volume of the scan when discounting their overlap), this sampling based strategy has a regularizing effect which prevents overfit-

ting on the relatively few training samples present in MRI datasets similar to applying dropout (107).

During inference on a trained model, there is less consideration to being conservative with the number of sampled patches. With no backpropagation step the model occupies less memory footprint on a GPU, and sub-second inference times can still be achieved with large number of patches. As such, we sample $\mathcal{P}' \gg \mathcal{P}$ patches proportional to their s-score with replacement and compute the final classification as the unweighted arithmetic mean of the classifier applied to each patch.

4.3.1 Predicting biological sex

We apply this framework to predict the biological sex of individuals from their diffusion MRI scan. This serves as a convenient task as this splits the population in half, and is known to be associated with differences in white matter microstructure as imaged by diffusion MRI (105; 104).

For this task, we implemented the patch-classifier f as a CNN with receptive field equal to a selected patch size of 32^3, corresponding to a cube of size of 40mm to a side due to the 1.25mm spatial resolution of the diffusion volumes.. The CNN consisted of four convolutional layers (see Table 4.1) for tabulated information of the model's layers), consisting of 3D convolution with a leaky ReLU activation followed by 3D max pooling, batch normalization (78), and spatial dropout (77), followed by a single fully connected layer with sigmoid activation.

The saliency function s was implemented as a fully connected network consisting of three hidden layers with 16 neurons each with leaky ReLU activation, with the final layer outputting a single unconstrained scalar. See Table 4.2 for further specifications of the model. It receives as input only the spatial coordinates of a patch, normalized to

Layer	Specifications	Output size
3D convolution	16 filters, kernel size 4	$29 \times 29 \times 29 \times 16$
Leaky ReLU	0.1 negative slope	
3D max pooling	kernel size 2	$14 \times 14 \times 14 \times 16$
Batch normalization		
Spatial dropout	P = 0.1	
3D convolution	24 filters, kernel size 3	$12 \times 12 \times 12 \times 24$
Leaky ReLU	0.1 negative slope	
3D max pooling	kernel size 2	$6 \times 6 \times 6 \times 24$
Batch normalization		
Spatial dropout	P = 0.1	
3D convolution	32 filters, kernel size 3	$4 \times 4 \times 4 \times 32$
Leaky ReLU	0.1 negative slope	
3D max pooling	kernel size 2	$2 \times 2 \times 2 \times 32$
Batch normalization		
Spatial dropout	P = 0.1	
3D convolution	32 filters, kernel size 2	$1 \times 1 \times 1 \times 32$
Leaky ReLU	0.1 negative slope	
Batch normalization		
Spatial dropout	P = 0.1	
Fully connected	1 filter, sigmoid activation	1

Table 4.1: Specification of patch classifier model.

Layer	Specifications
Fully connected	16 filters
Leaky ReLU	P = 0.1
Fully connected	16 filters
Leaky ReLU	P = 0.1
Fully connected	16 filters
Leaky ReLU	P = 0.1
Fully connected	1 filter, no bias

Table 4.2: Specification of saliency model.

the range $[-1, 1]$, as well as each of those coordinates squared to further help localize. The weighting factor λ of the saliency regularization in Eq. 4.3 was set to 0.2. A very important consideration is how this saliency s is initialized, as it represents our prior belief for which regions are most predictive. In this work, we prefer an uninformative prior which we achieve by initializing each of the dense layers with small values drawn from a normal distribution with standard deviation of 0.2. This results in an initial saliency that is fairly uniform, with the ratio of largest to smallest values around 1.2.

The data was normalized to zero mean and unit standard deviation, based on the distribution of training data. During training, the data was augmented with Gaussian noise with standard deviation 5% that of the data. This provided a crucial regularization to prevent the saliency from overfitting on locations with little signal. To further prevent overfitting, label smoothing of 0.05 was applied during training (108).

The model was trained end-to-end with an Adam optimzer with learning rate set at 10^{-5}, and parameters $\beta_1 = 0.7$ and $\beta_2 = 0.9$. We selected $|\mathcal{P}| = 8$ patches at every iteration with a batch size of 32. Models were trained for 5000 epochs each, at which time both validation and training loss had plateaued. Training was performed on a single NVIDA GeForce 2080 GPU.

4.4 Results

We present the saliency of two independently trained models to demonstrate the variability of results that can be obtained. The learned saliency maps from two models is visualized in Fig. 4.2, which preferentially identify the superior portion of the brain stem. These saliency maps are sharply defined and have discovered sparse portions of the brain scan which are sufficient for the classification task. Care must be taken in interpreting these results, as the saliency maps highlight the only center points of the

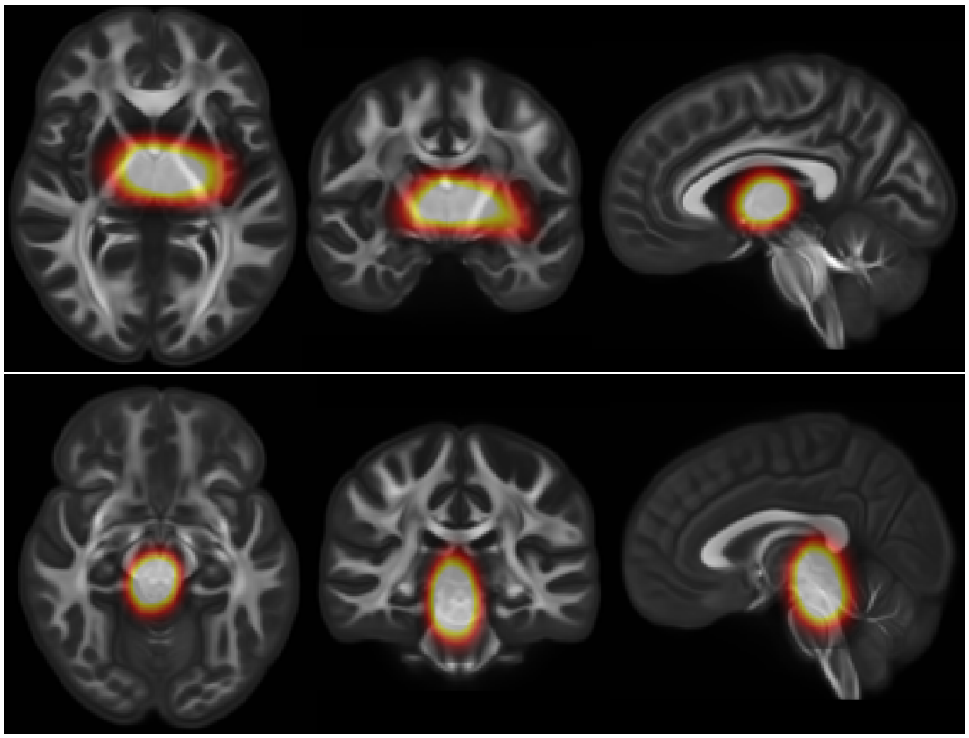


Figure 4.2: The learned saliency maps obtained from two independently trained models. A classification model considering only 40^3mm patches centered on these portions of the brain scan achieved 85.1% (top) and 78.7% (bottom) accuracy in predicting the biological sex of the 94 subjects in the held-out test set.

most informative patches as discovered by our framework. The receptive field of the CNN classifier f considers a volume 40mm to a side, so portions of the nearby brain structure is considered for the prediction.

To assess the performance of the overall model, we sampled $|\mathcal{P}'| = 10^{10}$ patches with replacement proportional to their saliency score. The classifier f was applied to each of these patches for every subject in the held-out test set, and the final prediction was the simple arithmetic mean across all those predictions per subject. This was repeated for five independently trained models to assess the variance in performance. These models ranged between 78.7% to 85.1% accuracy, with an average of 81.7%. The learned saliency was similar for each, with typical examples presented in Fig. 4.2.

4.5 Discussion

Our results demonstrate that voxels nearby to the superior portion of the brainstem as being preferentially discovered when classifying biological sex with a patch based classifier. This is consistent with prior literature, including that the thalamus, corpus callosum, and cingulum exhibit sex differences in their microstructure (109; 110; 111).

Our performance metrics are comparable to prior work in classifying sex from brain MRI scans, which typically achieves accuracy in sex prediction of 80-87% using functional MRI data solely (112; 113; 114), 83% when considering solely diffusion MRI (115), over 90% using structural T1- or T2-weighted images (116), and 96% considering morphological features derived from T2-weighted images (117).

In obtaining these results we initially had only an uninformative uniform prior of which brain regions to select, and the saliency was then learned jointly with the patch-based classifier. However, an interesting aspect of the proposed framework is that the saliency map s is independent of actual data, considering only the coordinates thereof. As such, it is trivial to select a more informative prior distribution by pre-training the saliency map s to preferentially guide the classifier towards those regions during optimization. For instance, this can be used to ignore certain portions of the brain to discover the 'second-most' informative regions of the brain by incorporating a prior that ignores where previous results have highlighted. This could also be used to ignore areas known to contain less signal or that are otherwise outside the brain, so as to prevent overfitting.

There is a limitation to the optimization function in Eq. 4.3, namely that singular regions will be defined as informative or not solely based on the predictive performance of that region in isolation. This precludes distant portions of the brain that are uncorrelated in the information they contain as pertains to the prediction that, when considered jointly, may outperform any given single region. This can be ameliorated by having the

patch-based classifier not output a prediction but instead a latent embedding vector for each region, allowing uncorrelated information from disparate regions of the brain to be modeled jointly. This would require a third component to the model which would output the final classification on the basis of the s-weighted average of embeddings. However, such an approach loses some of the explainability of the learned saliency maps, as it no longer assigns each region a pure score of its informativeness.

An underlying assumption to our framework is that there exists a spatially homogeneous region throughout the population that is sufficient for classification. This may not always hold, but the comparative success of a trained model and the salient regions it identifies can be used as to test the hypothesis of how well this assumption holds for given data.

4.6 Conclusion

In this chapter, we proposed a classification model for whole-brain scans which decomposes a CNN model into a patch-based classifier and an saliency map of which patches are most informative. This takes advantage of the spatial homogeneity present in MRI data, which is not present in natural images; with MRI data the same voxel will correspond to the same region of the brain for an entire population, given an accurate enough spatial normalization in particular. This allows for directly learning an explainable atlas of how informative brain regions are for classifying a given attribute of the subjects under consideration.

We apply our framework to the task of predicting the biological sex of diffusion MRI scans collected by the Human Connectome Project. Our results demonstrate that voxels nearby to the superior portion of the brain stem are preferentially selected as most informative. A patch-based classifier that was jointly trained with the saliency

map achieves 85.1% and 78.7% accuracy by two independent models when considering only those regions.

Chapter 5

Learning Cardiovascular Health Signatures

Bematists or *bematistae* were specialists in ancient Greece who were trained to measure distances by counting their steps.

Wikipedia

5.1 Introduction

When engaging in any physical task, the human body responds through a series of integrated changes in function that involves its physiologic systems, including the musculoskeletal, the cardiovascular, and the respiratory systems (118). Such responses may vary significantly due to environmental factors, yet when elicited in a controlled environment such as a 6-minute walk test carried out in lab settings, they allow inferring individual-specific physiological markers such as Resting Heart Rate (RHR), Maximal

Heart Rate, and Maximal Stroke Value. These markers are important in characterizing an individual's health and fitness status. For example, it is well known that cardio-respiratory fitness is inversely associated with all-cause mortality (119).

Recently, the advent and widespread adoption of wearable devices and fitness tracking apps (120) has enabled continuous, unobtrusive tracking of an individual behavior and physiological signals such as heart rate, physical activity, and sleep over time, with time resolution down to the minute-level and below. This has enabled population-scale physiological sensing (121).

In this chapter, we move beyond population-level aggregated sensing and set out to learn *individual* characteristics of cardiovascular responses by observing the relationship between behaviors such as sleep and physical activity and their associated changes in heart rate during the individuals everyday life. In absence of the controlled lab settings usually described in the physiology literature (118), we hypothesize that prolonged observation periods increase the likelihood of a behavior mimicking an in-lab test to spontaneously occur. For example, a brisk walk to the train station may be a good approximation of a 6-minute walk test. For this reason, we make use of attentioned models to pick up on such "natural experiments" that collectively contribute to shaping the envelope of an individual's cardiovascular response. In an analogy with control theory, we set out to learn the cardiovascular *transfer function* of an individual to capture the cardiac output for each possible (behavioral) input.

Though previous studies have leveraged representation learning to extract health-related features from wearable sensor data (122; 123), our work is unique in terms of dataset size (2.6×10^9 minutes of sensor measurements considered from 80k users over a span of one year), outputs (parsimonious individualized cardiovascular signatures output by attentioned convolutional autoencoders), and validation methods. We believe that accurately capturing cardiovascular response enables screening for abnormalities and de-

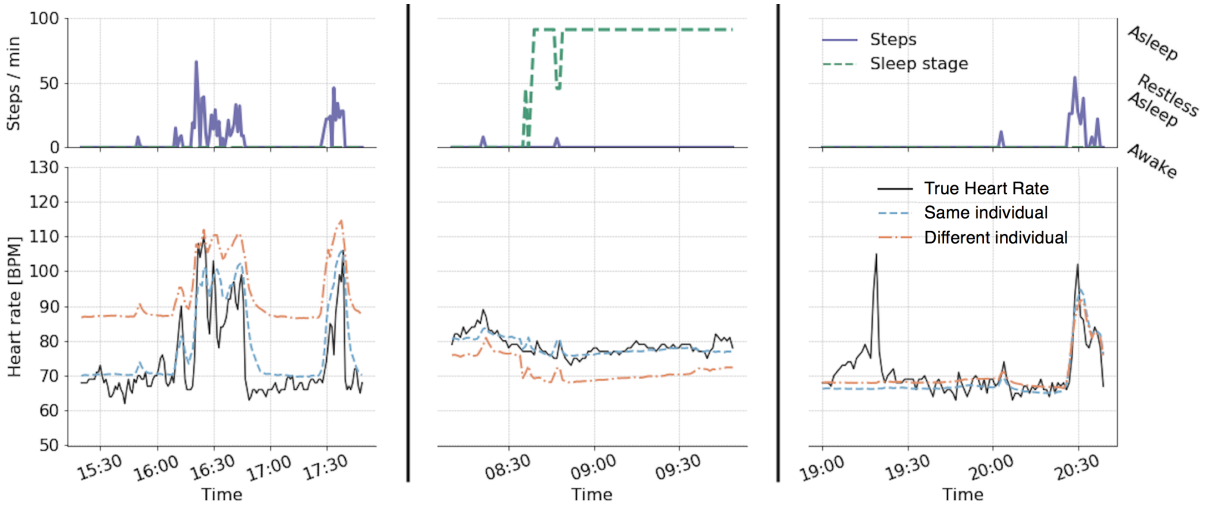


Figure 5.1: Example physical activity and sleep (upper row) and heart rate (bottom row) sensor data from three individuals, demonstrating how heart rate responds to onset of exercise (left column) and sleep (middle column). Changes in heart rate do not always occur due to physical activity (right column), with onset of anxiety or stress being potential unmeasured confounders. As expected, applying a signature from a different person (demonstrated in orange) results in increased reconstruction error.

tecting physiological changes over time unobtrusively in free living conditions.

5.2 Data

We select a cohort of 80,137 members of Achievement, a commercial reward platform. To be included in this study, users must have authorized sharing with Achievement of dense minute-level steps/sleep/heart rate activity logs from commercial activity trackers, such as Fitbit or Apple Watch. Following (124), to be included in the cohort a member must have at least 10 days worth of physical activity logs, with no more than 4 hours of unreported data per day, for one or both of the collection windows of January 2017 or 2018. Half of the members reported between 26,488 to 40,537 minutes per month, averaging 32,750 minutes. 82.8% of this cohort is female, with a median age and BMI of 31 and 28.3, respectively. All members with reported data in both of the two months

were assigned to the validation set (N=25,406). The remaining individuals were randomly assigned to either the training (N=43,784) or tuning sets (N=10,947).

The data from the activity trackers are minute-level measurements of a person's total step count and average heart rate, and if the wearer is asleep or restless asleep; see Figure 5.1 for sample data from three individuals. We scaled these measurements to the range $(0, 1)$ to speed up model training (125): the heart rate per-person is whitened to zero mean and unit standard deviation, and the step count values are log-transformed to handle the large spread of values as: $\text{steps}' = \log(\text{steps} + 1)/5$. The two sleep stages are encoded as separate binary channels. Missing data is imputed as mean heart rate of activity at awake, and no other data cleaning is performed.

5.3 Cardiovascular Signature Network

To learn a personalized cardiovascular response function, we consider a heart rate autoencoder (126) that is conditioned on the physical activity and sleep stages. The signature-encoder learns a signature of a person based on how their heart rate responds to physical activity, while the signature-decoder uses a learned signature to predict a person's heart rate based on their physical activity.

Encoder: The encoder model, as seen in Figure 5.2, learns a fixed-size signature from an arbitrarily length time-series. It consists of two WaveNet (127) convolutional neural network (CNN) blocks, W_1 and W_2, composed of seven dilated causal convolutional layers with residual connections and allow for modeling long-range temporal dependencies of up to 128 minutes, with 32 and 16 filters per layer, respectively. As opposed to recurrent layers, convolutions are typically faster to train especially when applied to very large sequences such as considered here. The encoder considers the physical activity channels and the heart rate signal separately in W_1 and W_2, which allows the encoder to jointly

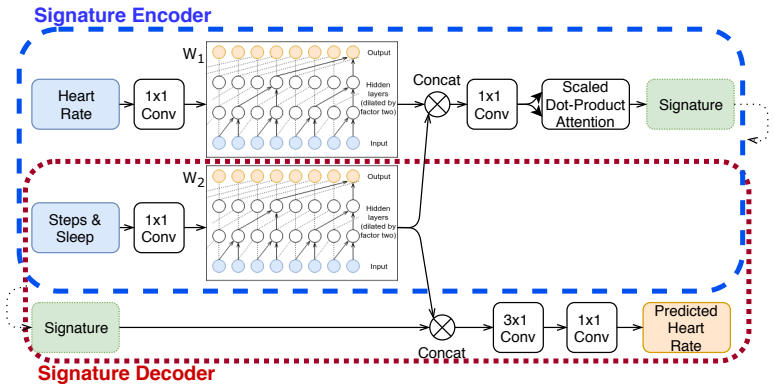


Figure 5.2: Diagram of proposed model architecture. The signature encoder predicts a cardiovascular signature from measured sensor data (top dashed box), and the signature decoder uses that same signature as well as physical activity data to predict the heart rate (bottom dotted box).

learn a latent physical activity representation with the decoder by sharing the weights of W_2. The outputs of W_1 and W_2 are concatenated together and a scaled dot-product attention mechanism (128) is applied to predict the cardiovascular signature with queries and keys of dimension $d_k = d_v = 8$ while the dimensionality of the values, d_v, is equal that of the signature size. Three separate convolutional layers of filter width 1 are applied to re-size the tensors appropriately.

Decoder: The decoder model consists of a single WaveNet block W_2, whose weights are tied to that of the encoder's, followed by two temporal convolutional layers. The output of W_2 at every time step is concatenated with a signature vector, and two temporal convolutional layers are then trained to predict the corresponding heart rate signal. The number of parameters unique to the decoder are kept to a minimum to force the signature to be as informative as possible.

Training: The two models are learned end to end by minimizing the average L_2 norm of the error in predicting heart rate, using the Adam optimizer (129) with default parameters ($\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-7}$). Missing data is imputed as mean heart rate of no activity at awake, though no loss is propagated corresponding

Table 5.1: Experimental results. The trained proposed model was validated on the 2017 data, and also using 2017 signatures applied to 2018 data. While varying signature sizes (results shown in left column) the full training set was used, and when varying training set size (results shown in right column) a size-32 signature was used.

Varying signature size			**Varying training set size**		
Size	**Validation error 2017 / 2018**		**# people**	**Validation error 2017 / 2018**	
4	0.295	0.385	500	0.319	0.400
8	0.291	0.385	2,000	0.306	0.391
16	0.283	0.394	5,000	0.298	**0.383**
32	0.279	0.393	20,000	0.285	0.387
64	0.288	**0.384**	43,784	**0.279**	0.393
128	**0.278**	0.395			

to these periods. The models are implemented in Keras (130) with a TensorFlow (131) backend. All hidden layers include ReLU activation functions (132), with the exception of the WaveNet blocks, which use gated activation units (127), and the output, which has no non-linearity. Training was done on mini-batches of size 16, for up to 30 epochs with an early stopping criteria if validation error was not observed to improve for five epochs.

5.4 Experimental results

Baseline models: We consider three baselines to compare our model to. The simplest predicts a persons mean heart rate at awake or asleep. The second uses XGBoost (133) with default parameters, trained on a single person to predict their heart rate based on the previous 120 minutes of physical activity. The third uses XGBoost again, but this time trained on a population of people rather than at the individual level. The performance of the baseline models can be seen in Table 5.2. Both XGBoost models are trained on the January 2017 activity window and validated on January 2018.

Table 5.2: The performance of the baseline models trained on 2017 data and validated on 2018 data.

Baseline model	**Validation error**
Awake/Sleep Mean	0.755
Individual XGBoost	0.445
Population XGBoost	0.539

Sensitivity analysis on signature size: We trained the proposed model with signature sizes of $\{4, 8, 16, 32, 64, 128\}$. As seen in Table 5.1, we observe that our model is robust to varying sizes of the cardiovascular signatures, with a decrease in validation error that levels off after a size of 16.

Effect of training set size: Our model leverages a population to learn a single persons cardiovascular transfer function. To understand the effect of the population on the model, we vary the training set size as fractions of the total (1%, 5%, 10%, 50%, 100%) and observe how well our model performs. As seen in Table 5.1, we observe a steady decrease in validation error as the training data is increased, culminating in a 14% better performance with a full dataset as opposed to only 1% of it.

Signature consistency: To assess test-retest reliability, a measure of *internal validity*, we consider how well a cardiovascular signature can be used to predict a persons heart rate from their physical activity a year later. For each individual in the validation set, we learn a signature from their signal measurements during January 2017 and apply that signature to predict their heart rate during January 2018. As compared to using a different persons signature, a person's own signature is significantly better at predicting their heart rate (Wilcoxon signed-rank test, $V = 2.6 \times 10^7$, $p < 10^{-16}$), with a median of 60% greater mean-square error when using another person's, randomly selected.

Predicting health conditions using signatures: To assess the *external validity* of the signatures, we tested whether they are associated with factors affecting car-

diovascular response, such as age and body mass index (BMI). We used an XGBoost model (133) trained on the learned size-32 validation signatures to predict if an individual is above/below median age of the cohort (31 years) with an AUC of 70.1% when trained on a random 70/30 split of the validation set. Predicting if a person is obese (BMI ≥ 30) from solely their signature achieves an AUC of 69.7%. Predicting the same outcomes using only an individual's resting heart rate results in significantly worse accuracy, with AUCs of 60.6% and 54.1%, respectively, demonstrating that signatures carry richer information about the relationship between physical activity and heart rate than the single RHR marker.

5.5 Discussion

It is informative to consider when a cardiovascular signature would *not* well predict a person's heart rate. Assuming the measuring conditions of the wearable device stay the same, this may happen when a person's cardiovascular response is hard to learn (e.g., short observation period, high missingness, or erratic behavior), when it changes (e.g., improvement/degradation of fitness), or when there are factors affecting HR that go beyond sleep and physical activity (e.g., stress endured during an interview, after taking medication, or having a meal). An example of where our model fails can be seen in the right-most column of Figure 5.1.

In future work we plan to explore the motifs surfaced by the attention component of the network, and study how they are related to health outcomes. From a methods perspective, future extensions will consider variational autoencoders to better condition the latent space of cardiovascular signatures as well as further hyper-parameter and architecture optimization.

Chapter 6

Conclusion

> Between stimulus and response there is a space. In that space is our power to choose our response. In our response lies our growth and our freedom.
>
> Viktor Frankl

This dissertation demonstrated several novel computational techniques to further characterize health data. Statistical and machine learning models were proposed that exploited key properties of the datasets to derive further insights from them. We considered two sources of data, namely diffusion MR images of the human brain and large-scale physical activity data from wearable sensors, where each had its own set of challenges. Each of these data sources holds enormous potential for bettering our health outcomes. Diffusion MRI has enabled in vivo study of the structural connectivity of the human brain and has been a catalyst for rapid advancements in our understanding of our neuroanatomy. Wearable activity trackers allow for massive scale study of our physical and mental health through long long periods of time. Deriving value and delivering impact

from these sets of data is challenging. Testing domain hypothesis on them at scale requires new computational approaches, and delivering these insights and models from a laboratory setting requires addressing rightful concerns about explanatory power.

In Chapter 2, we demonstrated how a novel distance metric for dyads of neighboring voxels between different scans can be built upon with a statistical framework to discover large regions conserved with a population under study. This methodology was applied on a population of 109 twins, and comparing them to a matched set of pairs of strangers (namely, the same individuals assigned to different pairs) we discovered nearly 4% of white matter as being associated with genetic similarity, and that this was primarily within deep white matter.

In Chapters 3 and 4, we considered two dual problems for characterising white matter, namely how informative attributes of a subject are for generating regions of diffusion imaged white matter (Chapter 3), and how informative regions of white matter are for predicting attributes of a person (Chapter 4). These last two problems were approached using newly developed methods building on recent literature from computer vision, generative modeling, and Bayesian uncertainty quantification as applied to deep neural networks. These methods allow for associating regions of white matter with traits and characteristics of interest across a population.

Lastly, in Chapter 5 we considered the promise of digital health devices for large-scale monitoring of cardiovascular health, developing a model that can learn from the few interesting and salient events that may occur in daily living conditions to predict a person's cardiovascular response from their physical activity. We demonstrated how such response functions can be meaningfully used to predict variables associated with cardiovascular health, which holds the promise of longitudinal tracking of cardiovascular health across large populations and enabling predictive interventions.

Throughout this dissertation, data-driven methodologies have been developed to meet

the constraints imposed by the sensor data to best develop an understanding of aspects relating to human health. A key focus has been on developing methods that do not solely improve on singular performance metrics, but also further characterize the data. These are steps towards bringing the superlative performance we have seen in the past decade of predictive modeling towards domains where erroneous data-driven decisions can not be accepted. There is an enormous potential for bettering our health and daily living, and we are called upon to address that.

Appendix A

Voxel-wise Population Differences

The results presented in Section 2.3 follow from the application of Eq. 2.5, which is an extension of Eq. 2.3 to two neighboring voxels. However, Eq. 2.3 is of independent interest and can be used to show similar results as Eq. 2.5.

Of the nearly one million white matter voxels we identify 33,003 voxels as significantly more similar within monozygotic (MZ) and dizygotic (DZ) twins than a matched control group of strangers ($p < 10^{-3}$, false discovery rate 1.3%) using Eq. 2.3 with $k = 1$ MDA peaks. These voxels are visualized in figure A.1. Most of these voxels, 23,566, overlap with the previous results in Section 2.3.

The voxels identified were further used to study similarities within MZ twins as compared to DZ twins and also in siblings as compared to a matched control group of strangers. Of the 33.0k voxels discovered to be more similar within twins than strangers we found 8,746 voxels that were significantly more similar within MZ than in DZ twins ($p < .05$, FDR 8.7%), and 5,244 voxels that were significantly more similar in siblings than in strangers ($p < .05$, FDR 19.2%). These two results overlapped to a small extent, or a total of 1,637 voxels.

We limit the null hypothesis space to only those 33.0k voxels discovered to be sig-

nificant previously as a whole-brain study did not identify any reasonable result with sufficiently low false discovery rate. We note that due to the reduced hypothesis space these results still assume a single model of similarity associated with genetically related pairs. We further note that these results are associated with much less statistical significance, indicating that this model does not generalize as well as that of Eq. 2.5 as seen in Fig. 2.10. In addition, the small intersection of the results is evidence for a need to model heterogeneity within the population, as mentioned in Section 2.4.

The microstructural orientation of the MDA distributions is important to distinguish between twins and the control population of strangers; the magnitude by itself only accounts for a small portion of these results. Repeating the experiment with a modification to Eq. 2.3 such that it only considers magnitude and not direction discovers 11,948 voxels as significantly more similar within MZ and DZ twins than in the control group ($p < 10^{-3}$, FDR 4.3%), of which only 6,235 were also identified as significantly similar in the previous test which included orientation data. These magnitude-only results are clustered in the corpus callosum and near the amygdalae.

Comparing the voxels identified as significant when applying Eq. 2.3 to Eq. 2.5 is complicated by the different number of hypothesis when considering voxels and voxel dyads in the two equations, respectively. An alternative middle ground between the two that considers within-voxel differences between subjects, as Eq. 2.3 does, but for voxel dyads, as Eq. 2.5 considers:

$$d(X,Y,u,v)=\frac{1}{2}\sum_{i=1}^{k}\Big(\min\left(\left\|X_u^i-Y_u^i\right\|,\left\|X_u^i+Y_u^i\right\|\right) + \min\left(\left\|X_v^i-Y_v^i\right\|,\left\|X_v^i+Y_v^i\right\|\right)\Big) \tag{A.1}$$

This differs from Eq. 2.5 by only comparing subjects X and Y within voxels u and v,

instead of across voxels. This relaxes the spatial coherence constraint while including up to 26 null hypothesis per voxel. Indeed, repeating the experiments of Section 2.3 with $k = 1$ peaks identifies 330,882 voxels dyads containing 84,147 unique voxels as significant ($p < 10^{-4}$, FDR 0.3%). These voxels are visualized in Fig. A.2, and nearly entirely encompass the previous results in Section 2.3 occurring in the same regions but taking larger extent.

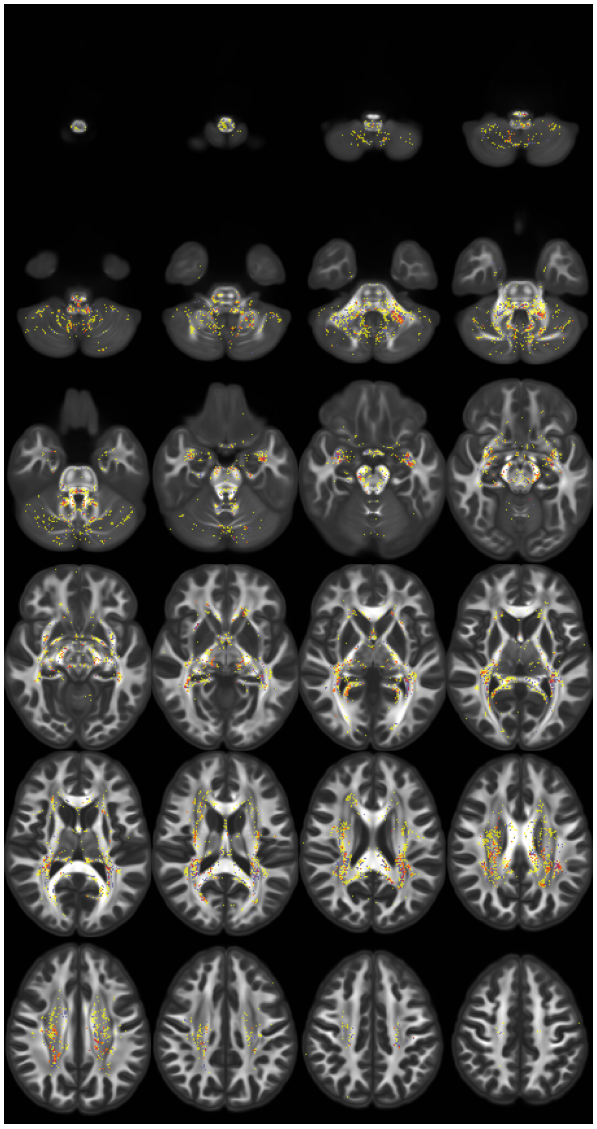


Figure A.1: Axial slices of voxels identified as significantly more similar within twins than strangers (yellow), and the subset of those voxels found to be significantly more similar within monozygotic compared to dizygotic twins (teal) and in siblings compared to strangers (orange), as computed using Eq. 2.3. Background image is of a population averaged Generalized Fractional Anisotrophy (GFA), where lighter regions indicate higher GFA values. Image created partly using ITK-SNAP (4).

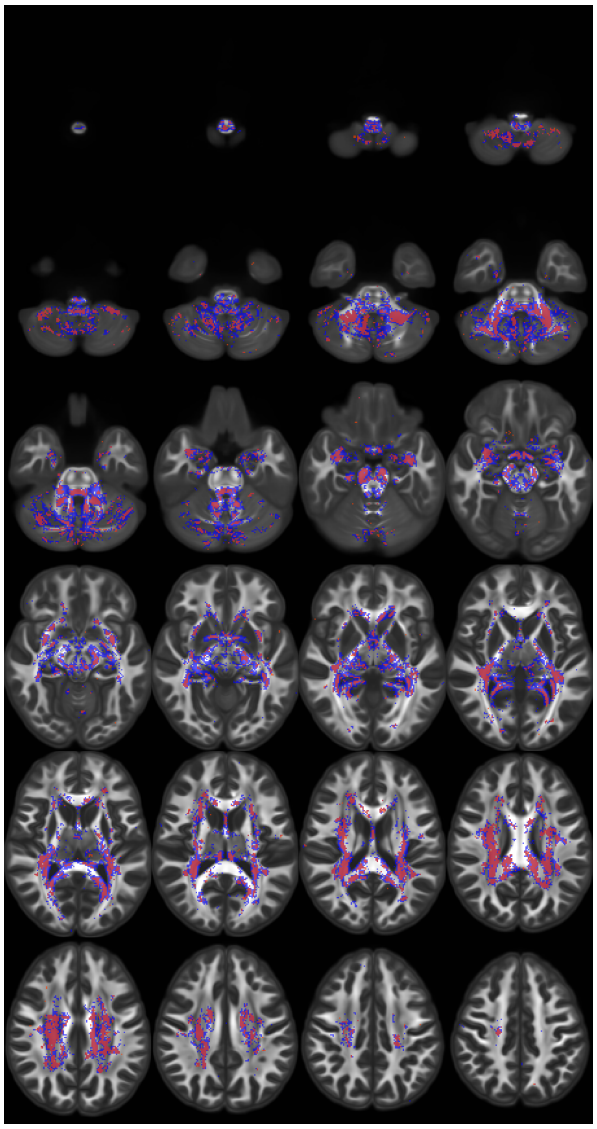


Figure A.2: Axial slices of voxels identified as significantly more similar within monozygotic and dizygotic twins than strangers as computed using Eq. A.1 (blue) and Eq. 2.5 (red). Background image is of a population averaged Generalized Fractional Anisotrophy (GFA), where lighter regions indicate higher GFA values. Image created partly using ITK-SNAP (4).

Appendix B

Controlling for Morphological Similarity

In Chapter 2, we seek to identify regions whose similarity between MZ and DZ twins is attributable to oriented white matter microstructure and not simply due to the morphology of the brain or systemic registration misalignment (50). Brain morphology is known to be heritable (134; 135). To that end we identify voxels in which twins have significantly more similar log-jacobian values than strangers do and exclude them from the analysis in this paper. The log-jacobian value of a voxel measures how much this voxel was expanded or contorted from a subject's native space to the normalized space in which the population analysis is performed in.

We define dissimilarity between a pair of subjects with respect to their log-jacobian values as their absolute difference. We compute a distribution of dissimilarities per voxel for twins and strangers and perform a Mann-Whitney U test (51) to test if MZ and DZ twins are significantly more similar than strangers in a given voxel. We identified 3.0k voxels that fit this criteria ($p < 10^{-3}$, FDR 28.9%) using a conservative threshold. Figure B.1 shows an overview of these voxels. Some of the regions identified as such have

previously been reported to have heritable anatomical structure (134).

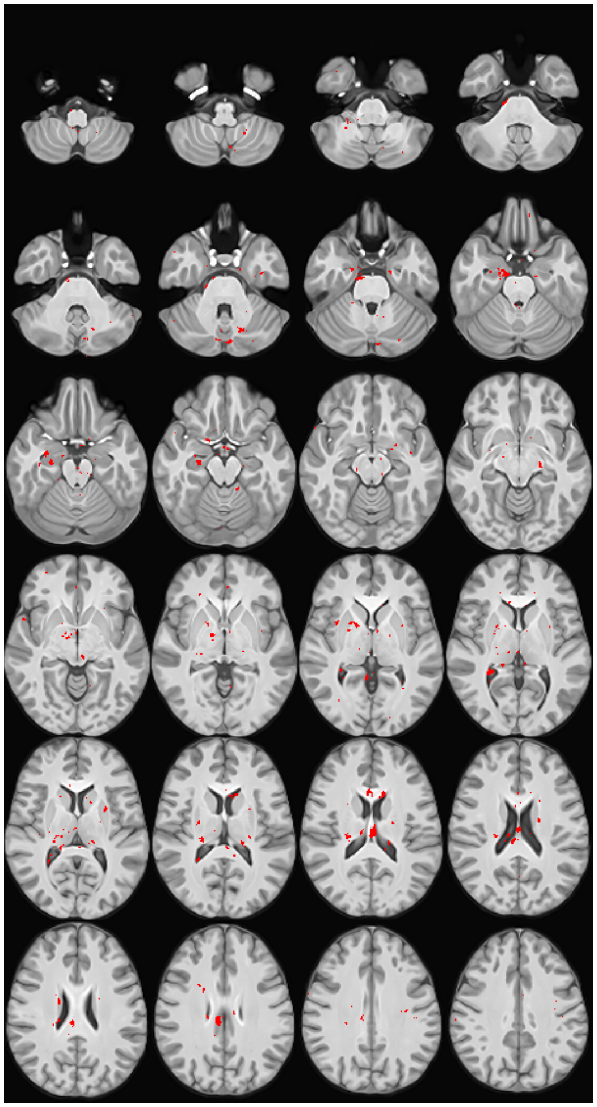


Figure B.1: Axial slices of voxels in red whose log-jacobian values from the registration process are found to be significantly more similar within monozygotic and dizygotic twins than strangers, suggesting possible morphological similarity. Background image is a population averaged T1 weighted MRI image. Image created partly using ITK-SNAP (4).

Bibliography

[1] H. T. Hallgrímsson, M. Cieslak, L. Foschini, S. T. Grafton, and A. K. Singh, *Spatial coherence of oriented white matter microstructure: Applications to white matter regions associated with genetic similarity*, *NeuroImage* **172** (2018) 390–403.

[2] H. T. Hallgrimsson, R. Sharan, S. T. Grafton, and A. K. Singh, *Estimating localized complexity of white-matter wiring with gans*, *Medical Imaging meets NeurIPS (MED-NeurIPS)* (2019).

[3] H. T. Hallgrímsson, F. Jankovic, T. Althoff, and L. Foschini, *Learning individualized cardiovascular responses from large-scale wearable sensors data*, *Machine Learning for Health workshop at NeurIPS* (2018).

[4] P. A. Yushkevich, J. Piven, H. Cody Hazlett, R. Gimpel Smith, S. Ho, J. C. Gee, and G. Gerig, *User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability*, *Neuroimage* **31** (2006), no. 3 1116–1128.

[5] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka, *et. al.*, *3d slicer as an image computing platform for the quantitative imaging network*, *Magnetic resonance imaging* **30** (2012), no. 9 1323–1341.

[6] L. J. Volz, M. Cieslak, and S. Grafton, *A probabilistic atlas of fiber crossings for variability reduction of anisotropy measures*, *Brain Structure and Function* **223** (2018), no. 2 635–651.

[7] F.-C. Yeh, P.-F. Tang, and W.-Y. I. Tseng, *Diffusion mri connectometry automatically reveals affected fiber pathways in individuals with chronic stroke*, *Neuroimage: Clinical* **2** (2013) 912–921.

[8] B. Baird, M. Cieslak, J. Smallwood, S. T. Grafton, and J. W. Schooler, *Regional white matter variation associated with domain-specific metacognitive accuracy*, *Journal of cognitive neuroscience* (2015).

[9] E. Bullmore and O. Sporns, *Complex brain networks: graph theoretical analysis of structural and functional systems*, *Nature reviews. Neuroscience* **10** (2009), no. 3 186.

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

[11] S. Jbabdi and H. Johansen-Berg, *Tractography: where do we go from here?*, *Brain connectivity* **1** (2011), no. 3 169–183.

[12] K. Maier-Hein, P. Neher, J.-C. Houde, M.-A. Cote, E. Garyfallidis, J. Zhong, M. Chamberland, F.-C. Yeh, Y. C. Lin, Q. Ji, *et. al.*, *Tractography-based connectomes are dominated by false-positive connections*, *bioRxiv* (2016) 084137.

[13] E. T. Bullmore and D. S. Bassett, *Brain graphs: graphical models of the human brain connectome*, *Annual review of clinical psychology* **7** (2011) 113–140.

[14] P. Hagmann, O. Sporns, N. Madan, L. Cammoun, R. Pienaar, V. J. Wedeen, R. Meuli, J.-P. Thiran, and P. Grant, *White matter maturation reshapes structural connectivity in the late developing human brain*, *Proceedings of the National Academy of Sciences* **107** (2010), no. 44 19067–19072.

[15] I. Deary, M. Bastin, A. Pattie, J. Clayden, L. J. Whalley, J. Starr, and J. Wardlaw, *White matter integrity and cognition in childhood and old age*, *Neurology* **66** (2006), no. 4 505–512.

[16] G. R. Poudel, J. C. Stout, L. Salmon, A. Churchyard, P. Chua, N. Georgiou-Karistianis, G. F. Egan, *et. al.*, *White matter connectivity reflects clinical and cognitive status in huntington's disease*, *Neurobiology of disease* **65** (2014) 180–187.

[17] N. G. Papadakis, D. Xing, G. C. Houston, J. M. Smith, M. I. Smith, M. F. James, A. A. Parsons, C. L.-H. Huang, L. D. Hall, and T. A. Carpenter, *A study of rotationally invariant and symmetric indices of diffusion anisotropy*, *Magnetic resonance imaging* **17** (1999), no. 6 881–892.

[18] U. C. Wieshmann, C. A. Clark, M. R. Symms, F. Franconi, G. J. Barker, and S. D. Shorvon, *Reduced anisotropy of water diffusion in structural cerebral abnormalities demonstrated with diffusion tensor imaging*, *Magnetic resonance imaging* **17** (1999), no. 9 1269–1274.

[19] M. Filippi, M. Cercignani, M. Inglese, M. Horsfield, and G. Comi, *Diffusion tensor magnetic resonance imaging in multiple sclerosis*, *Neurology* **56** (2001), no. 3 304–311.

[20] R. A. Kanaan, J.-S. Kim, W. E. Kaufmann, G. D. Pearlson, G. J. Barker, and P. K. McGuire, *Diffusion tensor imaging in schizophrenia*, *Biological psychiatry* **58** (2005), no. 12 921–929.

[21] D. J. Werring, A. T. Toosy, C. A. Clark, G. J. Parker, G. J. Barker, D. H. Miller, and A. J. Thompson, *Diffusion tensor imaging can detect and quantify corticospinal tract degeneration after stroke*, *Journal of Neurology, Neurosurgery & Psychiatry* **69** (2000), no. 2 269–272.

[22] B. P. Witwer, R. Moftakhar, K. M. Hasan, P. Deshmukh, V. Haughton, A. Field, K. Arfanakis, J. Noyes, C. H. Moritz, M. E. Meyerand, *et. al.*, *Diffusion-tensor imaging of white matter tracts in patients with cerebral neoplasm*, *Journal of neurosurgery* **97** (2002), no. 3 568–575.

[23] L. J. Volz, M. Cieslak, and S. T. Grafton, *A probabilistic atlas of fiber crossings for variability reduction of anisotropy measures*, *Brain Structure and Function* (Sep, 2017).

[24] V. J. Wedeen, P. Hagmann, W.-Y. I. Tseng, T. G. Reese, and R. M. Weisskoff, *Mapping complex tissue architecture with diffusion spectrum magnetic resonance imaging*, *Magnetic resonance in medicine* **54** (2005), no. 6 1377–1386.

[25] D. S. Tuch, *Q-ball imaging*, *Magnetic resonance in medicine* **52** (2004), no. 6 1358–1372.

[26] E. T. Tan, L. Marinelli, J. I. Sperl, M. I. Menzel, and C. J. Hardy, *Multi-directional anisotropy from diffusion orientation distribution functions*, *Journal of Magnetic Resonance Imaging* **41** (2015), no. 3 841–850.

[27] C. A. Greene, M. Cieslak, and S. T. Grafton, *Effect of different spatial normalization approaches on tractography and structural brain networks*, *Network Neuroscience* (2017).

[28] D. Christiaens, T. Dhollander, F. Maes, S. Sunaert, and P. Suetens, *The effect of reorientation of the fibre orientation distribution on fibre tracking*, in *Proceedings CDMRI 2012*, pp. 33–44, 2012.

[29] M. F. Glasser, T. S. Coalson, E. C. Robinson, C. D. Hacker, J. Harwell, E. Yacoub, K. Ugurbil, J. Andersson, C. F. Beckmann, M. Jenkinson, *et. al.*, *A multi-modal parcellation of human cerebral cortex*, *Nature* **536** (2016), no. 7615 171–178.

[30] D. Zhu, D. Zhang, C. Faraco, K. Li, F. Deng, H. Chen, X. Jiang, L. Guo, L. Miller, and T. Liu, *Discovering dense and consistent landmarks in the brain*, in *Information Processing in Medical Imaging*, pp. 97–110, Springer, 2011.

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

[32] D. Posthuma, E. De Geus, M. Neale, H. H. Pol, W. Baare, R. Kahn, and D. Boomsma, *Multivariate genetic analysis of brain structure in an extended twin design*, *Behavior genetics* **30** (2000), no. 4 311–319.

[33] J. S. Peper, R. M. Brouwer, D. I. Boomsma, R. S. Kahn, H. Pol, and E. Hilleke, *Genetic influences on human brain structure: a review of brain imaging studies in twins*, *Human brain mapping* **28** (2007), no. 6 464–473.

[34] Ö. de Manzano and F. Ullén, *Same genes, different brains: Neuroanatomical differences between monozygotic twins discordant for musical training*, *Cerebral Cortex* **28** (2017), no. 1 387–394.

[35] K.-K. Shen, S. Rose, J. Fripp, K. L. McMahon, G. I. de Zubicaray, N. G. Martin, P. M. Thompson, M. J. Wright, and O. Salvado, *Investigating brain connectivity heritability in a twin study using diffusion imaging data*, *NeuroImage* **100** (2014) 628–641.

[36] I. S. Häberling, G. Badzakova-Trajkov, and M. C. Corballis, *Asymmetries of the arcuate fasciculus in monozygotic twins: genetic and nongenetic influences*, *PloS one* **8** (2013), no. 1 e52315.

[37] N. Jahanshad, A. D. Lee, M. Barysheva, K. L. McMahon, G. I. de Zubicaray, N. G. Martin, M. J. Wright, A. W. Toga, and P. M. Thompson, *Genetic influences on brain asymmetry: a dti study of 374 twins and siblings*, *Neuroimage* **52** (2010), no. 2 455–469.

[38] P.-T. Yap, Y. Fan, Y. Chen, J. H. Gilmore, W. Lin, and D. Shen, *Development trends of white matter connectivity in the first years of life*, *PloS one* **6** (2011), no. 9 e24678.

[39] P. M. Thompson, T. D. Cannon, K. L. Narr, T. Van Erp, V.-P. Poutanen, M. Huttunen, J. Lönnqvist, C.-G. Standertskjöld-Nordenstam, J. Kaprio, M. Khaledy, *et. al.*, *Genetic influences on brain structure*, *Nature neuroscience* **4** (2001), no. 12 1253.

[40] N. Sadeghi, J. H. Gilmore, and G. Gerig, *Twin–singleton developmental study of brain white matter anatomy*, *Human brain mapping* **38** (2017), no. 2 1009–1024.

[41] M.-C. Chiang, K. L. McMahon, G. I. de Zubicaray, N. G. Martin, I. Hickie, A. W. Toga, M. J. Wright, and P. M. Thompson, *Genetics of white matter development: a dti study of 705 twins and their siblings aged 12 to 29*, *Neuroimage* **54** (2011), no. 3 2308–2317.

[42] D. C. Van Essen and K. Ugurbil, *The future of the human connectome*, *NeuroImage* **62** (2012), no. 2 1299–1310.

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

[44] D. A. Feinberg, S. Moeller, S. M. Smith, E. Auerbach, S. Ramanna, M. F. Glasser, K. L. Miller, K. Ugurbil, and E. Yacoub, *Multiplexed echo planar imaging for sub-second whole brain fmri and fast diffusion imaging*, *PloS one* **5** (2010), no. 12 e15710.

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

[46] F.-C. Yeh, V. J. Wedeen, and W.-Y. I. Tseng, *Generalized q-Sampling Imaging*, *IEEE Transactions on Medical Imaging* **29** (2010), no. 9 1626–1635.

[47] B. B. Avants, P. Yushkevich, J. Pluta, D. Minkoff, M. Korczykowski, J. Detre, and J. C. Gee, *The optimal template effect in hippocampus studies of diseased populations*, *Neuroimage* **49** (2010), no. 3 2457–2466.

[48] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, *Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain*, *Medical image analysis* **12** (2008), no. 1 26–41.

[49] B. Fischl, D. H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. van der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, A. Montillo, N. Makris, B. Rosen, and A. M. Dale, *Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain*, *Neuron* **33** (2002) 341–355.

[50] J. Kim, B. Avants, S. Patel, J. Whyte, B. H. Coslett, J. Pluta, J. A. Detre, and J. C. Gee, *Structural consequences of diffuse traumatic brain injury: a large deformation tensor-based morphometry study*, *Neuroimage* **39** (2008), no. 3 1014–1026.

[51] H. B. Mann and D. R. Whitney, *On a test of whether one of two random variables is stochastically larger than the other*, *The annals of mathematical statistics* (1947) 50–60.

[52] Y. Benjamini and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, *Journal of the royal statistical society. Series B (Methodological)* (1995) 289–300.

[53] C. R. Genovese, N. A. Lazar, and T. Nichols, *Thresholding of statistical maps in functional neuroimaging using the false discovery rate*, *Neuroimage* **15** (2002), no. 4 870–878.

[54] D. C. Van Essen, *A tension-based theory of morphogenesis and compact wiring in the central nervous system*, *Nature* **385** (1997), no. 6614 313.

[55] F.-C. Yeh, J. M. Vettel, A. Singh, B. Poczos, S. T. Grafton, K. I. Erickson, W.-Y. I. Tseng, and T. D. Verstynen, *Quantifying differences and similarities in whole-brain white matter architecture using local connectome fingerprints*, *PLoS computational biology* **12** (2016), no. 11 e1005203.

[56] G. Koch, R. Zemel, and R. Salakhutdinov, *Siamese neural networks for one-shot image recognition*, in *ICML Deep Learning Workshop*, vol. 2, 2015.

[57] S. I. Ktena, S. Parisot, E. Ferrante, M. Rajchl, M. Lee, B. Glocker, and D. Rueckert, *Distance metric learning using graph convolutional networks: Application to functional brain networks*, *arXiv preprint arXiv:1703.02161* (2017).

[58] K. Uğurbil, J. Xu, E. J. Auerbach, S. Moeller, A. T. Vu, J. M. Duarte-Carvajalino, C. Lenglet, X. Wu, S. Schmitter, P. F. Van de Moortele, *et. al.*, *Pushing spatial and temporal resolution for functional and diffusion mri in the human connectome project*, *Neuroimage* **80** (2013) 80–104.

[59] M. E. Moseley, Y. Cohen, J. Kucharczyk, J. Mintorovitch, H. Asgari, M. Wendland, J. Tsuruda, and D. Norman, *Diffusion-weighted mr imaging of anisotropic water diffusion in cat central nervous system.*, *Radiology* **176** (1990), no. 2 439–445.

[60] P. J. Basser, J. Mattiello, and D. LeBihan, *Mr diffusion tensor spectroscopy and imaging*, *Biophysical journal* **66** (1994), no. 1 259–267.

[61] R. F. Betzel and D. S. Bassett, *Generative models for network neuroscience: prospects and promise*, *Journal of The Royal Society Interface* **14** (2017), no. 136 20170623.

[62] E. Tang, C. Giusti, G. L. Baum, S. Gu, E. Pollock, A. E. Kahn, D. R. Roalf, T. M. Moore, K. Ruparel, R. C. Gur, *et. al.*, *Developmental increases in white matter network controllability support a growing diversity of brain dynamics*, *Nature communications* **8** (2017), no. 1 1–16.

[63] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial nets*, in *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[64] M. Arjovsky, S. Chintala, and L. Bottou, *Wasserstein generative adversarial networks*, in *International Conference on Machine Learning*, pp. 214–223, 2017.

[65] A. Kendall and Y. Gal, *What uncertainties do we need in bayesian deep learning for computer vision?*, in *Advances in neural information processing systems*, pp. 5574–5584, 2017.

[66] T. Karras, T. Aila, S. Laine, and J. Lehtinen, *Progressive growing of gans for improved quality, stability, and variation*, in *International Conference on Learning Representations*, 2018.

[67] Z. C. Lipton, *The mythos of model interpretability*, *Queue* **16** (2018), no. 3 31–57.

[68] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, *The secret sharer: Evaluating and testing unintended memorization in neural networks*, in *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pp. 267–284, 2019.

[69] K. Armanious, Y. Mecky, S. Gatidis, and B. Yang, *Adversarial inpainting of medical image modalities*, in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3267–3271, IEEE, 2019.

[70] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, *Synthetic data augmentation using gan for improved liver lesion classification*, in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pp. 289–293, IEEE, 2018.

[71] R. Tanno, D. E. Worrall, A. Ghosh, E. Kaden, S. N. Sotiropoulos, A. Criminisi, and D. C. Alexander, *Bayesian image quality transfer with cnns: exploring uncertainty in dmri super-resolution*, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 611–619, Springer, 2017.

[72] Y. Kwon, J.-H. Won, B. J. Kim, and M. C. Paik, *Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation*, *Computational Statistics & Data Analysis* **142** (2020) 106816.

[73] R. F. Betzel, A. Avena-Koenigsberger, J. Goñi, Y. He, M. A. De Reus, A. Griffa, P. E. Vértes, B. Mišic, J.-P. Thiran, P. Hagmann, *et. al.*, *Generative models of the human connectome*, *Neuroimage* **124** (2016) 1054–1064.

[74] J. A. Roberts, A. Perry, A. R. Lord, G. Roberts, P. B. Mitchell, R. E. Smith, F. Calamante, and M. Breakspear, *The contribution of geometry to the human connectome*, *Neuroimage* **124** (2016) 379–393.

[75] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, *Generative image inpainting with contextual attention*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5505–5514, 2018.

[76] A. L. Maas, A. Y. Hannun, and A. Y. Ng, *Rectifier nonlinearities improve neural network acoustic models*, in *Proc. icml*, vol. 30, p. 3, 2013.

[77] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, *Efficient object localization using convolutional networks*, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 648–656, 2015.

[78] S. Ioffe and C. Szegedy, *Batch normalization: Accelerating deep network training by reducing internal covariate shift*, in *International Conference on Machine Learning*, pp. 448–456, 2015.

[79] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, *Improved training of wasserstein gans*, in *Advances in neural information processing systems*, pp. 5767–5777, 2017.

[80] N. H. Stricker, B. Schweinsburg, L. Delano-Wood, C. E. Wierenga, K. J. Bangen, K. Haaland, L. R. Frank, D. P. Salmon, and M. W. Bondi, *Decreased white matter integrity in late-myelinating fiber pathways in alzheimer's disease supports retrogenesis*, *Neuroimage* **45** (2009), no. 1 10–16.

[81] D. K. Jones, M. Catani, C. Pierpaoli, S. J. Reeves, S. S. Shergill, M. O'Sullivan, P. Golesworthy, P. McGuire, M. A. Horsfield, A. Simmons, *et. al.*, *Age effects on diffusion tensor magnetic resonance imaging tractography measures of frontal cortex connections in schizophrenia*, *Human brain mapping* **27** (2006), no. 3 230–238.

[82] Y. Lu, X. Li, D. Geng, N. Mei, P.-Y. Wu, C.-C. Huang, T. Jia, Y. Zhao, D. Wang, A. Xiao, *et. al.*, *Cerebral micro-structural changes in covid-19 patients–an mri-based 3-month follow-up study*, *EClinicalMedicine* (2020) 100484.

[83] B. Jeurissen, A. Leemans, J.-D. Tournier, D. K. Jones, and J. Sijbers, *Investigating the prevalence of complex fiber configurations in white matter tissue with diffusion magnetic resonance imaging*, *Human brain mapping* **34** (2013), no. 11 2747–2766.

[84] F. Szczepankiewicz, S. Lasič, D. van Westen, P. C. Sundgren, E. Englund, C.-F. Westin, F. Ståhlberg, J. Lätt, D. Topgaard, and M. Nilsson, *Quantification of microscopic diffusion anisotropy disentangles effects of orientation dispersion from microstructure: applications in healthy volunteers and in brain tumors*, *NeuroImage* **104** (2015) 241–252.

[85] L. Torrey and J. Shavlik, *Transfer learning*, in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pp. 242–264. IGI global, 2010.

[86] S. U. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Çukur, *Image synthesis in multi-contrast mri with conditional generative adversarial networks*, *IEEE transactions on medical imaging* **38** (2019), no. 10 2375–2388.

[87] Y. Lei, J. Harms, T. Wang, Y. Liu, H.-K. Shu, A. B. Jani, W. J. Curran, H. Mao, T. Liu, and X. Yang, *Mri-only based synthetic ct generation using dense cycle consistent generative adversarial networks*, *Medical physics* **46** (2019), no. 8 3565–3581.

[88] T. T. Cai, T. Liang, and H. H. Zhou, *Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional gaussian distributions*, *Journal of Multivariate Analysis* **137** (2015) 161–172.

[89] R. Bellman, *Dynamic programming*, *Science* **153** (1966), no. 3731 34–37.

[90] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[91] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, *et. al.*, *Mastering the game of go without human knowledge*, *nature* **550** (2017), no. 7676 354–359.

[92] P. Sturmfels, S. Rutherford, M. Angstadt, M. Peterson, C. Sripada, and J. Wiens, *A domain guided cnn architecture for predicting age from structural brain images*, *arXiv preprint arXiv:1808.04362* (2018).

[93] J. H. Cole, R. P. Poudel, D. Tsagkrasoulis, M. W. Caan, C. Steves, T. D. Spector, and G. Montana, *Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker*, *NeuroImage* **163** (2017) 115–124.

[94] S. Qiu, P. S. Joshi, M. I. Miller, C. Xue, X. Zhou, C. Karjadi, G. H. Chang, A. S. Joshi, B. Dwyer, S. Zhu, *et. al.*, *Development and validation of an interpretable deep learning framework for alzheimer's disease classification*, *Brain* (2020).

[95] R. Anirudh and J. J. Thiagarajan, *Bootstrapping graph convolutional neural networks for autism spectrum disorder classification*, in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3197–3201, 2019.

[96] M. Ghafoorian, N. Karssemeijer, T. Heskes, M. Bergkamp, J. Wissink, J. Obels, K. Keizer, F.-E. de Leeuw, B. van Ginneken, E. Marchiori, *et. al.*, *Deep multi-scale location-aware 3d convolutional neural networks for automated detection of lacunes of presumed vascular origin*, *NeuroImage: Clinical* **14** (2017) 391–399.

[97] M. Lyksborg, O. Puonti, M. Agn, and R. Larsen, *An ensemble of 2d convolutional neural networks for tumor segmentation*, in *Scandinavian Conference on Image Analysis*, pp. 201–211, Springer, 2015.

[98] M. Maleki, M. Teshnehlab, and M. Nabavi, *Diagnosis of multiple sclerosis (ms) using convolutional neural network (cnn) from mris*, *Global Journal of Medicinal Plant Research* **1** (2012), no. 1 50–54.

[99] D. Zikic, Y. Ioannou, M. Brown, and A. Criminisi, *Segmentation of brain tumor tissues with convolutional neural networks*, *Proceedings MICCAI-BRATS* (2014) 36–39.

[100] J. Bernal, K. Kushibar, D. S. Asfaw, S. Valverde, A. Oliver, R. Martí, and X. Lladó, *Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review*, *Artificial intelligence in medicine* **95** (2019) 64–81.

[101] C. Wachinger, M. Brennan, G. Sharp, and P. Golland, *On the importance of location and features for the patch-based segmentation of parotid glands*, in *MICCAI Workshop on Image-Guided Adaptive Radiation Therapy*, 2014.

[102] F. Eitel, J. P. Albrecht, F. Paul, and K. Ritter, *Harnessing spatial mri normalization: patch individual filter layers for cnns*, *Medical Imaging meets NeurIPs* (2019).

[103] D. Castelvecchi, *Can we open the black box of ai?*, *Nature News* **538** (2016), no. 7623 20.

[104] J. van Hemmen, I. M. Saris, P. T. Cohen-Kettenis, D. J. Veltman, P. Pouwels, and J. Bakker, *Sex differences in white matter microstructure in the human brain predominantly reflect differences in sex hormone exposure*, *Cerebral Cortex* **27** (2017), no. 5 2994–3001.

[105] M. Ingalhalikar, A. Smith, D. Parker, T. D. Satterthwaite, M. A. Elliott, K. Ruparel, H. Hakonarson, R. E. Gur, R. C. Gur, and R. Verma, *Sex differences in the structural connectome of the human brain*, *Proceedings of the National Academy of Sciences* **111** (2014), no. 2 823–828.

[106] J. Zhang, Y. Zhao, H. Li, and C. Zong, *Attention with sparsity regularization for neural machine translation and summarization*, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **27** (2018), no. 3 507–518.

[107] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, *Improving neural networks by preventing co-adaptation of feature detectors*, *arXiv preprint arXiv:1207.0580* (2012).

[108] R. Müller, S. Kornblith, and G. E. Hinton, *When does label smoothing help?*, in *Advances in Neural Information Processing Systems*, pp. 4694–4703, 2019.

[109] K. Menzler, M. Belke, E. Wehrmann, K. Krakow, U. Lengler, A. Jansen, H. Hamer, W. H. Oertel, F. Rosenow, and S. Knake, *Men and women are different: diffusion tensor imaging reveals sexual dimorphism in the microstructure of the thalamus, corpus callosum and cingulum*, *Neuroimage* **54** (2011), no. 4 2557–2562.

[110] R. Westerhausen, C. Walter, F. Kreuder, R. A. Wittling, E. Schweiger, and W. Wittling, *The influence of handedness and gender on the microstructure of the human corpus callosum: a diffusion-tensor magnetic resonance imaging study*, *Neuroscience letters* **351** (2003), no. 2 99–102.

[111] R. Westerhausen, F. Kreuder, S. D. S. Sequeira, C. Walter, W. Woerner, R. A. Wittling, E. Schweiger, and W. Wittling, *Effects of handedness and gender on macro-and microstructure of the corpus callosum and its subregions: a combined high-resolution and diffusion-tensor mri study*, *Cognitive brain research* **21** (2004), no. 3 418–426.

[112] E. Dhamala, K. W. Jamison, M. R. Sabuncu, and A. Kuceyeski, *Sex classification using long-range temporal dependence of resting-state functional mri time series*, *Human Brain Mapping* **41** (2020), no. 13 3567–3579.

[113] C. Zhang, C. C. Dougherty, S. A. Baum, T. White, and A. M. Michael, *Functional connectivity predicts gender: Evidence for gender differences in resting brain connectivity*, *Human brain mapping* **39** (2018), no. 4 1765–1776.

[114] S. Weis, K. R. Patil, F. Hoffstaedter, A. Nostro, B. T. Yeo, and S. B. Eickhoff, *Sex classification by resting state brain connectivity*, *Cerebral cortex* **30** (2020), no. 2 824–835.

[115] D.-L. Feis, K. H. Brodersen, D. Y. von Cramon, E. Luders, and M. Tittgemeyer, *Decoding gender dimorphism of the human brain using multimodal anatomical and diffusion mri data*, *Neuroimage* **70** (2013) 250–257.

[116] L. Yuan, F. Chen, L.-L. Zeng, L. Wang, and D. Hu, *Gender identification of human brain image with a novel 3d descriptor*, *IEEE/ACM transactions on computational biology and bioinformatics* **15** (2015), no. 2 551–561.

[117] Z. Luo, C. Hou, L. Wang, and D. Hu, *Gender identification of human cortical 3-d morphology using hierarchical sparsity*, *Frontiers in human neuroscience* **13** (2019) 29.

[118] G. Brooks, T. Fahey, and T. White, *Physiologic responses and long-term adaptations to exercise*, *Exercise physiology: human bioenergetics and its applications. 2nd ed. Mountain View (CA): Mayfield Publishing Co* (1996) 61–77.

[119] M. K, H. S, C. P, P. D, N. SE, and J. W, *Association of cardiorespiratory fitness with long-term mortality among adults undergoing exercise treadmill testing*, *JAMA Network Open* **1** (2018), no. 6 e183605–.

[120] M. S. Patel, L. Foschini, G. W. Kurtzman, J. Zhu, W. Wang, C. A. Rareshide, and S. M. Zbikowski, *Using wearable devices and smartphones to track physical activity: initial activation, sustained use, and step counts across sociodemographic characteristics in a national sample*, *Annals of internal medicine* **167** (2017), no. 10 755–757.

[121] T. Althoff, E. Horvitz, R. W. White, and J. Zeitzer, *Harnessing the web for population-scale physiological sensing: A case study of sleep and performance*, in *Proceedings of the 26th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, 2017.

[122] T. Quisel, L. Foschini, A. Signorini, and D. C. Kale, *Collecting and analyzing millions of mhealth data streams*, in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1971–1980, ACM, 2017.

[123] B. Ballinger, J. Hsieh, A. Singh, N. Sohoni, J. Wang, G. H. Tison, G. M. Marcus, J. M. Sanchez, C. Maguire, J. E. Olgin, and M. J. Pletcher, *Deepheart: Semi-supervised sequence learning for cardiovascular risk prediction*, in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, 2018.

[124] J. H. Migueles, C. Cadenas-Sanchez, U. Ekelund, C. D. Nyström, J. Mora-Gonzalez, M. Löf, I. Labayen, J. R. Ruiz, and F. B. Ortega, *Accelerometer data collection and processing criteria to assess physical activity and other outcomes: a systematic review and practical considerations*, *Sports medicine* **47** (2017), no. 9 1821–1845.

[125] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, *Efficient backprop*, in *Neural networks: Tricks of the trade*, pp. 9–48. Springer, 2012.

[126] Y. Bengio, A. Courville, and P. Vincent, *Representation learning: A review and new perspectives*, *IEEE transactions on pattern analysis and machine intelligence* (2013).

[127] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, *Wavenet: A generative model for raw audio.*, in *SSW*, p. 125, 2016.

[128] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin, *Attention is all you need*, in *Advances in Neural Information Processing Systems*, 2017.

[129] D. Kingma and J. Ba, *Adam: A method for stochastic optimization*, *CoRR* (2014) [arXiv:1412.6980].

[130] F. Chollet *et. al.*, "Keras." `https://keras.io`, 2015.

[131] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et. al.*, *Tensorflow: a system for large-scale machine learning.*, in *OSDI*, vol. 16, pp. 265–283, 2016.

[132] V. Nair and G. E. Hinton, *Rectified linear units improve restricted boltzmann machines*, in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010.

[133] T. Chen and C. Guestrin, *Xgboost: A scalable tree boosting system*, in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, ACM, 2016.

[134] J. S. Oppenheim, J. E. Skerry, and M. S. Gazzaniga, *Magnetic resonance imaging morphology of the corpus callosum in monozygotic twins*, *Annals of neurology* **26** (1989), no. 1 100–104.

[135] A. G. Jansen, S. E. Mous, T. White, D. Posthuma, and T. J. Polderman, *What twin studies tell us about the heritability of brain development, morphology, and function: A review*, *Neuropsychology review* **25** (2015), no. 1 27–46.