université PARIS-SACLAY

# Predictive Modelling in Diffusion MRI: Benchmarking, Interpretability, and a Cortical Microstructure Application

*Modélisation prédictive en IRM de diffusion : analyse comparative, interprétabilité et application à la microstructure corticale*

**Thèse de doctorat de l'Université Paris-Saclay**

École doctorale n°580 Sciences et Technologies de l'Information et de la Communication (STIC)
Spécialité de doctorat: Sciences du traitement du signal et des images

Graduate school: Informatique et sciences du numérique

Référent: Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche **Inria Saclay-Île-de-France (Université Paris-Saclay, Inria)**, sous la direction de **Demian WASSERMANN**, Directeur de Recherche, le co-encadrement de **Reuben DORENT**, Chercheur.

**Thèse soutenue à Paris-Saclay, le 9 octobre 2026, par**

**Gabriela GÓMEZ JIMÉNEZ**

## Composition du Jury

Membres du jury avec voix délibérative

| | |
|---|---|
| Sylvain CHEVALLIER<br>Professeur, LISN-CNRS/Inria Saclay | Président |
| Camille MAUMET<br>Directrice de Recherche, Inria Rennes | Rapporteuse & Examinatrice |
| Roberto TORO<br>Directeur de Recherche, Institut Pasteur | Rapporteur & Examinateur |
| Oscar ESTEBAN<br>Professeur, HES-SO Valais-Wallis (Suisse) | Examinateur |

Thèse de doctorat

NNT : 2026UPASG047

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

**Titre:** Modélisation prédictive en IRM de diffusion : analyse comparative, interprétabilité et application à la microstructure corticale

**Mots clés:** Analyse comparative, Cerveau, IRM de diffusion, Apprentissage automatique, Apprentissage prédictif, Reproductibilité

**Résumé:** Comprendre comment le cerveau nous permet de penser, de ressentir, d'apprendre et d'interagir avec le monde dépend largement des neurones et de leurs connexions. L'IRM de diffusion permet d'étudier cette organisation in vivo en mesurant les déplacements de l'eau dans les tissus, bien que cette mesure reste indirecte : des modèles mathématiques sont nécessaires pour traduire le signal de diffusion en une mesure de microstructure, de sorte que ce que l'on apprend dépend des hypothèses de ces modèles. Le tissu cérébral se divise globalement en substance blanche, bien organisée et relativement facile à étudier, et en substance grise, qui contient les corps cellulaires, les dendrites et un réseau dense de connexions locales, et qui est plus difficile à modéliser et moins bien comprise. Cette thèse cherche à comprendre comment les changements de microstructure sont liés à la cognition, au comportement et à d'autres résultats, une question simple mais difficile à résoudre : les résultats sont difficiles à comparer entre études aux méthodes légèrement différentes, les résultats obtenus sur un jeu de données ne se généralisent pas toujours à un autre, et la plupart des travaux antérieurs se sont concentrés sur la substance blanche ou ont décrit la substance grise à l'aide de mesures simples qui ne capturent qu'une partie limitée de l'information qu'elle porte.
Cette thèse aborde ces problèmes à travers trois chapitres. Le premier propose un cadre de benchmark en libre accès qui standardise le prétraitement, la préparation des données et la prédiction au sein d'un même pipeline, montrant que les modèles d'apprentissage automatique prédisent efficacement plusieurs résultats à partir de l'IRM de diffusion, sans révéler quelles régions cérébrales sous-tendent ces prédictions. Le second comble cette lacune à l'aide d'un cadre de sélection de stabilité, testant si les régions importantes restent cohérentes selon les représentations des données, échantillons d'entraînement et modèles, et en les comparant à la littérature existante. Le troisième évalue dans quelle mesure la cognition peut être prédite à partir de la seule microstructure de la substance grise, à l'aide de la probabilité de retour à l'origine (RTOP), une mesure plus fine reflétant la densité de l'empaquetage tissulaire. Nous montrons que la RTOP corticale prédit les capacités cognitives individuelles, plus fortement que les mesures émotionnelles ou psychosociales, et que ce résultat se généralise à un jeu de données indépendant.

**Title:** Predictive Modelling in Diffusion MRI: Benchmarking, Interpretability, and a Cortical Microstructure Application

**Keywords:** Benchmark, Brain, diffusion MRI, Machine Learning, Predictive Learning, Reproducibility

**Abstract:** Understanding how the brain lets us think, feel, learn, and interact with the world depends largely on neurons and their connections. Diffusion MRI lets us study this organisation in-vivo by measuring how water moves within tissue, though this measurement is indirect: mathematical models are needed to turn the diffusion signal into a measure of microstructure, so what we learn depends on the assumptions these models make. Brain tissue divides broadly into white matter, well-organised and relatively easy to study, and gray matter, which contains cell bodies, dendrites, and dense local connections, and is harder to model and less well understood. This thesis asks how changes in microstructure relate to cognition, behaviour, and other outcomes, a simple question that is hard to answer: results are difficult to compare across studies with slightly different methods, findings from one dataset often do not generalise to another, and most prior work has focused on white matter or described gray matter using simple measures that miss much of the information it carries.
This thesis addresses these problems across three chapters. The first introduces an open-source benchmarking framework standardising preprocessing, data preparation, and prediction within a single pipeline, showing that machine learning models predict several outcomes well from diffusion MRI, without revealing which brain regions drive these predictions. The second addresses this gap, using a stability selection framework to test whether important regions remain consistent across data representations, training samples, and models, and comparing them with the existing literature. The third asks how well cognition can be predicted from gray matter microstructure alone, using the return-to-origin probability (RTOP), a finer measure reflecting how tightly packed brain tissue is. We show that cortical RTOP predicts individual cognitive ability, more strongly than emotional or psychosocial measures, and that this holds in an independent dataset.

*A mi familia.*

# Acknowledgements

I want to start thanking the support and guidance of my supervisors, Demian and Reuben for the guidance and exchanges through these years. Thanks to Demian for the opportunity. For opening the doors of diffusion MRI and to be part of a team such MIND. Thanks to Reuben too for the discussions and for the support during the lasts years of my PhD. To both, thank you for your expertise, patience, and encouragement; it has been a journey with its ups and downs, it has not been easy, but I am greatful, happy and proud of the work done.

I would not be writing this without the support of my colleagues and friends and unfortunately I cannot mention all of them (otherwise these acknowledgements would be longer than the thesis itself). But I want to thank all the people from the MIND, SODA and other teams at INRIA. Arriving to an office with you, chatting, laughing and exchanging really made the difference. Thank you so much for the support, the help and the encouragement, really, it might sound like a cliche, but it would have been even harder without you. However, I want to make a special mention to Fernanda. She has been one of the first people I met when I arrived to the lab and she has been a great support and friend since then. Having her around, through the whole journey, with a heart like hers has been always a source of comfort and joy. Other people from outside of this lab but that has also been a great support and without whom this journey would have also been harder is Paulino. Thank you so much for the trips in the RER, for the talks, for the support and for the friendship.

Benoît, I don't think there are enough words to express my absolute gratitude for what you have done during these years. Without your support, your guidance, your patience and your friendship, I would not be here writing this. Thank you for being a constant source of motivation and inspiration and helping me see the positive side of things when I was struggling.

Por último, me gustaría agradecer a mi madre y toda mi familia por su apoyo desde la distancia. Tener a gente que cree en ti y te apoya, aunque no entiendan mucho lo que haces o sea desde la distancia, es una fuente de motivación y fuerza que no tiene precio. Así que muchas gracias a todos por estar ahí durante todo este proceso.

Thank you everyone for being next to me while I grew as a researcher. Muchísimas gracias a todos por todo el apoyo dado.

# Contents


**1 Introduction** **5**

1.1 Diffusion MRI as a window to microstructure . . . . . 9

1.1.1 From MRI to diffusion MRI . . . . . 9

1.1.2 Why diffusion reflects tissue microstructure . . . . . 10

1.1.3 Why diffusion MRI is valuable for computational analysis . . . 12

1.2 Signal processing and Machine Learning for diffusion MRI . . . . . 13

1.2.1 Signal preprocessing . . . . . 13

1.2.2 Feature extraction from diffusion MRI . . . . . 15

1.2.3 Predictive modelling with machine learning . . . . . 18

1.3 Contributions of this thesis . . . . . 20

**2 DiffBench: A Reproducible Benchmark in Diffusion MRI Prediction** **23**

2.1 Introduction . . . . . 23

2.2 Related Work . . . . . 25

2.2.1 Microstructural representations and feature design . . . . . 25

2.2.2 Diffusion MRI microstructure in predictive modelling . . . . . 27

2.2.3 Machine learning and deep learning approaches for dMRI . . . 28

2.2.4 Benchmarking and Reproducibility . . . . . 29

2.2.5 Gaps Addressed by DiffBench . . . . . 30

2.3 Methods . . . . . 31

2.3.1 Datasets . . . . . 32

2.3.2 DiffBench Preprocessing . . . . . 33

2.3.3 Data Preparation . . . . . 33

2.3.4 Prediction and validation . . . . . 38

2.3.5 Methodological Limitations . . . . . 40

2.4 DiffBench Software Architecture . . . . . 41

2.4.1 Data Preparation . . . . . 42

2.4.2 Model Interface and Learning Backends . . . . . 45

2.4.3 Experiment Infrastructure . . . . . 46

2.4.4 Design and Benchmark Philosophy . . . . . 47

2.5 Results . . . . . 48

2.6 Conclusion . . . . . 52

**3 Stability Selection for Interpretable Diffusion MRI** **55**
3.1 Introduction . . . 55
3.2 Related Work . . . 57
3.2.1 Region based analysis in neuroimaging . . . 57
3.2.2 Interpretability and explainability in ML for neurosciences . . . 58
3.3 Methods . . . 60
3.3.1 Data Representation . . . 61
3.3.2 Sparse Predictive Models . . . 62
3.3.3 Stability-Based Region Importance . . . 65
3.3.4 Interpretation of Stability Maps . . . 67
3.3.5 Limitations . . . 68
3.4 Results . . . 69
3.5 Conclusion . . . 77

**4 Predicting Cognition from Cortical Microstructure With Diffusion MRI** **79**
4.1 Introduction . . . 79
4.2 Related Work . . . 81
4.2.1 The brain-behaviour gap and the case for gray microstructure. 81
4.2.2 Diffusion MRI and RTOP: a candidate biomarker of cortical microstructure . . . 82
4.2.3 Into predictive modelling . . . 83
4.3 Methodology . . . 84
4.3.1 Datasets . . . 84
4.3.2 Preprocessing and data preparation . . . 84
4.3.3 Behavioural measurements . . . 85
4.3.4 Model . . . 85
4.4 Results . . . 86
4.5 Conclusion . . . 91

**Conclusions and Perspectives** **95**

# Synthèse en français

Comprendre comment le cerveau nous permet de penser, de ressentir, d'apprendre et d'interagir avec le monde reste une question importante en neurosciences. L'IRM de diffusion permet d'étudier l'organisation du cerveau *in vivo* en mesurant les déplacements des molécules d'eau dans les tissus. Ces déplacements sont influencés par les structures cellulaires et contiennent donc des informations sur la microstructure cérébrale. Cependant, cette mesure reste indirecte : des modèles mathématiques sont nécessaires pour traduire le signal de diffusion en indicateurs de microstructure, et les informations obtenues dépendent des hypothèses de ces modèles.

Le cerveau est principalement composé de substance blanche et de substance grise. La substance blanche contient des fibres axonales relativement bien organisées, ce qui facilite son étude avec l'IRM de diffusion. La substance grise, composée notamment des corps cellulaires, des dendrites et de nombreuses connexions locales, présente une organisation plus complexe et reste moins bien comprise.

Cette thèse cherche à comprendre comment les variations de la microstructure cérébrale sont liées aux caractéristiques démographiques, à la cognition et au comportement. Plusieurs difficultés se présentent. D'abord, les études utilisent des méthodes différentes de prétraitement, de préparation des données et de modélisation, ce qui rend leurs résultats difficiles à comparer. Ensuite, les résultats obtenus sur un jeu de données ne se généralisent pas toujours à un autre, ce qui pose la question de leur fiabilité. Enfin, les études prédictives se concentrent principalement sur la substance blanche, tandis que la substance grise est souvent décrite à l'aide de mesures simples, comme le volume ou l'épaisseur corticale, qui ne capturent qu'une partie de son information microstructurelle.

Cette thèse aborde ces limites à travers trois contributions principales.

La première contribution est le développement d'un cadre de benchmark en libre accès qui rassemble le prétraitement, la préparation des données et l'apprentissage automatique dans un pipeline commun et reproductible. Il permet de comparer différentes méthodes dans des conditions cohérentes et d'évaluer l'influence des choix méthodologiques sur les performances prédictives. Les résultats montrent que les modèles d'apprentissage automatique appliqués à l'IRM de diffusion peuvent obtenir de bonnes performances pour plusieurs tâches. Cependant, ces performances ne permettent pas, à elles seules, d'identifier les régions cérébrales qui contribuent aux prédictions.

La deuxième contribution propose une approche fondée sur la sélection de stabilité pour étudier la cohérence des régions cérébrales identifiées par les modèles prédictifs. Nous utilisons des modèles parcimonieux et répétons les analyses en faisant varier les représentations des données, les échantillons d'entraînement et les modèles employés. Cette méthode permet d'évaluer si certaines régions sont sélectionnées de manière cohérente malgré ces changements et de comparer les résultats à la littérature existante. Elle vise ainsi à rendre l'interprétation des modèles plus fiable, au-delà des résultats d'une seule analyse.

La troisième contribution étudie dans quelle mesure la microstructure de la substance grise permet, à elle seule, de prédire les capacités cognitives individuelles. Nous utilisons la probabilité de retour à l'origine (RTOP), une mesure issue de l'IRM de diffusion qui apporte des informations sur les restrictions du déplacement de l'eau et l'organisation des tissus. Les résultats montrent que la RTOP corticale contient des informations prédictives sur les capacités cognitives individuelles, avec des associations plus fortes pour la cognition que pour les caractéristiques émotionnelles ou psychosociales. Ces résultats se généralisent également à un jeu de données indépendant, acquis dans des conditions différentes.

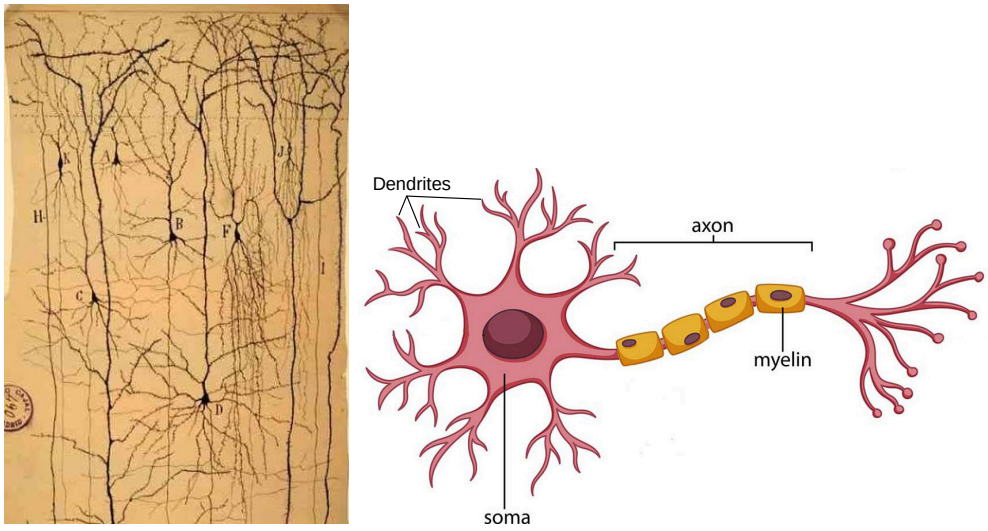
En conclusion, cette thèse propose un cadre reproductible pour comparer les méthodes prédictives en IRM de diffusion, une approche pour évaluer la stabilité des régions cérébrales importantes pour les prédictions, et de nouvelles observations sur le lien entre la microstructure corticale et la cognition. Elle montre que la microstructure cérébrale, et particulièrement celle de la substance grise, constitue une source d'information prometteuse pour mieux comprendre les différences individuelles dans les fonctions cognitives.

# Chapter 1

# Introduction

Understanding the human brain remains an active question of modern neuroscience. Despite more than a century of advances in neuroanatomy, physiology, and imaging, the question of how the structure of our physiology relates to perception, thought, emotion, and behaviour continues to drive research across neuroscience, and computational science [Gao and Ganguli, 2015; Paillette et al., 2021; Smith, 1988]. The foundations of our current understanding of brain organization were established in large part by Santiago Ramón y Cajal and his work in neuroanatomy [Zergeroğlu and Nalçacı, 2015]. Through meticulous histological observations and remarkably detailed anatomical drawings (Figure 1.1-a), in spite of not having access to modern imaging or computational tools, Ramón y Cajal showed that the nervous system is composed of discrete individual cells that communicate through specialized connections, rather than forming a single continuous cellular network, establishing the neuron doctrine that transformed neuroscience. His work revealed not only the morphology of neurons but also the complex organization of their connections, laying the foundations for understanding the brain as a highly structured communication network.

Neurons constitute the fundamental functional units of the nervous system. Figure 1.1-b shows a schematic representation of a neuron, highlighting its principal anatomical components, including the soma, dendrites, and axon. Each neuron consists of a cell body (soma), where cellular metabolism and signal integration occur, a branching tree of dendrites that receives information from neighbouring neurons, and a single axon that propagates electrical impulses toward distant targets. Communication between neurons takes place at specialized connections known as synapses, where electrical activity is converted into chemical signals capable of influencing the activity of other neurons. Axons are frequently wrapped by a lipid-rich insulating, called myelin, which increases the speed and efficiency of neural communication. Alongside neurons, several classes of glial cells provide metabolic support, immune defence, and structural maintenance, contributing actively to the organization and function of neural tissue. Together, these cellular components define the brain's microstructure. Their density, orientation, spatial arrangement, and connectivity determine




(a) (b)

Figure 1.1: **Neuronal morphology.** a) Historical drawing by Santiago Ramón y Cajal illustrating several distinct neuron types found in the nervous system, each varying in the shape and extent of its dendritic tree and axon. b) Schematic representation of a generic neuron, highlighting its main components: the soma, dendrites and axon. Image (b) was adapted from [Johns Hopkins Medicine, 2024].

how information flows through neural circuits and ultimately shape cognition and behaviour. Alterations in this microscopic organization happens throughout normal brain development and ageing [Raz and Rodrigue, 2006], and are marks of numerous neurological and psychiatric disorders. Characterizing brain microstructure is therefore essential for understanding both healthy brain function and disease. However, linking macroscale network organisation to microscale cellular architecture requires methods capable of resolving brain structure at much finer spatial resolution.

At the macroscopic scale, the brain is broadly organized into two complementary tissue types that reflect this underlying cellular architecture. Figure 1.2 shows a slice of the human brain, with the white and gray matter distinctions and an insight on their microstructure. Gray matter comprises the cell bodies of neurons, dendrites, and synapses, and is concentrated primarily in the cerebral cortex (the highly folded outer surface of the brain) as well as in a range of subcortical structures including the hippocampus, the basal ganglia, the thalamus, and the cerebellum. The cortex itself is parcellated into distinct regions (Figure 1.3), each associated with particular cognitive or sensorimotor functions, from the primary visual and auditory areas to the prefrontal regions that support executive control, decision making, and social behaviour. White matter, by contrast, lies beneath the cortical surface and consists predominantly of myelinated axonal bundles that form the long-range connections between gray matter regions [Fields, 2008]. These pathways enable the rapid exchange of information across the brain, and their organization is far from random. Specific tracts reliably connect particular cortical and subcortical areas, forming the struc-

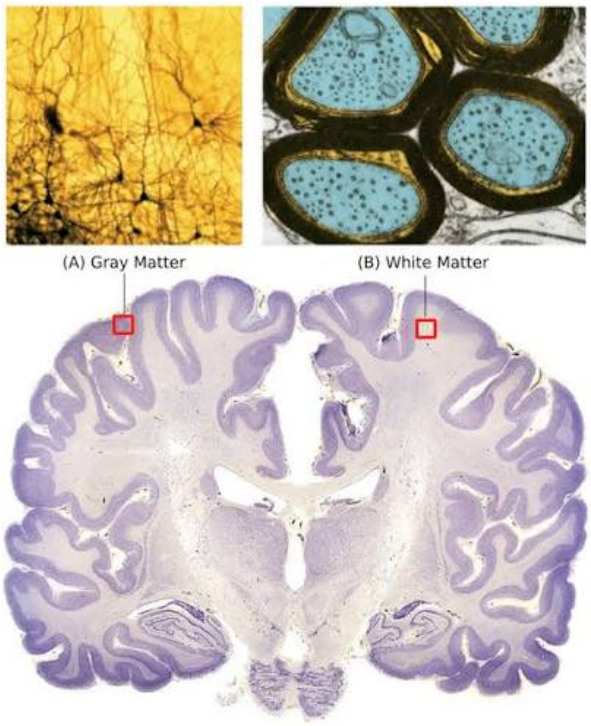



Figure 1.2: **Histological section of gray and white matter.** A stained post-mortem brain slice showing (A) gray matter, comprising cortical cell bodies, and (B) white matter, composed of myelinated axonal tracts, revealing their distinct cellular organisation. Image adapted from [Diez, 2018].

tural stage on which functional brain networks are built [Sporns, 2016; Fornito et al., 2016]. Changes to either gray or white matter (whether through neurodegeneration, demyelination, or developmental alteration) can profoundly affect the cognitive and behavioural functions that depend on the affected circuits.

Historically, one of the first ways of studying this cellular architecture was post-mortem histology. Figure 1.2 illustrates this approach, showing a histological section of a brain slice alongside a finer detailed views of gray and white matter, highlighting their differences at cellular organization and tissue composition. These techniques involve cutting thin sections of fixed brain tissue, treating them with histological dyes to highlight specific cellular components, and examining them under a microscope [Annese, 2012; Alkemade et al., 2023]. These studies have been essential because they established the cytoarchitectural maps of the cortex, identified the major white matter tracts, and revealed the pathological signatures of diseases including Alzheimer's, Parkinson's, and multiple sclerosis [Yuste, 2015]. However, histological approaches are inherently limited. They require access to tissue that can only be obtained after death, they capture a single moment in time, and they cannot be used to study how brain structure changes over the course of development, ageing, or disease in a living individual [Raz and Rodrigue, 2006; Alkemade et al., 2023]. This inability

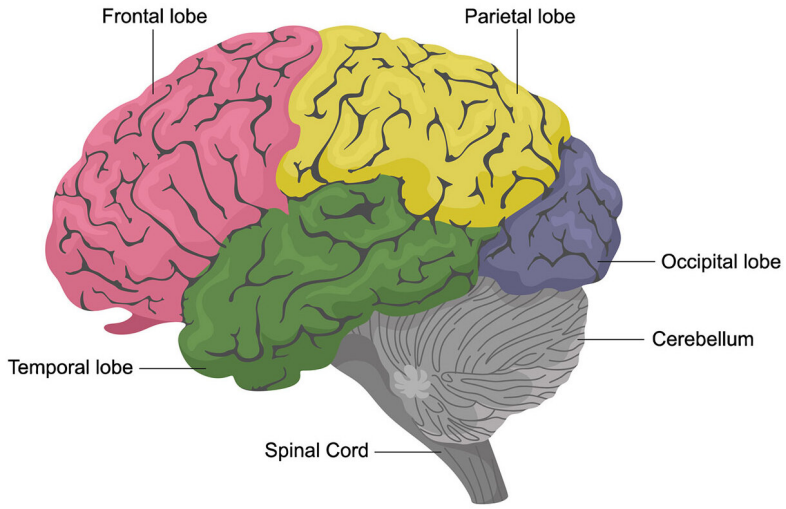



Figure 1.3: **The lobes of the cerebral cortex.** The cortex is divided into four main lobes, frontal, parietal, temporal, and occipital, each broadly associated with distinct cognitive and sensorimotor functions, from visual and auditory processing to executive control and decision making. Image adapted from [Johns Hopkins Medicine, 2024].

to observe the brain over time, or to relate structural properties to behaviour and cognition measured in the same person at the same moment, represents a fundamental constraint that post-mortem methods alone cannot overcome.

The development of non-invasive neuroimaging techniques transformed this landscape by making it possible to study brain structure and function repeatedly and without harm [Lerch et al., 2017]. Modalities such as electroencephalography (EEG) and magnetoencephalography (MEG) capture the electrical and magnetic activity of neuronal populations with millisecond precision [He and Liu, 2008], while positron emission tomography (PET) maps metabolic activity and receptor distributions [Raichle, 1998]. Among these neuroimaging modalities, magnetic resonance imaging (MRI) has become one of the most influential tools in modern neuroscience. It provides detailed, three-dimensional images of brain anatomy and functional activity, allowing researchers to study how brain connections and structure changes throughout development, ageing, and neurological conditions [Lerch et al., 2017]. Among MRI-based techniques, diffusion MRI is one of the methods for exploring tissue microstructure in vivo. By measuring the microscopic displacement of water molecules as they interact with cellular structures, it provides indirect but sensitive measures of the microstructural architecture that are inaccessible to conventional MRI [Le Bihan, 2003; Novikov, 2021; Alexander et al., 2019].

The combination of these imaging advances with computational modelling and machine learning has opened new possibilities for understanding the brain at scale. Large neuroimaging datasets, acquired across diverse populations and clinical groups, now make it possible for us to ask how brain microstructure varies across individuals,

how it changes with age or disease, and whether it can predict cognitive or clinical outcomes [Fornito et al., 2016; Woo et al., 2017]. Diffusion MRI is particularly well positioned to contribute to these questions, because it generates high-dimensional representations of brain organization that serve as inputs to predictive models. The following sections build on this opportunity by describing computational methods for extracting reliable, meaningful information from diffusion MRI data and using it to predict individual differences in brain structure and clinical status. They first introduce the physical basis of diffusion MRI and the models used to estimate microstructural features, then discuss the signal processing and machine learning methods that transform these features into predictions, and finally place the methodological contributions presented here within the broader research landscape.

## 1.1 Diffusion MRI as a window to microstructure

Diffusion MRI is sensitive to the microscopic organization of tissue itself [Le Bihan, 2003; Novikov et al., 2019]. This sensitivity arises from the interaction between water molecules and the cellular structures that surround them, an interaction that leaves measurable traces in the MRI signal and that varies systematically with the underlying tissue architecture [Beaulieu, 2002]. As a result, diffusion MRI can detect subtle changes in axonal organization, or cellular density that are invisible to conventional imaging, making it particularly valuable for studying development, ageing, and disease in living individuals [Alexander et al., 2019; Assaf and Pasternak, 2008]. The following subsections describe the physical basis of this sensitivity, explain how the motion of water encodes information about tissue organization, and discuss why the resulting measurements constitute a rich and flexible source of features for computational and predictive analyses of the brain.

### 1.1.1 From MRI to diffusion MRI

Magnetic resonance imaging (MRI) is a non-invasive technique that uses the magnetic properties of hydrogen nuclei, which are abundantly present in the water molecules of biological tissues, to generate detailed images of internal anatomy [Brown et al., 2014]. By applying carefully designed sequences of magnetic field pulses, an MRI scanner can selectively probe tissue properties and generate images whose contrast reflects local differences in tissue composition and structure. Conventional MRI contrast is primarily determined by relaxation properties of water (denoted $T_1$ and $T_2$) that differ between tissue types, allowing an easy visualisation of anatomical structures such as the cortex, white matter, and subcortical regions.

Diffusion MRI (dMRI) extends this framework by making the acquired signal sensitive to the microscopic displacement of water molecules over a brief observation window, typically of the order of tens of milliseconds [Stejskal and Tanner, 1965; Le Bihan, 2003]. This is achieved by incorporating an additional pair of magnetic field gradients into the pulse sequence, known as the diffusion-encoding gradient. These

gradients effectively encode the position of water molecules at two successive time points, so that the final signal reflects how far and in which direction molecules have moved between the two measurements [Basser and Jones, 2002]. Because the diffusion-encoding gradient has both a strength and an orientation, it is naturally described as a vector denoted $\mathbf{g}$, whose magnitude corresponds to the gradient strength $G$ and whose direction, often expressed as a unit vector referred to as the b-vector, determines the specific direction along which molecular displacement is measured. The degree of signal attenuation produced by this encoding is summarised by a scalar quantity known as the b-value [Le Bihan and Breton, 1985],

$$b = \gamma^2 G^2 \delta^2 (\Delta - \delta/3), \tag{1.1}$$

depends on the gyromagnetic ratio of the hydrogen nucleus ($\gamma$), the strength of the diffusion-encoding gradient ($G$), the duration of the gradient ($\delta$), and timing between the two pulses ($\Delta$) of the applied gradients, and the orientation of these gradients defines the direction along which displacement is measured [Jones et al., 2013]. Since each acquired image corresponds to one specific combination of gradient direction (the b-vector) and b-value, a complete dMRI dataset is built from many such images, acquired across multiple gradient directions and, increasingly, multiple b-values. Together, these measurements provide a multi-dimensional characterisation of water mobility throughout the brain, forming the basis for the microstructural models introduced in the remainder of this chapter.

### 1.1.2 Why diffusion reflects tissue microstructure

The basis for using dMRI to study brain organization lies in a simple physical principle: water molecules are in continuous random motion driven by thermal energy, known as Brownian motion [Einstein, 1906]. Figure 1.4 illustrates this motion in two different environments. In free water, where there are no physical barriers to molecular motion, diffusion is isotropic; water molecules move equally in every direction and can travel relatively long distances during the observation period. Inside biological tissue, however, the situation is different. Cellular structures, including membranes, axons, dendrites, myelin sheaths, and the extracellular matrix, obstruct, restrict, or redirect water movement, so that the displacement of a molecule over a given time reflects the geometry of the local cellular environment it is in [Beaulieu, 2002].

This close relationship between water diffusion and the surrounding cellular environment is what makes diffusion MRI sensitive to brain microstructure. Figure 1.5 illustrates how the organization of white and grey matter leads to distinct diffusion profiles. In white matter, long bundles of parallel myelinated axons impose a strongly directional constraint on water movement, molecules diffuse along the axis of the fibres but are substantially obstructed when attempting to cross them [Beaulieu, 2002; Assaf and Pasternak, 2008]. This directional preference, known as anisotropic diffusion, is particularly pronounced in well-organized white matter tracts and makes them especially suitable for characterization with dMRI. Grey matter presents a more

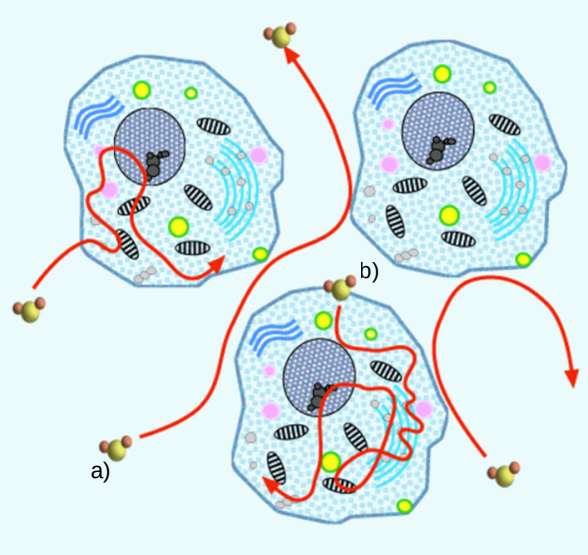



Figure 1.4: **Water diffusion in biological tissue.** Water molecules undergo random Brownian motion, but their displacement is constrained by cellular structures. In free water (a), molecules diffuse isotropically and free without restriction; while inside cells (b) their motion is constrained by membranes and other obstacles, leading to anisotropic diffusion that reflects the underlying microstructure. Image adapted from Elster [2026].

complex cellular organization composed primarily of neuronal cell bodies, dendrites, synapses, and glial cells. This less coherent cellular architecture leads to a more isotropic diffusion profile. However, diffusion is still partially constrained by cellular density, making dMRI informative although more challenging to interpret [Novikov et al., 2019; Alexander et al., 2019].

The diffusion patterns observed in white and grey matter illustrate a more general principle than the basis of diffusion MRI. Rather than visualizing cells or fibres directly, the dMRI technique measures how the collective motion of water molecules is influenced by the surrounding tissue architecture. Each voxel contains a huge population of water molecules distributed across many cells and extracellular spaces, and the measured diffusion signal represents the average behaviour of all these molecules over the diffusion time [Novikov et al., 2019]. Consequently, the signal does not provide a direct image of the underlying microstructure. Instead, it reflects the statistical effects that cellular features such as axonal orientation, membrane density, myelination, and cell packing have on water diffusion. This indirect relationship is precisely what makes diffusion MRI a window into tissue microstructure. Changes in the cellular environment, such as alterations in axonal density, modify the diffusion of water molecules and therefore leave characteristic signatures in the measured signal. However, these tissue properties cannot be observed directly from the diffusion-weighted images alone. Interpreting them requires mathematical models that relate the measured diffusion signal to biologically meaningful descriptors of tissue organization, transforming signal

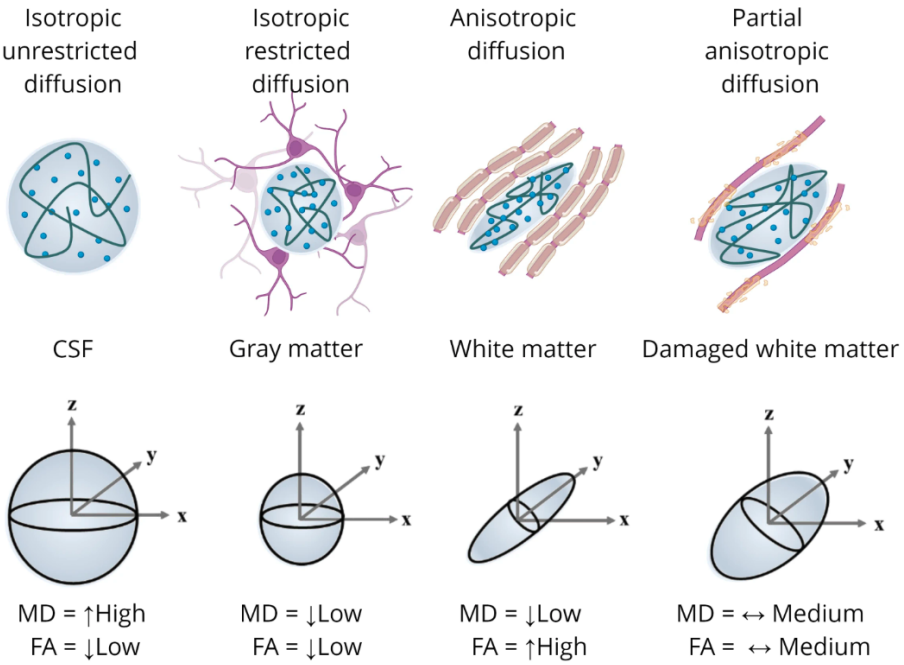



Figure 1.5: **Water diffusion in tissue and its relation to the DTI microstructural model.** Water diffusion is shaped by surrounding cellular structures: CSF shows unrestricted, isotropic diffusion (high MD, low FA); gray matter shows isotropic but restricted diffusion (low MD, low FA); white matter shows strongly anisotropic diffusion along aligned fibres (low MD, high FA); and damaged white matter shows an intermediate profile reflecting disrupted fibre organisation. Diffusion tensor ellipsoids for each tissue type are shown below. Image adapted from Bedell [2026].

variations into quantitative estimates of brain microstructure [Alexander et al., 2019; Jelescu and Budde, 2017]. This modelling step, which transforms the raw diffusion signal into quantitative biomarkers, will be extended in the following section.

### 1.1.3 Why diffusion MRI is valuable for computational analysis

Beyond its biophysical interest, diffusion MRI has proven to be a particularly rich source of quantitative information for neuroscience and clinical research. Diffusion-derived measures are sensitive to a wide range of biologically and clinically meaningful processes, including brain development, healthy ageing and cognitive performance [Jones et al., 2013; Alexander et al., 2019]. Changes in white matter integrity, fibre organization, or cellular density associated with conditions such as Alzheimer's disease, or brain injury can all present measurable alterations in diffusion-derived metrics, making dMRI a valuable technique worth exploring [Assaf and Pasternak, 2008; Novikov et al., 2019].

From a computational perspective, a key property of diffusion MRI is that it produces high-dimensional representations of brain organization. A single participant's dMRI dataset can be summarized as voxel-wise parameter maps, regions-of-interest, white matter tract profiles, cortex meshes or whole-brain connectivity matrices, each providing a different level of description of the underlying microstructure [Alexander et al., 2019]. These representations are well-suited for machine learning where diffusion-derived measures are used as input features and paired with cognitive, demographic, or clinical outcomes, allowing the training of predictive models that learn the mapping from brain structure to individual characteristics [Woo et al., 2017; Abraham et al., 2014]. This perspective, in which dMRI data are used as a measure instrument to generate features for predictive modelling, is the one adopted throughout this thesis.

## 1.2 Signal processing and Machine Learning for diffusion MRI

Diffusion MRI provides a rich, non-invasive insight of brain microstructure by measuring how the MR signal changes under different diffusion encodings [Johansen-Berg and Behrens, 2013]. However, the acquired diffusion-weighted images do not directly quantify biological tissue properties [Le Bihan, 2014; Novikov et al., 2019]. Instead, they represent noisy physical measurements that are affected by subject motion, scanner imperfections, acquisition variability, and other imaging artifacts [Schilling et al., 2025; Tax et al., 2019]. Transforming these measurements into biologically meaningful biomarkers therefore requires a computational pipeline with three main steps: signal preprocessing to correct acquisition artifacts, diffusion modelling to estimate quantitative microstructural parameters, and machine learning methods to relate these biomarkers to behavioural, cognitive, or clinical outcomes [Alexander et al., 2019; Novikov et al., 2019; Abraham et al., 2014; Nielsen et al., 2020; Erickson et al., 2017].

### 1.2.1 Signal preprocessing

The first stage of this pipeline aims to remove acquisition-related artifacts while preserving the underlying biological signal. Compared with conventional structural MRI, diffusion MRI is considerably more susceptible to image degradation because the strong diffusion-encoding gradients increase sensitivity to several sources of error [Jones et al., 2013; Alexander et al., 2019; Andersson and Sotiropoulos, 2016a].

Several artifacts commonly affect diffusion MRI acquisitions. Thermal noise reduces the signal-to-noise ratio, particularly at high diffusion weightings [Veraart et al., 2016c]. Rapid switching of the diffusion-encoding gradients induces eddy currents, which generate geometric distortions and spatial misalignment between diffusion-weighted volumes [Andersson and Sotiropoulos, 2016a]. Subject motion between successive diffusion-weighted volumes introduces spatial misalignment, while magnetic

susceptibility differences at tissue interfaces generate nonlinear geometric distortions in echo-planar imaging [Andersson and Sotiropoulos, 2016a]. Additional artifacts, including Gibbs ringing and radiofrequency bias-field inhomogeneities, further degrade image quality if left uncorrected [Andersson and Sotiropoulos, 2016a; Veraart et al., 2016c; Kellner et al., 2016].

To minimize these effects, modern preprocessing pipelines apply a sequence of correction steps before any diffusion model is fitted. Although implementations differ slightly between software packages, a typical pipeline includes thermal-noise removal, Gibbs-ringing correction, correction of motion, eddy-current and susceptibility-induced distortions, bias-field correction, brain extraction, and, when required, spatial registration to a common anatomical space [Andersson and Sotiropoulos, 2016a; Jones et al., 2013; Cieslak et al., 2021]. These procedures aim to ensure that differences observed in the diffusion signal primarily reflect tissue microstructure rather than acquisition artifacts [Alexander et al., 2019; Novikov et al., 2019]. Therefore, careful preprocessing is essential, as residual artifacts may propagate through subsequent analyses, biasing both estimated microstructural parameters and predictive models [Tax et al., 2019; Fortin et al., 2018; Jones et al., 2013]. Standardized software frameworks such as FSL, MRtrix3, DIPY, and QSIPrep have become widely adopted because they provide robust, reproducible implementations of these preprocessing steps and facilitate harmonized analyses across studies [Jenkinson et al., 2012; Tournier et al., 2019; Garyfallidis et al., 2014; Cieslak et al., 2021].

Preprocessing produces a set of artifact-corrected diffusion-weighted images that more faithfully represent the underlying diffusion signal. Nevertheless, these images still do not directly quantify biological properties such as axonal density, fibre orientation, or cellular complexity. Rather, they describe how the MR signal attenuates under different diffusion encodings, reflecting the interaction between water molecules and the surrounding tissue [Basser and Jones, 2002; Le Bihan, 2003; Novikov et al., 2019]. Extracting biologically meaningful information therefore requires mathematical models that relate the measured diffusion signal to the underlying tissue architecture [Alexander et al., 2019; Novikov et al., 2019; Jelescu and Budde, 2017].

From this perspective, diffusion MRI can be regarded as an inverse problem [Alexander et al., 2019; Novikov et al., 2019]. The acquisition provides only a finite set of diffusion-weighted measurements sampled throughout q-space, whereas the tissue microstructure responsible for generating those measurements remains unknown. Diffusion MRI models define the forward relationship between tissue organization and the expected MR signal, allowing quantitative microstructural parameters to be estimated by fitting the model to the observed data [Alexander et al., 2019; Novikov et al., 2019; Jelescu and Budde, 2017]. These estimated parameters, rather than the diffusion-weighted images themselves, constitute the biomarkers that are subsequently used in neuroscience and clinical studies.

### 1.2.2 Feature extraction from diffusion MRI

Once the diffusion-weighted images have been corrected for acquisition artifacts, the next stage of the pipeline consists of extracting quantitative features that represent brain microstructure. These features are obtained by fitting mathematical models to the diffusion signal, which allow biologically meaningful parameters, such as tissue anisotropy, cellular density, or fibre orientation, to be estimated even though they cannot be measured directly [Basser and Jones, 2002; Novikov et al., 2019]

Diffusion models can adopt different strategies to relate the measured diffusion signal and the underlying tissue microstructure [Novikov et al., 2019; Jelescu and Budde, 2017]. At one end of the spectrum, signal representations describe the diffusion signal empirically using flexible mathematical functions without explicitly modelling tissue architecture. Statistical models instead characterize the probability distribution of water displacement, providing a compact description of diffusion while making only limited assumptions about the underlying tissue. Biophysical models go a step further by representing the tissue as a collection of compartments, such as intra- and extracellular spaces or neurites, each with distinct diffusion properties, allowing model parameters to be interpreted in terms of specific microstructural features. These approaches are complementary modelling philosophies, and many modern diffusion models combine ideas from more than one of them.

**Diffusion Tensor Imaging (DTI).** DTI is the simplest and most widely used signal representation [Basser et al., 1994; Basser and Pierpaoli, 2011]. It assumes that, within each voxel, the displacement of water molecules follows a three-dimensional Gaussian distribution, meaning that molecules are, on average, equally likely to move any given distance in a given direction, following a bell-shaped probability profile. Under this assumption, the relationship between the measured signal and the diffusion-encoding gradient $\mathbf{g}$ can be written as

$$S(b, \mathbf{g}) = S_0 exp(-b\mathbf{g}^T\mathbf{D}\mathbf{g}), \tag{1.2}$$

where $S_0$ is the signal measured without any diffusion weighting, $b$ is the diffusion weighting factor introduced earlier, and $\mathbf{D}$ is a symmetric, positive-definite $3 \times 3$ diffusion tensor describing how diffusion varies with direction within the voxel.

Diagonalizing $\mathbf{D}$ gives three eigenvalues ($\lambda_1 \geq \lambda_2 \geq \lambda_3$), which describe the rate of diffusion along three orthogonal directions and from which two widely used scalar measures are derived. The first, mean diffusivity,

$$\text{MD} = \frac{\lambda_1 + \lambda_2 + \lambda_3}{3}, \tag{1.3}$$

summarises the overall rate of water diffusion within the voxel, averaged across all

directions. The second, fractional anisotropy,

$$\mathrm{FA} = \sqrt{\frac{3}{2}} \frac{\sqrt{(\lambda_1 - \mathrm{MD})^2 + (\lambda_2 - \mathrm{MD})^2 + (\lambda_3 - \mathrm{MD})^2}}{\sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}}, \tag{1.4}$$

quantifies how directionally dependent this diffusion is, ranging from zero, when diffusion is equally free in all directions, to one, when diffusion is essentially restricted to a single direction, as is typically the case within coherently organised fibre bundles. FA and MD remain among the most widely reported diffusion biomarkers in clinical and neuroscience research. However, the assumption of a single, well-defined diffusion direction per voxel breaks down in regions containing multiple crossing fibre populations, a limitation that has motivated the development of more flexible models. Figure 1.5 represents how these descriptors are characterized in different tissue microstructures.

**Diffusion Kurtosis Imaging (DKI).** Diffusion Kurtosis Imaging (DKI) extends the tensor model by capturing deviations from the Gaussian diffusion assumed by DTI, which arise in real tissue due to cell membranes, and other microstructural barriers that restrict and compartmentalise water movement [Jensen et al., 2005]. In DTI, the signal is assumed to decay exponentially, and at a constant rate, as the b-value increases; in practice, however, this decay tends to curve, particularly at the higher b-values increasingly used in modern acquisitions, showing that water movement is not perfectly free. DKI captures this behaviour by adding a second term to the signal equation,

$$S(b) = S_0 \exp\left(-bD + \frac{1}{6} b^2 D^2 K\right), \tag{1.5}$$

where $D$ is the apparent diffusivity along the gradient direction $\mathbf{g}$, that is, the effective diffusion coefficient that would be measured if diffusion were assumed to be Gaussian along that specific direction, similar to a single-direction version of the diffusivities described above for DTI. The additional term, $K$, is the apparent kurtosis, a dimensionless quantity that captures how much the signal decay curves away from this simple, straight-line behaviour on a logarithmic scale. A value of zero indicates purely Gaussian diffusion, as assumed by DTI, while increasingly positive values indicate a more pronounced departure from it, typically reflecting greater microstructural complexity, such as tissue heterogeneity or restricted diffusion within small compartments. Because estimating this curvature requires measuring the signal at more than one non-zero b-value, DKI provides additional sensitivity to tissue complexity that a model assuming purely Gaussian diffusion cannot capture.

**Mean Apparent Propagator MRI (MAP-MRI).** Rather than describing diffusion within a small number of predefined tissue compartments, MAP-MRI aims to reconstruct the full three-dimensional diffusion propagator, $P(\mathbf{r})$, the probability that a water molecule undergoes a given displacement $\mathbf{r}$ during the diffusion time [Özarslan

et al., 2013]. This propagator is related to the diffusion signal $E(\mathbf{q})$, measured across many gradient directions and strengths encoded by the vector $\mathbf{q}$, through a Fourier relationship,

$$P(\mathbf{r}) = \int E(\mathbf{q}) \, \exp(-2\pi i \, \mathbf{q} \cdot \mathbf{r}) \, \mathrm{d}\mathbf{q}, \tag{1.6}$$

meaning that, in principle, if the diffusion signal were measured across the entire $q$-space, the displacement probability of water molecules could be recovered directly and without assuming any particular tissue geometry. In practice, only a finite, discrete set of measurements is available, so MAP-MRI approximates $E(\mathbf{q})$ using a fixed set of mathematical functions, called Hermite functions, weighted by a small set of coefficients estimated from the data [Özarslan et al., 2013]. This is conceptually similar to describing a smooth curve by summing a small number of standard basis shapes rather than by measuring every point on the curve directly, and it allows the diffusion propagator to be estimated efficiently even from limited data.

Once these coefficients have been estimated, several complementary microstructural descriptors can be computed directly from them, without needing to re-estimate the propagator from scratch. The return-to-origin probability (RTOP), for instance, corresponds to the value of the propagator at zero displacement, $P(\mathbf{0})$, and can be shown to depend only on the zeroth-order coefficient of the Hermite expansion, providing a simple, closed-form measure of how strongly water diffusion is restricted overall [Özarslan et al., 2013]. A related measure, the return-to-axis probability (RTAP), is obtained analogously by restricting this calculation to displacements occurring in the plane perpendicular to a principal diffusion axis, making it more specifically sensitive to restriction imposed by structures such as axons. Because MAP-MRI does not require assuming a particular tissue geometry in advance, it offers a richer and less assumption-dependent characterisation of microstructure than compartmental models, though this flexibility comes at the cost of requiring denser sampling of $q$-space than simpler models such as DTI or DKI.

**Neurite Orientation Dispersion and Density Imaging (NODDI).** NODDI belongs to the biophysical family and decomposes the signal into three tissue compartments: intracellular space (modelled as sticks representing axons and dendrites), extracellular space (modelled as anisotropic Gaussian diffusion), and a free isotropic water compartment representing cerebrospinal fluid [Zhang et al., 2012]. The total signal is a weighted sum,

$$S = (1 - f_{\mathrm{iso}}) \Big( f_{\mathrm{ic}} \, S_{\mathrm{ic}} + (1 - f_{\mathrm{ic}}) \, S_{\mathrm{ec}} \Big) + f_{\mathrm{iso}} \, S_{\mathrm{iso}}, \tag{1.7}$$

where $f_{\mathrm{ic}}$ and $f_{\mathrm{iso}}$ denote the intracellular and isotropic volume fractions, respectively.

**Soma and Neurite Density Imaging (SANDI).** SANDI extends the compartmental approach beyond white matter by explicitly including a spherical compartment representing cell bodies (soma), in addition to neurite and extracellular compartments

[Palombo et al., 2020]. This addition is particularly relevant in grey matter, where cell bodies occupy a substantial volume fraction and conventional stick-based models tend to underestimate tissue complexity. Because distinguishing soma from neurites requires sensitivity to very short diffusion times and high diffusion weightings, SANDI typically relies on multi-shell acquisitions with high b-values, which is worth keeping in mind when comparing across datasets with different protocols.

**Tractography and structural connectomes.** Beyond voxel-wise parameter estimation, the local fibre orientation information provided by these models, particularly tensor- or fibre orientation distribution-based estimates, can be propagated across the brain through tractography algorithms to reconstruct white matter pathways [Mori et al., 1999; Tournier et al., 2019]. Aggregating these streamlines into regional connectivity matrices yields structural connectomes, which characterize large-scale brain organization and have become an important complementary representation alongside voxel- and region-based microstructural metrics [Sporns et al., 2005].

Taken together, these models illustrate that the same diffusion-weighted dataset can output different quantitative maps depending on the assumptions embedded in the fitting procedure [Jelescu and Budde, 2017; Novikov et al., 2019]. The extracted features, whether voxel-wise parameter maps, regional summaries, tract-specific metrics, connectivity matrices, or, minimally processed signals are the input variables used for the statistical and machine learning methods discussed in the following section. Understanding the principles and limitations of each model is therefore not a minor methodological detail, but a prerequisite for correctly interpreting any downstream biomarker or predictive result.

### 1.2.3 Predictive modelling with machine learning

One of the objectives of predictive modelling from neuroimaging data is to learn relationships between neuroimaging-derived features and variables of interest, such as sex, age, or cognitive performance, among many other potential targets. In this setting, a datasets typically consist of diffusion mri signals for each participant alongside demographic or clinical information. When a model is optimized to predict a target variable $\mathbf{y}$ from a matrix of input features $\mathbf{X}$, the problem is framed as supervised learning: the model is fitted on labeled examples in order to learn a mapping from inputs to outputs. The goal is to achieve a good generalization to unseen data by capturing meaningful associations between brain structure and the outcome of interest [Bishop and Nasrabadi, 2006; Hastie et al., 2009].

Neuroimaging studies have relied on machine learning algorithms, including linear and logistic regression, support vector machines, and random forests [Cortes and Vapnik, 1995; Breiman, 2001]. These methods remain widely used because they are computationally efficient, perform well on relatively small datasets, and are inherently interpretable, making them particularly attractive in neuroscience applications. Furthermore, software libraries such as scikit-learn have made these algorithms easily

accessible and reproducible within the scientific community [Pedregosa et al., 2011].

The fast development of deep learning algorithms has substantially expanded the range of predictive models available for neuroimaging analysis, speeding the improvement of predictive capabilities. Convolutional neural networks (CNNs), originally developed for computer vision, automatically learn hierarchical feature representations from imaging data and have demonstrated strong performance in tasks including image classification, segmentation, reconstruction, and disease prediction [LeCun et al., 2015; Litjens et al., 2017]. A natural extension of this framework to non-Euclidean data is provided by graph convolutional networks (GCNs) [Kipf and Welling, 2016], which operates on graph-structured representations of the brain, such as structural or functional connectivity networks, where nodes correspond to brain regions and edges encode the strength or nature of their interactions [Bronstein et al., 2017]. This representation is particularly well-suited to neuroimaging, as it allows the model to exploit both the properties of individual regions and the relational structure connecting them; GCNs have consequently been applied to tasks including disease classification, population-level phenotype prediction, and brain age estimation [Parisot et al., 2018; Ktena et al., 2017]. More recently, transformer-based architectures [Vaswani et al., 2017] and self-supervised learning approaches [Chen et al., 2020b] have enabled models to exploit large collections of unlabeled neuroimaging data, reducing the dependence on manual feature engineering and improving generalization across datasets and acquisition protocols [Dosovitskiy et al., 2020; Adeli et al., 2026].

Building upon these advances, foundation models have recently emerged as a promising paradigm for neuroimaging. These models are pre-trained on large, heterogeneous imaging datasets using self-supervised objectives, and are subsequently adapted to specific downstream tasks through fine-tuning or transfer learning. By learning general-purpose representations of brain anatomy and microstructure, foundation models have the potential to improve predictive performance and reduce dependence on large labeled cohorts, which remain comparatively scarce in diffusion MRI [Moor et al., 2023].

Regardless of model complexity, all predictive frameworks follow the same fundamental workflow: diffusion MRI data are preprocessed, informative features are extracted, a predictive model is trained on labeled examples, and its performance is evaluated on previously unseen data. The principal differences between modelling approaches therefore lie not in the overall pipeline, but in how they represent the input data, learn relationships between features and outcomes, and generalize beyond the training dataset.

Despite this progress, several challenges continue to limit the translation of predictive models into reliable neuroscience and clinical tools. A first limitation concerns sample size: diffusion MRI datasets are typically far smaller than those used in computer vision, which restricts the complexity of models that can be trained without overfitting and motivates the use of multi-site consortia or transfer learning to compensate for limited data [Abraham et al., 2014; Marek et al., 2022]. A second, closely related issue is scanner- and site-dependent variability: differences

in scanner hardware, pulse sequences, and acquisition parameters introduce systematic, non-biological variation in diffusion measures, an effect commonly referred to as the *site* effect, which can be partially mitigated through statistical harmonization techniques such as ComBat or through learning acquisition-invariant representations [Fortin et al., 2018; Tax et al., 2019]. Heterogeneous acquisition protocols across cohorts introduce additional challenges, since differences in the number and distribution of diffusion gradients, $b$-values, and spatial resolution affect the diffusion models described in the previous section in different ways, making direct pooling of features across datasets nontrivial [Tax et al., 2019; Schilling et al., 2025]. A further concern is demographic and clinical representativeness, since models trained on cohorts that are biased in age, sex, or disease severity risk learning false associations that fail to generalize to underrepresented populations, an issue increasingly recognized as a barrier to fair and clinically deployable neuroimaging models [Sadikov et al., 2025; Fortin et al., 2018; Varoquaux and Cheplygina, 2022].

Beyond these data-related challenges, methodological and interpretability concerns are equally important. Many high-performing models behave as black boxes, making it difficult to relate their predictions back to interpretable microstructural or anatomical features, which is a significant obstacle in clinical settings where understanding the basis of a prediction is often as important as its accuracy [Litjens et al., 2017]. Robust validation practices, including the use of independent held-out test sets, cross-validation strategies that respect family or site structure, and external replication across cohorts, are therefore essential to ensure that reported performance reflects genuine generalization rather than overfitting to dataset-specific artifacts [Abraham et al., 2014]. Equally important are transparent and reproducible research practices, including the public release of code, trained models, and analysis pipelines whenever possible, to facilitate independent verification and comparison of results.

Taken together, these challenges indicate that improving preprocessing pipelines, harmonization strategies, representation learning, and the robustness and interpretability of predictive models remains a central and active focus of diffusion MRI research. Addressing them is not merely a technical refinement but a prerequisite for translating diffusion-derived biomarkers into tools that are reliable across scanners, sites, and populations, motivating many of the methodological developments presented throughout this thesis.

## 1.3 Contributions of this thesis

Although diffusion MRI has become an important tool for studying brain microstructure, several methodological challenges limit its use for robust and reproducible machine learning research. Predictive studies often employ different datasets, preprocessing pipelines, diffusion MRI models, and evaluation protocols, making it difficult to compare published results or determine which methodological choices genuinely improve predictive performance. Furthermore, even when accurate predictive models are obtained, relatively little attention has been devoted to understanding whether

the brain regions identified by these models represent stable neurobiological signals or artifacts of specific analytical choices. Finally, despite the increasing number of advanced diffusion MRI models, many recently proposed microstructural biomarkers remain insufficiently explored for predicting complex cognitive phenotypes.

This thesis addresses these challenges through three complementary contributions.

1. **A benchmarking framework for diffusion MRI predictive modelling.** The first contribution is DiffBench, an open, reproducible benchmarking framework for evaluating machine learning methods on diffusion MRI data. Inspired by benchmarking initiatives that have become standard practice in machine learning, DiffBench provides standardized datasets, evaluation protocols, preprocessing strategies, diffusion representations, and predictive models within a unified experimental framework. By enabling fair comparisons across methods, it promotes reproducibility, facilitates methodological development, and establishes common evaluation standards for diffusion MRI prediction.

2. **An interpretability framework based on stability selection.** Predictive accuracy alone provides limited insight into the neurobiological mechanisms underlying model predictions. The second contribution therefore introduces a stability-selection framework for interpreting sparse machine learning models applied to diffusion MRI. The proposed approach aggregates feature selections across multiple experimental perturbations to identify cortical regions that consistently contribute to prediction. This framework allows regional importance to be evaluated in terms of robustness and reproducibility, providing a more reliable basis for biological interpretation.

3. **Evaluation of advanced diffusion microstructure using RTOP for cognitive prediction.** The third contribution investigates the predictive value of Return-To-the-Origin Probability (RTOP), an advanced diffusion MRI biomarker derived from MAP-MRI, for modelling individual differences in cognitive performance. Unlike conventional diffusion metrics, RTOP is designed to capture aspects of tissue microstructure related to cellular restriction and complexity. This study evaluates whether such advanced microstructural information improves cognitive prediction and provides further insight into the relationship between brain microstructure and cognition.

Together, these contributions address three complementary aspects of diffusion MRI machine learning: how predictive models should be evaluated, how their decisions should be interpreted, and how advanced microstructural representations can be used to study brain function. Collectively, they contribute methodological tools and experimental evidence that support more reproducible, interpretable, and biologically meaningful predictive modelling in diffusion MRI.

# Chapter 2

# DiffBench: A Reproducible Benchmark in Diffusion MRI Prediction

## 2.1 Introduction

Diffusion magnetic resonance imaging (dMRI) enables in vivo measurement of tissue microstructure and provides biologically meaningful biomarkers sensitive to axonal organization, myelination, and tissue complexity [Novikov et al., 2019]. Leveraging these microstructural descriptors for predictive modeling has become increasingly common, with applications ranging from demographic estimation [Chen et al., 2024; He et al., 2022; Chen et al., 2020a; de Mooij et al., 2018; Dibaji et al., 2024], clinical diagnosis [Egle et al., 2022; Wen et al., 2021], and cognitive characterization [Kotikalapudi et al., 2025; Menon et al., 2020]. The combination of diffusion imaging and machine learning is therefore often viewed as a powerful framework for data-driven neurobiological inference.

While many studies demonstrate the promising predictive power of dMRI, reported performance varies widely across studies, tasks, and datasets, making direct comparisons difficult. This variability arises from multiple sources. Preprocessing pipelines differ substantially; microstructural information is represented in diverse ways ranging from voxel-wise maps [Chen et al., 2024] and ROI summaries [Menon et al., 2020] to learned embeddings [Chen et al., 2023a; He et al., 2022] and evaluation protocols vary in data splits, cross-validation schemes, and hyperparameter selection. Limited code and configuration sharing further restrict reproducibility. Consequently, reported performance differences may reflect inconsistencies in implementation and evaluation rather than genuine modeling improvements.

Addressing replication challenges in brain structure–behavior research [Wu et al., 2023; Genon et al., 2022; Rosenblatt et al., 2024], is essential to control these method-

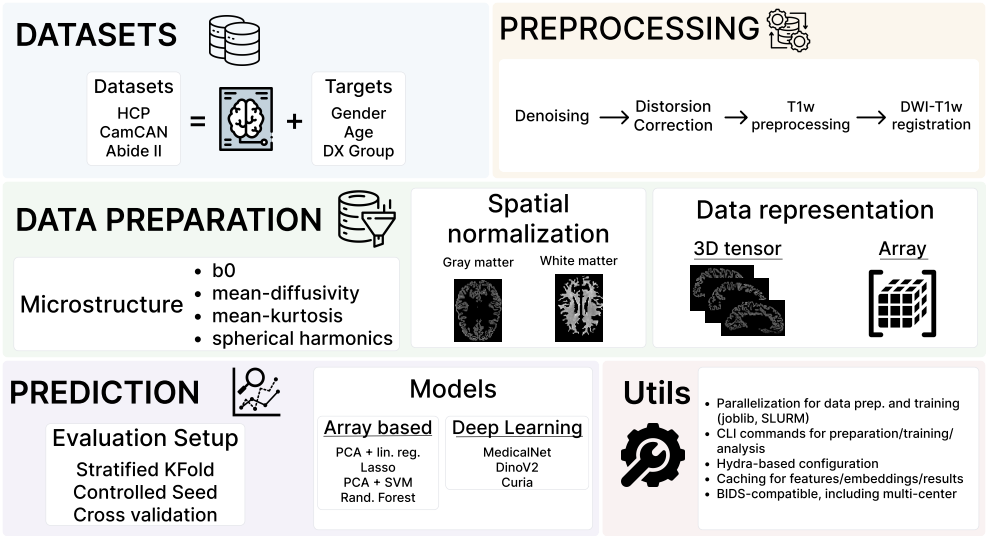



Figure 2.1: **End-to-end overview of the DiffBench pipeline.** Summary of the full benchmark pipeline as a sequence of five components: datasets, preprocessing, data preparation, prediction and validation, and supporting software utilities. Publicly available diffusion MRI datasets and their associated phenotypic variables enter the pipeline first, they are passed through a shared preprocessing stage to generate diffusion derivatives, and are then transformed into multiple data representations, including microstructural features and spatially represented images, which serve as inputs to the predictive models described in later sections.

ological sources of variability and have comparable evaluation standards for dMRI prediction. Accordingly, there has been increasing interest in reproducible benchmarking frameworks within specific clinical contexts or imaging modalities [Wen et al., 2021], and large-scale analyses have assessed the replicability of dMRI-based brain–behavior associations across phenotypes [Rosenblatt et al., 2024]. However, these efforts do not jointly vary microstructural representations, tissue compartments, model families, datasets, and predictive tasks within a unified experimental design. As a result, several fundamental questions remain unsolved: how representation and model choice interact with task characteristics, whether white and gray matter contribute differently across datasets and prediction targets, and under what conditions diffusion-derived features provide consistent predictive value beyond non-diffusion baselines.

To address these limitations, we introduce **DiffBench**, an open-source, and reproducible benchmark for dMRI prediction. DiffBench enables controlled comparison of preprocessing, spatial registration, microstructural representation, model family, and validation within a unified framework applied across multiple public datasets. Specifically, DiffBench (i) unifies preprocessing across cohorts; (ii) compares various tissue-specific microstructural representations; (iii) includes machine learning models and deep foundation architectures; and (iv) evaluates performance across classification and regression tasks. Our analysis shows that performance differences are often as sensitive to representation and preprocessing choices as to model family, highlight-

ing the need for a standardized benchmark. The key contribution of DiffBench is therefore not only the results reported, but the benchmarking framework itself.

## 2.2 Related Work

Diffusion MRI is one of the main modalities for non-invasive characterisation of brain microstructure. It offers quantitative biomarkers sensitive to axonal organisation, myelination, and tissue complexity [Novikov et al., 2019]. Leveraging these descriptors for predictive modelling has become increasingly common across demographic, clinical, and cognitive targets. However, the field faces a recurring challenge: reported performance variability across studies. Chen et al. [2024] and Dibaji et al. [2024] report comparable results for sex classification using 3D CNNs, but differences in the datasets used make direct performance comparisons difficult. Similarly, Cardenas-Hernandez et al. [2025] and ElNakieb et al. [2021] both address ASD classification, yet Cardenas-Hernandez et al. [2025] achieves accuracy comparable to the best-performing model in ElNakieb et al. [2021] despite evaluating only two models rather than eight. Even though the two studies use the same dataset and evaluate the same models and feature types, the reported performance of those models is not identical. This shows that it is often unclear how much variability comes from genuine differences in microstructural information versus inconsistencies in preprocessing, feature extraction, and evaluation design. In this section we review the literature across four dimensions relevant to DiffBench (i) microstructural representations, (ii) prediction tasks, (iii) predictive models, and (iv) benchmarking practices.

### 2.2.1 Microstructural representations and feature design

How microstructural information is represented is as consequential as the choice of predictive model, yet it has received comparatively less systematic attention. A predictive pipeline typically combines (i) a microstructural model, (ii) a spatial registration strategy, and (iii) a feature aggregation scheme, each introducing independent sources of variability.

**Microstructural models.** Diffusion MRI microstructure can be represented using a range of mathematical models that differ in biological specificity, acquisition requirements, and sensitivity to tissue complexity. This paragraph reviews the main classes of microstructural representations used in diffusion MRI, with the goal of establishing the design space explored in DiffBench. A detailed description of the specific implementations used in our framework is provided in the Methods section. Diffusion models range from tensor-based approaches (DTI [Basser et al., 1994], DKI [Jensen et al., 2005]) to multi-compartment models that explicitly represent different tissue components (e.g., NODDI [Zhang et al., 2012], SANDI [Palombo et al., 2020]), model-free representations based on spherical harmonics [Tournier et al., 2004] and advanced propagator models as MAP-MRI [Özarslan et al., 2013]. Tensor-based models provide

compact, interpretable scalars but rely on Gaussian assumptions that do not hold in regions with complex fibre geometry.

**Spatial registration.** Cross-subject spatial registration strategies differ substantially between tissue types. For white matter, alignment is typically achieved by warping each subject's diffusion scalar maps to a common template space [Smith et al., 2006], or by registering to a standardised white matter atlas such as the JHU atlas [Mori et al., 2008], allowing corresponding pathways to be compared across subjects despite individual anatomical variability. For cortical gray matter, surface-based alignment better respects cortical folding geometry and reduces partial-volume contamination [Greve and Fischl, 2009], but surface-projected diffusion features remain comparatively rare in the literature, despite evidence that gray matter microstructure carries independent predictive value [Menon et al., 2020; de Mooij et al., 2018]. Overall, studies tend to focus on white or gray matter registration separately, and direct comparisons of registration strategies across both tissue types remain relatively limited [Chen et al., 2024; Menon et al., 2020; Kotikalapudi et al., 2025].

**Spatial representations.** Once spatially aligned, diffusion-derived microstructural measures can be represented at different levels of anatomical granularity.

Voxel-wise representations preserve maximal spatial specificity and enable localisation of predictive signal at fine anatomical scales. However, they introduce extremely high dimensionality (e.g., in HCP, each subject is represented by 3D volumes of size 145×174×145, i.e., 3.66 million voxels per volume) and remain sensitive to residual registration errors, particularly in multi-site datasets [Chen et al., 2024].

Region-of-interest (ROI) representations aggregate voxel-wise values within predefined anatomical parcels or atlases, substantially reducing dimensionality while improving robustness and interpretability. ROI aggregation is therefore particularly attractive for machine learning approaches operating in relatively small neuroimaging cohorts. Prior work has shown that region-specific microstructural summaries derived from MAP-MRI can capture behaviourally relevant variation in cognitive control [Menon et al., 2020].

Tract-based representations further constrain features along anatomically defined white matter pathways. TBSS [Smith et al., 2006] projects scalar maps onto a group-wise skeleton, enabling tract-level comparison with reduced sensitivity to residual registration error. Atlas-derived tract summaries, such as those computed from the JHU white matter atlas [Mori et al., 2008], preserve pathway-specific information while remaining computationally tractable. More detailed tractometry frameworks such as AFQ-Insight [Kruper et al., 2025] extend this approach using within-tract feature profiles coupled with QSIPrep-based preprocessing pipelines.

For cortical gray matter, surface-based representations provide anatomically informed alternatives to volumetric averaging. Diffusion scalar maps can be projected onto the cortical surface and aggregated within parcels such as the Schaefer atlas [Schaefer et al., 2018], yielding compact representations that better respect cortical

geometry.

### 2.2.2 Diffusion MRI microstructure in predictive modelling

After deriving microstructural representations from diffusion MRI, they can be directly used as input features for predictive modelling tasks.

**Sex and age prediction.** Sex classification and brain age estimation are canonical benchmarks in neuroimaging, providing well-defined targets with large reference cohorts and extensive prior work across multiple imaging modalities, particularly structural T1-weighted MRI. Deep learning on volumetric diffusion data has demonstrated that white matter contains discriminative signal for sex classification [Chen et al., 2024], and end-to-end representations can avoid handcrafted feature design while retaining spatial specificity. However, a persistent limitation across such studies is the absence of non-diffusion or structural MRI baselines. Structural MRI or brain size has repeatedly been shown to achieve strong predictive performance for sex [Mendes et al., 2021; Dibaji et al., 2024] suggesting that much of the discriminative signal may be from macroscopic anatomical differences rather than microstructural properties.

For brain age, diffusion MRI captures microstructural variation complementary to volumetric morphology [Chen et al., 2020a; de Mooij et al., 2018]. Yet inter-site variability and acquisition differences are major sources of performance instability, often contributing as much variance as model choice [Chen et al., 2020a]. Domain adaptation strategies can mitigate these effects, but they require multi-site data that is not always available. Non-diffusion contrasts such as T2-weighted imaging and structural MRI have also been shown to support brain age prediction [Hwang et al., 2021; Peng et al., 2021], providing an important reference point. Since the $b_0$ signal is T2 contrast, similar performance between $b_0$ and diffusion-derived metrics would suggest that predictive signal reflects macroscopic tissue properties rather than microstructure specifically.

**Clinical prediction.** In clinical neuroimaging, DTI-derived metrics such as FA and MD remain the most commonly used features for classification and prognosis tasks across conditions including ASD, Alzheimer's disease, and cerebrovascular disease [Cardenas-Hernandez et al., 2025; Egle et al., 2022]. A recurring observation is that simpler representations often perform comparably to more complex ones: FA and MD can match the performance of fuller DTI feature sets due to redundancy among correlated tensor-derived scalars [Cardenas-Hernandez et al., 2025], and MD median alone performs competitively with more elaborate diffusion quantification strategies across cohorts [Egle et al., 2022]. These findings suggest that the marginal value of more complex representations is task- and dataset-dependent, and cannot be assumed without controlled evaluation.

**Cognitive prediction.** Beyond clinical tasks, diffusion microstructure has also been linked to cognitive phenotypes to understand why individuals differ in how they think, remember, and reason, and how these differences relate to variation in brain structure and function [Genon et al., 2022; He et al., 2020]. Early work using tractography reported associations between white matter integrity and general intelligence across major white matter tracts [Penke et al., 2012; Filley and Fields, 2016]. Building on this, more recent studies have extended this framework to predictive modelling in which diffusion-derived features are used to estimate cognitive performance at the individual level, although reported effect sizes often fail to replicate across independent cohorts [Ooi et al., 2022]. At the level of diffusion-specific representations, advanced multi-shell and more expressive representations such as MAP-MRI can capture behaviourally relevant variation in cortical gray matter beyond what DTI provides [Menon et al., 2020], and large-scale replicability analyses [Kotikalapudi et al., 2025] show that brain-phenotype associations based on diffusion metrics are statistically fragile.

Multimodal studies have compared the predictive utility of anatomical, diffusion, and functional connectivity features can all contain behaviorally relevant information, with functional connectivity often achieving the highest predictive performance across large cohorts such as HCP and ABCD [Ooi et al., 2022]. Following this cohort variability, recent large-scale replication analyses have shown that associations between diffusion-derived metrics and cognitive or behavioural phenotypes are often unstable across cohorts, with no single metric consistently generalising across datasets [Kotikalapudi et al., 2025]. Taken together, these findings suggest that while diffusion MRI contains signals relevant to cognition, the robustness and generalisability of diffusion-based cognitive prediction remain limited, particularly when evaluated across datasets, modelling choices, and analytical pipelines. This motivates the need for standardised, multi-dataset evaluation frameworks capable of systematically assessing model stability and generalisation rather than relying on single-cohort studies to move the diffusion mri field beyond proxy tasks.

### 2.2.3 Machine learning and deep learning approaches for dMRI

After outlining the primary predictive tasks derived from diffusion MRI, we next review the range of modelling approaches employed, spanning machine learning and deep learning techniques.

Machine learning methods, principally SVMs, random forests, and linear models, remain widely used in dMRI-based prediction due to their robustness in small-to-medium neuroimaging datasets and their interpretable feature spaces. These methods have been applied across tasks including AD classification, ASD diagnosis, and cognitive prediction [Wen et al., 2021; Cardenas-Hernandez et al., 2025; Menon et al., 2020], typically with ROI-averaged or voxel-wise diffusion scalars as inputs. Hyperparameter selection in such studies is frequently ad hoc or based on trial and error, and evaluation protocols vary substantially across studies, limiting cross-study com-

parisons.

Deep learning methods have increasingly been applied to neuroimaging data, offering greater representational expressivity at the cost of more data needed for better predictions and higher computational requirements. Recent studies have demonstrated the feasibility of learning predictive representations directly from diffusion-derived measures. Chen et al. [2024] evaluated 2D CNNs, 3D CNNs, and Vision Transformers for sex classification in the Human Connectome Project using white-matter maps of fractional anisotropy (FA), mean diffusivity (MD), and mean kurtosis (MK). Their results showed that both convolutional and transformer-based architectures achieved high classification accuracy, while Vision Transformers trained from scratch performed substantially worse, highlighting the importance of pretraining. Transfer learning has also been explored for diffusion MRI-based brain age prediction, where Chen et al. [2020a] pretrained convolutional networks on the Cam-CAN cohort and transferred them to independent datasets, achieving mean absolute errors between 4.78 and 5.64 years across external cohorts. More recently, Dibaji et al. [2024] employed a 3D CNN on T1-weighted MRI for sex classification across multiple cohorts, demonstrating robust generalisation across heterogeneous datasets. Collectively, these studies suggest that deep learning can effectively leverage diffusion MRI microstructural information for predictive modelling, although differences in datasets, preprocessing pipelines, representations, and evaluation protocols make direct comparison across methods challenging.

### 2.2.4 Benchmarking and Reproducibility

Reproducibility challenges in neuroimaging extend beyond individual studies. Brain-behaviour effect sizes are typically small in population-representative samples, and reported associations frequently fail to replicate across independent cohorts [Genon et al., 2022]. Prediction accuracies vary from near-chance [ElNakieb et al., 2021] to near-perfect [ElNakieb et al., 2021; Cardenas-Hernandez et al., 2025] across the literature, with methodological variability (as microstructural computation, preprocessing, tissue type representation) identified as a primary driver rather than genuine differences in the underlying biology [Wu et al., 2023]. External validation requires substantially larger samples than are typically available, and effect sizes are systematically inflated when evaluated on small held-out sets [Rosenblatt et al., 2024].

Within dMRI, Wen et al. [2021] proposed a reproducible framework for AD classification, demonstrating that methodological choices including feature type, registration, and smoothing influence classification performance as much as the classifier itself. Similarly, preprocessing decisions in sex and age prediction from white matter dMRI affect results as strongly as representation choice, with no single configuration consistently dominating [He et al., 2022]. These findings establish that methodological choices are a primary source of performance variability, but existing studies each vary only one or two dimensions of their methodology (e.g., feature type, preprocessing) within a single task, dataset, and tissue type, limiting the generalisability of their

conclusions. This has led to an ongoing effort within the diffusion MRI community to identify methodological choices that are consistently robust across tasks and datasets.

Table 2.1 summarises key studies applying diffusion MRI microstructure to predictive modelling across various tasks and datasets, highlighting the diversity of approaches and the gaps that DiffBench is designed to address.

Table 2.1: Summary of predictive modelling studies using dMRI across datasets-tasks

| Paper | Dataset | Task | Microstr. | Model | Tissue | Metric | Value | Code |
|---|---|---|---|---|---|---|---|---|
| Chen et al. (2024) | HCP | Sex prediction | FA, MD, MK | 2D CNN, 3D CNN, ViT | WM | Accuracy | 0.66–0.91 | No |
| Menon et al. (2020) | HCP | Cognitive control | MAP-MRI (RTOP) | CCA, LR | GM | $R^2$ | 0.31 | No |
| Kotikalapudi et al. (2025) | HCP | Cognition | DTI | Linear models | WM | Replicability | 36% | Yes |
| Chen et al. (2020) | NTUH, IXI-HH, IXI-Guys | Brain age | DTI | Transfer learning (training Cam-CAN) | WM | MAE years | 4.78, 5.35, 5.64 | Yes |
| de Mooij et al. (2018) | Cam-CAN | Brain age | FA, GM volume | Linear regression | GM, WM | Correlation | 0.48 | No |
| Dibaji et al. (2024) | Cam-CAN, ADNI, OASIS3, CC359 | Sex classification | FA, GM volume | 3D CNN | T1-weighted MRI | Correlation | 0.92, 0.86, 0.858, 0.851 | Yes |
| Cardenas et al. (2025) | ABIDE II | ASD classification | DTI (FA, MD) | Random Forest | WM | Accuracy | 0.9737 | No |
| Hwang et al. (2021) | Internal, IXI | Brain age | T2-weighted MRI | CNN | Whole brain | MAE years | 4.22, 9.96 | No |
| ElNakieb et al. (2021) | ABIDE II | ASD classification | DTI (FA, MD, AD, RD, TSkew) | LSVM | WM | Mean Acc. | 0.99 | No |

### 2.2.5 Gaps Addressed by DiffBench

Across the literature reviewed above, we identified four methodological gaps.

**Absence of non-diffusion baselines.** Most studies do not compare diffusion-derived features against non-diffusion inputs such as the $b_0$ signal, making it impossible to determine how much predictive performance reflects microstructure-specific information versus macroscopic anatomical or T2 contrast effects.

**Single-tissue evaluations.** Gray and white matter are rarely evaluated within the same framework, despite evidence that they carry complementary predictive signals and may behave differently across tasks and representations.

**Single-dataset and single-task evaluations.** The majority of studies rely on one cohort and one prediction target, providing limited evidence of cross-cohort generalisability or task-dependence of representation choices.

**Heterogeneous and non-reproducible evaluation practices.** Hyperparameter selection strategies, cross-validation designs, and preprocessing choices differ substantially across studies, preventing reliable comparison of results across the literature.

DiffBench addresses all four gaps within a single unified framework. It evaluates gray and white matter jointly across three cohorts and three prediction tasks, includes non-diffusion $b_0$ baselines alongside microstructural metrics, applies uniform preprocessing and nested cross-validation across both classical and deep learning model families, and is released as an open-source pipeline to support reproducible comparison and extension by the community.

## 2.3 Methods

In this section, we introduce **DiffBench**, a comprehensive and extensible benchmarking framework for evaluating predictive modelling approaches using diffusion MRI (dMRI). The benchmark is designed to address a key challenge in the field: predictive performance is often influenced not only by the choice of model, but also by differences in preprocessing pipelines, feature extraction strategies, data representations, and evaluation protocols. As a result, comparisons across studies are frequently difficult to interpret and reproduce. DiffBench provides a unified framework in which these sources of variability are controlled, enabling systematic and fair comparisons between alternative modelling approaches.

The overall organisation of the benchmark is illustrated in Figure 2.1. DiffBench is structured as a modular pipeline comprising five main components: datasets, preprocessing, data preparation, prediction and validation, and software utilities. The workflow begins with publicly available dMRI datasets and associated phenotypic variables. A preprocessing pipeline is then applied to generate diffusion derivatives across datasets. These derivatives are subsequently transformed into multiple data representations, including microstructural features and spatially represented imaging data, which serve as inputs to predictive models.

The benchmark supports both machine learning and deep learning approaches and evaluates them under a consistent validation framework for regression and classification tasks. By standardising preprocessing, microstructural feature generation, and evaluation procedures, DiffBench aims to isolate the contribution of the predictive

models themselves, allowing differences in performance to be attributed more directly to modelling choices rather than upstream methodological variations. Finally, the framework is implemented as a modular and reproducible software package that facilitates the integration of new datasets, feature representations, and predictive models.

The following sections describe each component of the benchmark in detail.

### 2.3.1 Datasets

To ensure reproducibility, DiffBench leverages three publicly available datasets covering both healthy and clinical populations.

Dataset selection was driven by three criteria: (i) public availability, enabling reproducibility of the benchmark; (ii) sufficient cohort size to support predictive modelling; and (iii) widespread use in the literature, facilitating comparison with previous studies.

We identified 6 datasets that satisfy these criteria, including UK Biobank, ABCD, HCP, Cam-CAN, ABIDE, and ADNI. Due to data-access and administrative constraints, UK Biobank and ABCD were not included in the current release of the benchmark. Consequently, DiffBench focuses on HCP and Cam-CAN as representative large-scale datasets of healthy populations, while ABIDE II was included to provide a clinically relevant classification task. The combination of these datasets enables the evaluation of predictive models across diverse populations, age ranges, and prediction objectives, including both regression and clinical classification tasks.

The following sections describe the included datasets, their acquisition parameters, and the prediction tasks considered.

**HCP**: The Human Connectome Project (HCP) Young Adult dataset [Van Essen et al., 2013] targets sex and age. It includes 1,036 healthy subjects after filtering (474 males, 562 females; ages 22-37). HCP data were acquired on a customized Siemens 3T Connectome scanner using a multi-shell Spin-Echo EPI sequence (TR/TE = 5520/89.5 ms, 1.25 mm isotropic, multiband factor 3; $b$-values = 1000, 2000, 3000 s/mm$^2$, 90 directions/shell).

**Cam-CAN**: The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) dataset [Shafto et al., 2014] targets sex and age. It includes 629 healthy subjects (316 males, 313 females; ages 18-88), scanned on a Siemens 3T TIM Trio using a 2D twice-refocused SE EPI sequence (TR/TE = 9100/104 ms, 2.0 mm isotropic; $b$-values = 1000, 2000 s/mm$^2$, 30 directions each).

**ABIDE II**: The Autism Brain Imaging Data Exchange II (ABIDE II) dataset [Di Martino et al., 2017] targets Autism Spectrum Disorder (ASD) diagnosis. It includes 218 subjects (186 males, 32 females; ages 5-64), including 119 with ASD and 99 controls. As a multi-site dataset, ABIDE II data were acquired on various 3T scanners with highly heterogeneous, site-specific single-shell diffusion protocols.

Across the included sites, spatial resolutions range from 0.94 to 3.0 mm, TR/TE from 5200-20244/78-101 ms, $b$-values include 1000, 1500, and 2500 s/mm$^2$, and diffusion directions vary between 32 and 64. This variability provides a realistic scenario for evaluating model robustness to scanner heterogeneity.

### 2.3.2 DiffBench Preprocessing

A key design principle of DiffBench is to control for preprocessing variability across datasets and models. By applying a unified, fixed preprocessing pipeline to all datasets, differences in predictive performance can be attributed to model architecture and microstructural representation rather than to upstream processing choices. Evaluating or optimising the preprocessing pipeline itself is explicitly outside the scope of this benchmark.

Except for HCP, which is already preprocessed, dMRI data are preprocessed using the package[1] which containerises various tools such as dipy [Garyfallidis et al., 2014], sMRIPrep [Esteban et al., 2019], FSL [Smith et al., 2004], ANTs [Avants et al., 2009], FreeSurfer [Dale et al., 1999] and AFN [Cox, 1996]. The preprocessing follows a standard dMRI pipeline addressing noise, artifacts, geometric distortions, and anatomical alignment in sequence. Images are first denoised with the Marchenko-Pastur PCA (`MPPCA`) method [Veraart et al., 2016b; Cordero-Grande et al., 2019] implemented in `DiPy`. This denoising procedure identifies the principal components corresponding to noise and then removes them from the dMRI data [Veraart et al., 2016b]. After `MPPCA` denoising, Gibbs ringing artifacts [Veraart et al., 2016a] are also removed with `DiPy`. Next, we correct the distortions due to inhomogeneities of the magnetic field with FSL's `eddy` correction [Andersson and Sotiropoulos, 2016b]. Finally, this denoised and distortion-corrected dMRI data are registered to the anatomical T1w image preprocessed with `sMRIPrep`, using FreeSurfer's *bbregister* [Greve and Fischl, 2009] with 12 degrees of freedom and boundary-based registration cost function.

Figure 2.2 illustrates the preprocessing pipeline applied to all datasets in DiffBench except HCP, which is already preprocessed.

### 2.3.3 Data Preparation

**Microstructural models.** Acquired diffusion MRI signal does not directly reflect underlying biological structure. To extract biological meaningful information, mathematical models are needed to relate the diffusion signal to the microstructural properties of the brain. These mathematical models vary significantly according to the assumptions they make about the diffusion process and the level of microstructural information they can capture. Simple models such as DTI assume Gaussian diffusion and provide robust summary measures that can be estimated from relatively limited data. On the other hand, more advanced representations, such as MAP-MRI, relax the Gaussian assumption and reconstruct the diffusion propagator from basis

[1] `https://github.com/gabrielapgomezji/diff_benchmark`

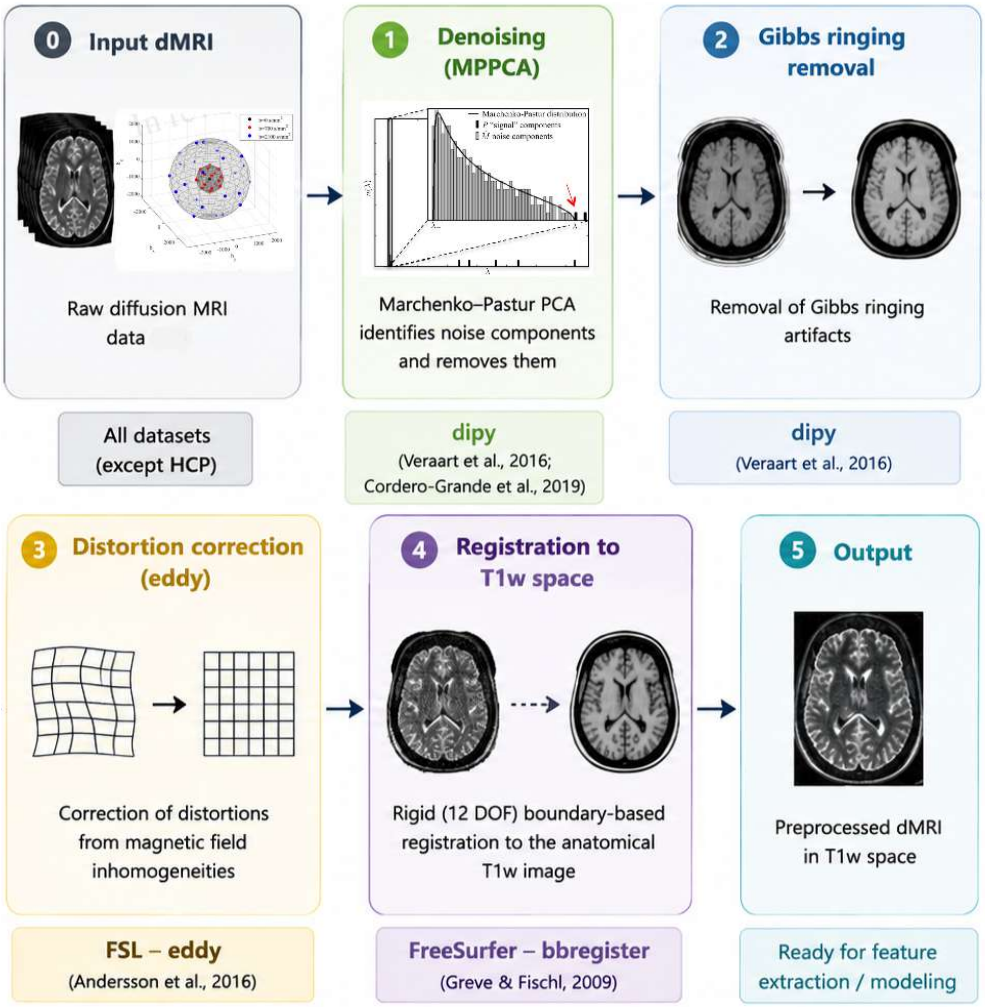



Figure 2.2: **Standardised preprocessing pipeline for DiffBench.** This figure shows the sequence of preprocessing steps applied to all DiffBench datasets except HCP, which is provided already preprocessed and is therefore not included.

functions in q-space, enabling the derivation of richer descriptors of tissue microstructure and diffusion complexity. In Diffbench, we focus on three widely used diffusion representations which provide complementary information while remaining broadly applicable and computationally efficient.

- **Diffusion Tensor Imaging (DTI)** [Basser et al., 1994] characterizes diffusion with a 3D Gaussian distribution represented mathematically by a 3x3 symmetric tensor. From this tensor it is possible to derive metrics such as Fractional Anisotropy (FA) or Mean Diffusivity (MD). DTI's Gaussian nature makes it work well in regions where fiber orientations are uniform, but it struggles to capture complex microstructural configurations such as crossing fibers [Mori and Van Zijl, 2002] since it is hard to capture multiple directions.

- **Diffusion Kurtosis Imaging (DKI)** [Jensen et al., 2005] extends DTI by accounting non-Gaussian diffusion behaviour. It includes a fourth-order kurtosis tensor that captures deviations from Gaussian diffusion, providing additional information about tissue complexity and heterogeneity. Metrics derived from DKI are Mean Kurtosis (MK), Axial Kurtosis (AK) and Radial Kurtosis (RK) among others, since they capture non-Gaussian diffusion effects, are therefore, often more sensitive to microstructural alterations than DTI metrics.

- **Spherical Harmonics (SH)** [Tournier et al., 2004] is a model-free representation without assuming a specific compartmental model. It represents diffusion as continuous functions on the surface of a sphere; it extends the diffusion signal into a weighted sum of basis functions defined on angular directions. SH-based approaches can capture complex fiber configurations, such as crossing fibers, and provide a more flexible representation of the diffusion signal compared to DTI and DKI.

**Microstructural representation.** To characterise the diffusion signal, DiffBench implements a set of rotationally invariant microstructural metrics derived from established diffusion models. Although each model provides multiple derived metrics (e.g., fractional anisotropy for DTI, or axial and radial kurtosis for DKI), this benchmark focuses on a reduced set of representative descriptors. Restricting the analysis to these descriptors allows for consistent comparison across models while limiting redundancy among highly correlated diffusion metrics. These metrics were selected because of its wide use in previous studies [Hwang et al., 2021; He et al., 2022; Chen et al., 2023a; Cardenas-Hernandez et al., 2025], that range from diffusion magnitude to more complex descriptors. All metrics were computed using `DiPy` [Garyfallidis et al., 2014], which also provides straightforward access to additional microstructural metrics not included in this benchmark. The selected metrics, their fitting procedures, and any dataset-specific considerations are described below.

- **Baseline ($b = 0$):** All volumes with $b \leq 50$ s/mm$^2$ were averaged to produce a T2-weighted baseline image. This metric requires no model fitting and was computed identically across all three datasets. The ventricular mean $b_0$ was used as the normalisation reference. This signal is included as a reference to disentangle diffusion-driven predictive effects from underlying T2 contrast.

- **Mean Diffusivity (MD):** MD is derived from Diffusion Tensor Imaging (DTI) [Basser et al., 1994] and reflects the overall magnitude of water diffusion. Tensors were fitted using the weighted least-squares (WLS) estimator (*TensorModel* in `DiPy`). For HCP and Cam-CAN, all available shells were used jointly for tensor estimation; for ABIDE II, the single available shell was used. The ventricular mean MD was used as the normalisation reference.

- **Mean Kurtosis (MK):** MK is derived from Diffusion Kurtosis Imaging (DKI) [Jensen et al., 2005] and quantifies deviations from Gaussian diffusion,

serving as a proxy for microstructural heterogeneity and tissue complexity. Kurtosis tensors were fitted using WLS (*DiffusionKurtosisModel* in `DiPy`). Stable estimation requires at least two non-zero b-value shells. The datasets HCP ($b = 1000, 2000, 3000$ s/mm$^2$) and Cam-CAN ($b = 1000, 2000$ s/mm$^2$) satisfy the requirement, whereas ABIDE II single-shell acquisitions do not; MK was therefore not computed for ABIDE II. The ventricular mean MK was used as the normalisation reference.

- **Spherical Harmonics (SH) Power:** SH power is a model-free, rotationally invariant scalar summary of the diffusion-weighted signal. Only non-$b_0$ gradient directions are retained. Spherical harmonic coefficients are fitted to the normalised signal using the *sf_to_sh* function from `DiPy` with a *descoteaux07* basis and maximum order $l_{\max} = 6$. SH power is defined as the $\ell_2$-norm of all SH coefficients per voxel, $P = \|\mathbf{c}\|_2$, which captures the total angular energy of the diffusion signal across all harmonic orders, combining mean signal level and angular variation. We note that this differs from a pure anisotropy index, which would exclude the zeroth-order ($\ell = 0$) coefficient and isolate angular variation alone. Our choice of retaining all harmonic orders is deliberate. $P$ provides mean signal level and angular variation, but this makes it sensitive to tissue properties that no single model-based metric captures in isolation. As a model-free descriptor, it serves as a complementary baseline against which the model-based metrics can be benchmarked. To ensure consistent angular sampling across datasets, only the non-$b_0$ volumes of the first shell ($b = 1000$ s/mm$^2$) were used for HCP and Cam-CAN; for ABIDE II, all available non-$b_0$ volumes were used. Because SH power scale differs fundamentally from diffusivity-based metrics, ventricular normalisation was not applied; the $b_0$-normalisation step serves as a built-in intensity reference. SH power was computed for all three datasets.

Table 2.2 summarises the characteristics of the metrics in the benchmark.

Table 2.2: Comparison of microstructural metrics included in DiffBench.

| Metric | What it measures | Model | Shell requirement |
|---|---|---|---|
| $b_0$ | T2-weighted tissue water content; no diffusion weighting | None | $b \approx 0$ only |
| MD | Mean water displacement; sensitive to cellularity, oedema, free water | DTI (WLS) | Single shell |
| MK | Deviation from Gaussian diffusion; proxy for microstructural heterogeneity | DKI (WLS) | Two shells minimum |
| SH power | Total diffusion-weighted signal energy: mean level + angular variation | None (model-free) | Single shell |

To reduce inter-subject instensity variability, mitigate free-water effects and scanner-dependent intensity scaling [Pasternak et al., 2009], DiffBench implements ventricular normalisation: for each subject and each microstructural representation, voxel-wise diffusion measures are divided by the mean value extracted from cerebrospinal fluid within the ventricles. Finally, extreme values are clipped at the 99th percentile to ensure numerical stability during model training.

**Spatial registration.** We apply distinct spatial registration strategies for gray and white matter to enable anatomically consistent comparisons across subjects and datasets while respecting tissue-specific geometry.

*Gray Matter.* Gray matter microstructure is represented on the cortical surface reconstructed from T1w images with FreeSurfer. Cortical meshes (white and pial surfaces) are parcellated with the Desikan–Killiany atlas [Desikan et al., 2006]. Diffusion scalar maps are computed in the native diffusion space and rigidly aligned to the anatomical image. Voxelwise diffusion measures are then projected onto the individual cortical surface with ribbon-constrained sampling, preserving vertex-wise resolution ($\sim$ 64k vertices per hemisphere) [Greve and Fischl, 2009]. This surface-based strategy respects cortical geometry, reduces partial-volume effect, maintains high spatial specificity and enables consistent spatial registration across subjects.

*White Matter.* White matter microstructure is represented with Tract-Based Spatial Statistics (TBSS) [Smith et al., 2006]. Diffusion scalar maps are nonlinearly registered to standard space and projected onto a group-wise white matter skeleton, which captures the core of major fiber bundles and minimizes residual misalignment and partial-volume effects. Then, we summarize microstructural values by extracting the mean for 48 major tracts from the JHU White-Matter Tractography Atlas [Mori et al., 2008]. This normalisation enforces anatomical correspondence across subjects and reduces variability associated with subject-specific tractography. While this approach sacrifices voxel-level granularity compared to the cortical surface representation, it provides a compact, reproducible, and anatomically interpretable feature set that is robust in multi-site settings.

**Data Representation.** Two complementary data representations of the computed microstructure metrics are provided to support both deep learning and machine learning pipelines. Volumetric tensor representations and a flatterned feature vector representation. The volumetric tensor preserves anatomical detail, while the parcellated vector representation provides compact and robust features across datasets and acquisition protocols.

*3D Tensor Representation.* For the 3D tensor representation, the microstructural metrics scalar maps are needed. All scalar maps are first computed in native diffusion space and stored as 3D NIfTI images in the subject's anatomical space. These volumetric maps constitute the tensor based representation, preserving full spatial structure enabling direct use with 3D and 2D architectures.

*Array Representation.* The vectorized representation is derived from the volumetric maps and specific for each spatial registration representation. Computed by projecting the scalar maps onto anatomically defined structures, following the spatial registration framework described above. For white matter, scalar maps are projected onto the TBSS skeleton to ensure alignment across subjects in standard space. Mean values are then projected to the 48 tracts of the JHU ICBM-DTI atlas, producing tract-specific features that summarize microstructure along major fiber bundles. While gray matter is organised in left and right hemisphere, tracts are further organ-

ised including also a midline partition. For gray matter, diffusion metrics are sampled onto the cortical surface using ribbon-constrained projection between the white and pial surfaces reconstructed with Free Surfer. Vertex-wise values are then aggregated according to the Schaefer-2018 atlas (*fsLR* or *fsaverage* space depending on dataset) to obtain region-wise summaries per hemisphere. The choice of region-wise representation was motivated by the high dimensionality of the vertex-wise representation which would lead to overfitting in machine learning models and to allow a more direct comparison with the tract-based representation of white matter in terms of representation dimensionality. This produces a compact feature vector indexed by cortical parcels, suitable for machine learning models. Further studies could explore the use of vertex-wise or embedding representations for machine learning models.

### 2.3.4 Prediction and validation

**Model descriptions.** DiffBench includes a broad range of predictive models previously used in dMRI studies [Wen et al., 2021; Cardenas-Hernandez et al., 2025], from machine learning pipelines to foundation models.

*Machine Learning.* Models (Ridge, Lasso, Random Forests, and SVM), implemented in `scikit-learn`, were evaluated both with the original features and with PCA-based dimensionality reduction applied prior to model fitting [Menon et al., 2020]. All models follow a common training protocol; input features are standardised to zero mean and unit variance using `StandardScaler` prior to fitting, ensuring that regularisation penalties are applied uniformly across features with different scales. Hyperparameters were selected via exhaustive grid search (`GridSearchCV`) with 5-fold cross-validation, optimising the corresponding metric according to the prediction task.

For Ridge and logistic regression with $\ell_2$ penalty, the regularisation strength was searched over a log-uniform grid of 21 values for $\alpha$ (or equivalently $C$ for the classification formulation). Lasso regression and $\ell_1$-penalised logistic regression (fitted with the `saga` solver) were tuned over the same grid. When PCA-based dimensionality reduction was applied, the number of retained components was optimised alongside the regularisation hyperparameter.

Random Forest models were tuned over tree depth, the fraction of features considered at each split, and the minimum number of samples per leaf. When paired with PCA, the number of components was additionally searched over the same number of components as Ridge and logistic models. Support Vector Machines used linear kernels (*LinearSVC* and *LinearSVR*) to scale to high-dimensional inputs, with the regularisation parameter $C$; when paired with PCA, we also defined the number of components. Compared with other models, the SVM search space was intentionally reduced due to the computational cost associated with exhaustive hyperparameter optimisation on high-dimensional data.

*Deep Learning.* Deep learning approaches have become popular for diffusion MRI

analysis [Chen et al., 2023a; Dibaji et al., 2024]. To consider these types of methods within DiffBench, we include three representative deep learning models considering recent foundation models and supervised volumetric architectures.

DINOv2 [Oquab et al., 2023] is a self-supervised Vision Transformer (ViT) trained on a large-scale image collection. The model partitions each image into patches that are processed by a transformer encoder to produce patch embeddings and a global image representation.

Curia [Dancette et al., 2025] adopts the same Vision Transformer architecture and self-supervised learning framework from DINOv2, but is pretrained on clinical CT and MRI images collected from routine clinical practice. As a result, Curia provides medical imaging representations learned from large-scale radiological data.

For both DINOv2 and Curia, diffusion volumes are decomposed into 2D slices extracted along the x-axis. DINOv2 was trained on RGB images. Therefore, to ensure compatibility with the pretrained weights, each diffusion slide was repeated across the three channels of the input image. The same strategy was adopted for Curia, following the 2D-CNN framework of [Chen et al., 2024]. Slice-level embeddings are obtained from the frozen transformer backbone and aggregated before being passed to a task-specific prediction head. Only the prediction head is trained, while all pretrained backbone parameters remain fixed throughout optimisation.

In contrast, MedicalNet [Chen et al., 2019] is a 3D convolutional neural network based on volumetric ResNet architecture. MedicalNet extends standard residual networks to three-dimensional medical images through 3D convolutional kernels and residual blocks, enabling hierarchical feature extraction directly from volumetric inputs. Unlike the transformer-based models, MedicalNet was pretrained using supervised learning on a large collection of annotated medical imaging datasets and subsequently trained end-to-end on each downstream task. Full diffusion volumes are provided as input, and all network parameters are updated during training.

The decision to freeze DINOv2 and Curia while training MedicalNet end-to-end was motivated by computational considerations. Consequently, the benchmark evaluates these models as feature extractors rather than fully adapted diffusion MRI foundation models. This represents an important limitation of the current study, as fine-tuning large pretrained transformers on diffusion MRI data could potentially output improved performance.

All deep learning models share a common training backend implemented in `PyTorch`. Classification tasks are optimised using cross-entropy loss, whereas regression tasks use mean squared error loss. Models are trained using Adam with exponential learning-rate decay (factor $\gamma = 0.95$ per epoch) and $\ell_2$ regularisation through weight decay. Training and evaluation follow a cross-validation protocol, with a fixed fraction of the training data held out as a validation set for monitoring purposes (20% by default).

Further details are fully available in the codebase[2].

[2] https://github.com/gabrielapgomezji/diff_benchmark

**Prediction tasks and evaluation metrics.** DiffBench includes both classification and regression tasks, spanning three cohorts and two prediction targets.

*Classification.* Binary classification tasks include Autism Spectrum Disorder (ASD) diagnosis (ASD vs. Typically Developing) using the ABIDE II cohort [Cardenas-Hernandez et al., 2025], and sex prediction in the HCP and Cam-CAN cohorts [Chen et al., 2024; Dibaji et al., 2024]. Performance is reported using balanced accuracy, which accounts for class imbalance. Machine Learning models optimise balanced accuracy directly during hyperparameter selection, while deep learning models are trained with cross-entropy loss and evaluated post-hoc using balanced accuracy. This distinction is due to differences in the optimization mechanisms of the respective model types. Machine Learning algorithms do not typically rely on gradient-based loss minimization during hyperparameter tuning, allowing balanced accuracy to be used directly as the model selection criterion. However, deep learning models require a differentiable objective for gradient-based optimization, making cross-entropy loss a standard training choice, while balanced accuracy is retained as the primary evaluation metric to ensure consistent comparison across methods.

*Regression.* Brain age prediction [Chen et al., 2020a; de Mooij et al., 2018; Hwang et al., 2021] is evaluated as a regression task on the healthy Cam-CAN cohort. Performance is quantified with mean absolute error (MAE) and $R^2$. Machine Learning models select hyperparameters by minimising MAE under 5-fold cross-validation; deep learning models are trained with mean squared error (MSE) loss on whitened targets, with predictions rescaled to the original age range at inference time.

**Validation Methodology.** All models are evaluated under a common 5-fold cross-validation scheme. Data-driven operations are fitted, for each fold, exclusively on the training set and applied to the held-out test fold, preventing data leakage. For machine learning models, hyperparameters are additionally tuned via an inner 5-fold cross-validation loop nested within each training fold. For deep learning models, a validation set is further held out from the training fold (20% by default) for monitoring during training, while the test fold remains entirely unseen until final evaluation. Both classification and regression folds are stratified by sex to ensure representative splits across tasks. To guarantee reproducibility, random seeds are fixed throughout

### 2.3.5 Methodological Limitations

Although the proposed benchmark covers end-to-end the diffusion MRI predictive modelling pipeline, from raw data to model evaluation, and includes a variety of models, datasets and microstructural representations, there are methodological limitations that need to be addressed.

The focus of this work is the prediction power of diffusion MRI microstructural representations. Consequently, preprocessing steps such as denoising, motion correction, and eddy current correction are not explicitly analysed, despite their known influence on diffusion metrics and downstream analyses [Andersson and Sotiropoulos, 2016b;

Veraart et al., 2016b]. Regarding the microstructural representations, we included a limited set of commonly used diffusion metrics and a model-free SH power descriptor to make it easier to compare across tissue types and with existing literature. However, there are many other microstructural representations and models that could be included in future extensions of the benchmark, including high-order approaches such as MAP-MRI or biophysical models (e.g., NODDI, CHARMED). Regarding spatial registration and data representation, we prioritise anatomically consistent representations for gray and white matter. While this choice ensures methodological consistency across experiments, it also limits a more exhaustive comparison of alternative representation strategies specific to each tissue type. The current benchmark further focuses exclusively on predictive performance. Analyses of feature representations and model interpretability (for linear methods), or exploring parameter-efficient and full-model fine-tuning strategies (for Deep Learning) are left for future work. The prediction tasks included in DiffBench were selected as an initial validation step rather than as an exhaustive survey of diffusion MRI's predictive scope. These tasks offer well-established ground truth labels, large reference cohorts, and a substantial body of prior work against which benchmark results can be calibrated [Chen et al., 2024; Dibaji et al., 2024; Chen et al., 2020a; Cardenas-Hernandez et al., 2025]. We acknowledge that sex and age prediction may be partially driven by macroscopic morphological differences rather than microstructural properties specific to diffusion imaging, and that ASD classification represents a challenging low-signal regime; the comparable performance of the non-diffusion $b_0$ baseline on certain tasks further underscores this point. Nevertheless, establishing reproducible, cross-cohort baselines on these canonical tasks is a necessary precondition for extending the framework to prediction targets where diffusion microstructure is expected to provide unique discriminative value (such as clinically relevant outcomes, white matter pathology, or cognitive phenotypes). We identify this as a primary direction for future work. In addition, cross-site harmonisation issues in multi-centre datasets (e.g., ABIDE-II) are not explicitly addressed in the present study.

Overall, DiffBench is designed with an emphasis on modularity, reproducibility, and extensibility, enabling it to be adopted and extended by the community. This design philosophy ensures that new preprocessings pipelines, datasets and models can be integrated easily while maintaining consistent experimental standards. In the following section, we describe its implementation and the organisation of its components to support flexible and scalable experimentation.

## 2.4 DiffBench Software Architecture

The design of DiffBench was motivated by several practical challenges that slow down reproducible evaluation in diffusion MRI predictive modelling. First, although numerous studies report promising predictive performance, publicly available implementations are often incomplete or unavailable, making independent validation and

extension difficult. Second, methodological differences in preprocessing, feature extraction, and evaluation protocols frequently confound comparisons across datasets and studies. Third, the fast development of learning architectures creates a need for a benchmarking system that can adapt to methodological innovation without requiring substantial software redesign.

To address these challenges, DiffBench was developed as an open-source, modular, dataset-agnostic and extensible framework for diffusion MRI microstructure analysis and predictive modelling. The framework aims to standardize common components of the pipeline while preserving flexibility for alternative representations, datasets, and modelling approaches. In particular, preprocessing and feature extraction procedures are shared across datasets whenever possible, enabling more meaningful comparisons between cohorts and reducing implementation-related variability. By relying on publicly available datasets, the benchmark further provides a common reference point that allows researchers to evaluate new methods under comparable conditions, even when their primary data cannot be openly shared.

A central design objective was to facilitate systematic comparison across microstructural representations. Beyond commonly used diffusion metrics, the framework is designed to support extension to other representations such as MAP-MRI and other diffusion models, allowing their predictive value to be assessed with common baselines as DTI within a unified evaluation setting.

To support these goals, the architecture explicitly separates (i) microstructure computation and spatial projection, (ii) representation construction, and (iii) model training and evaluation. This separation allows the same extracted features to be reused across models and experimental settings, minimizes redundant preprocessing, and reduces confounding effects arising from implementation variability. The framework further integrates run identifiers, configuration fingerprints, and structured outputs to ensure that benchmark results remain traceable and reproducible.

The following subsections describe the software architecture and its main components in detail.

### 2.4.1 Data Preparation

The data preparation step is handled by `BasePreparationPipeline`, the abstract base class, which defines the convention for data validation, microstructure computation, and feature loading. This base class is subclassed by specific pipelines for array-based, volumetric, and mesh-based representations, each implementing the necessary logic for their respective data formats. The preparation pipeline is responsible for ensuring that raw diffusion inputs are correctly formatted, that microstructure features are computed according to specified diffusion parameters, and that these results are cached. This design enforces a standardized data flow, agnostic to the dataset, from raw inputs to benchmark-ready features, while allowing for representation-specific processing and quality control steps.
Figure 2.3 represents the workflow of this data preparation pipeline.

**Data Needs and Dataset Interfaces.** Data passing allows BIDS-compliant datasets, HCP-style folder organization, and multi-site BIDS collections. A dataset-configuration dataclass encapsulates all dataset-specific parameters, including paths, diffusion parameterization, surface-space targets, and tissue type. The `data_reading` parameter of this dataclass specifies the expected organization of raw diffusion inputs, and the pipeline ensure that required inputs are present and correctly formatted. For BIDS data, the pipeline constructs a cached BIDS layout per site, supporting scalable indexing and multi-center harmonization. Subject identity is consistently normalized (e.g., stripping `sub-` prefixes), and a demographics pipeline provides canonical column mapping, multi-site annotations, and optional binarization of categorical variables. This enables reproducible target definitions and stratification. A subject alignment step enforces intersection between available brain features and demographics, ensuring that training data are consistent across representations and data sources.

**File Validation.** The file validation is two-fold: (i) it ensures that all required diffusion inputs are present and correctly formatted, and (ii) it checks for the presence of precomputed microstructure features in the cache before performing expensive computations. If the microstructural features are cached, the pipeline loads them directly; this and the remaining steps of the data preparation pipeline are skipped. When microstructure features have not been yet computed, preprocessing begins by verifying required diffusion inputs (DWI, b-values, b-vectors, segmentation, and surface files). The framework then resamples anatomical parcellations into DWI voxel space to construct tissue masks and ventricular masks. Microstructure computation is explicitly parameterized by diffusion timing (e.g., $\Delta$, $\delta$, and optionally per-shell $\Delta$), enabling consistent handling of heterogeneous acquisition protocols. Quality control is embedded in the pipeline through file validation, mask erosion to reduce partial-volume contamination, and logging of missing or invalid inputs. For BIDS datasets, surface projection is performed in native space and resampled to a template (fsLR or fsaverage) using sphere-based nearest-neighbor mapping, ensuring compatibility with cross-subject template-space analyses.

**Microstructure Feature Extraction.** The framework implements a registry of diffusion scalar metrics with a shared interface. Current metrics include MD (DTI), MK (DKI), SH power, and b0 intensity maps. Each metric is computed in masked voxel space, with optional ventricular normalization to mitigate inter-subject intensity variability. The metric registry enables extension to new microstructure measures without altering downstream pipelines. After computation, metrics are saved as volumetric NIfTI derivatives and immediately projected to cortical surfaces or the TBSS skeleton depending on tissue type, ensuring a standardized representation for subsequent analyses.

**Spatial Registration and Anatomical Projection.** Spatial registration is ex-

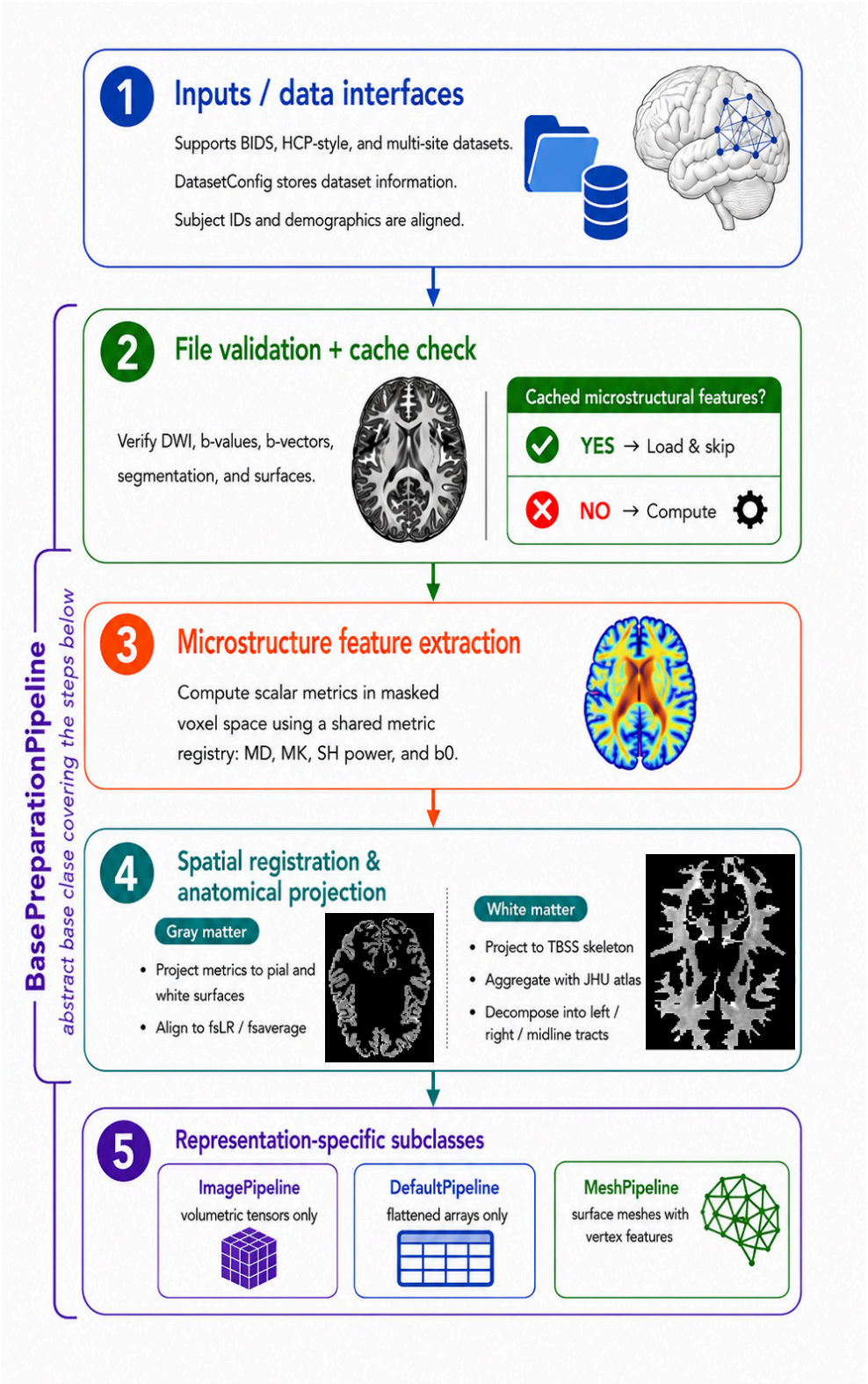



Figure 2.3: **Data preparation: from raw diffusion volumes to model-ready features.** An abstract base pipeline, `BasePreparationPipeline`, defines a common sequence of steps, validating the input data, computing microstructural metrics from the specified diffusion parameters, and projecting these metrics onto anatomically meaningful representations, while enforcing consistent spatial registration and quality control throughout. This shared base is then extended by representation-specific subclasses handling array-based, volumetric, and mesh-based data formats.

plicitly handled at the projection stage. For gray matter, volumetric metrics are sampled onto the pial and white cortical surfaces and aggregated across depths to reduce partial-volume effects. For white matter, the framework projects metrics onto the TBSS skeleton and performs tract-wise aggregation using the JHU ICBM-DTI-81 atlas, with hemisphere-aware decomposition into left, right, and midline tracts. Cortical parcellation relies on the Schaefer 2018 atlas, resampled onto fsLR or fsaverage surfaces via TemplateFlow with a local-first caching strategy, thus supporting reproducible template alignment even in offline environments. Parcel labels are maintained as globally unique identifiers across hemispheres, which is essential for parcel-level modeling and interpretability. It is possible to find this and the previous step in `utils_dmri_metric` file of the codebase.

**Data Representations.** Three primary representations are supported: (i) volumetric tensors (NIfTI paths loaded on demand), (ii) flattened feature vectors derived from parcel or tract aggregation, and (iii) cortical surface meshes with per-vertex microstructure features. These represenntations are derived from the `BasePreparationPipeline`. The corresponding child classes are therefore `ImagePipeline`, `DefaultPipeline`, and `MeshPipeline`. We need to note that, in order to compute the array-like representations, we need to first compute the images too so `DefaultPipeline` gets both outputs (volumes and array features) while `ImagePipeline` only gets the volumes. The mesh representation provides a structured graph with vertex features, parcel labels, and mesh connectivity serialized into Parquet files. This design decouples heavy geometric data from training loops while preserving compatibility with graph neural network frameworks. For image-based models, volumetric tensors are normalized and optionally resized; for classical ML, features are flattened into standardized arrays. This multi-representation design allows direct comparison between voxel-, surface-, and region-based approaches, while enabling memory-performance tradeoffs suited to diverse computational environments.

### 2.4.2 Model Interface and Learning Backends

Models are instantiated by pairing a backbone with a prediction head or, for classical ML models, creates a scikit-learn estimator with the corresponding prediction head already embedded. The models are configured in `model_configurations` where the `get_model` function passes all the model and backend parameters to the `create_trainer` to create the corresponding model for the experiment.

The prediction head abstraction supports both standard MLP heads and structured parcel-based heads with group-regularized penalties, facilitating interpretability and sparsity at the parcel level. Training backends implement a consistent API (`fit`, `predict`) across scikit-learn and PyTorch while preserving differences in optimization and logging. The PyTorch trainers handle target standardization for regression tasks, structured regularization, and augmentation-aware caching. This design allows new models to be integrated with minimal implementation effort while preserving consis-

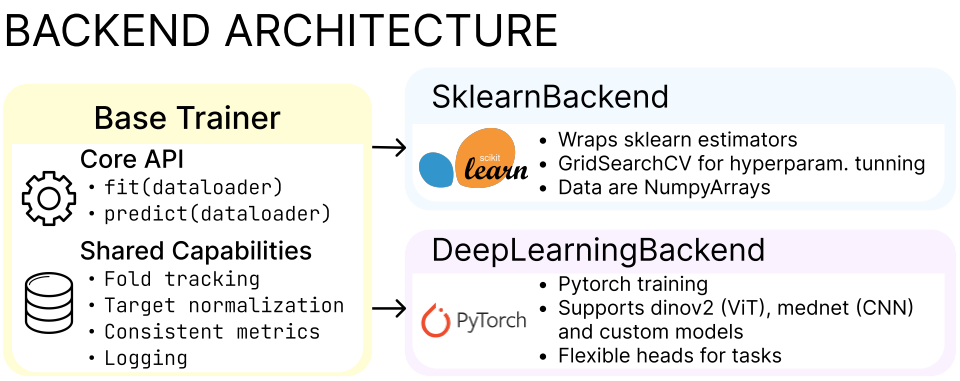



Figure 2.4: **Backend architecture: a unified interface for machine learning and deep learning models.** Each model is built by pairing a backbone with a prediction head that supports sparsity and interpretability at the level of individual brain parcels; for classical machine learning models, this pairing is wrapped directly into a scikit-learn estimator. All models are trained and evaluated through the same `fit`/`predict` interface, with framework-specific behaviour, such as target standardisation, structured regularisation, and augmentation-aware caching, handled internally by the corresponding training backend.

tent training and evaluation semantics.

Figure 2.4 illustrates the architecture and characteristics of the learning backends. What are the base capabilities of the backends, and how they are extended to support both types of trainings (ML and deep learning).

### 2.4.3 Experiment Infrastructure

To ensure reproducible and scalable experimentation, the framework provides dedicated infrastructure for experiment management, execution, and result tracking.
**Experiment Management and Reproducibility.** Reproducibility is enforced via configuration fingerprints and deterministic run identifiers derived from model, dataset, and target settings (`make_run_id` ensuring to create a readable identifier). Each experiment stores its full configuration and metadata (including host and job identifiers) under a structured output directory, alongside fold-level metrics and predictions in Parquet format. The framework supports caching of expensive features (e.g., frozen vision or mesh backbones) with explicit augmentation bookkeeping, ensuring that identical configurations reuse precomputed embeddings while documenting augmentation counts and input normalization conventions. Logging utilities record training curves and evaluation metrics, enabling both debugging and auditability. Cross-validation splits are stratified and seeded, and optional training-set subsampling is deterministic, ensuring that learning-curve experiments are reproducible.

**Scalability and Extensibility.** The pipeline is designed to scale across local exe-

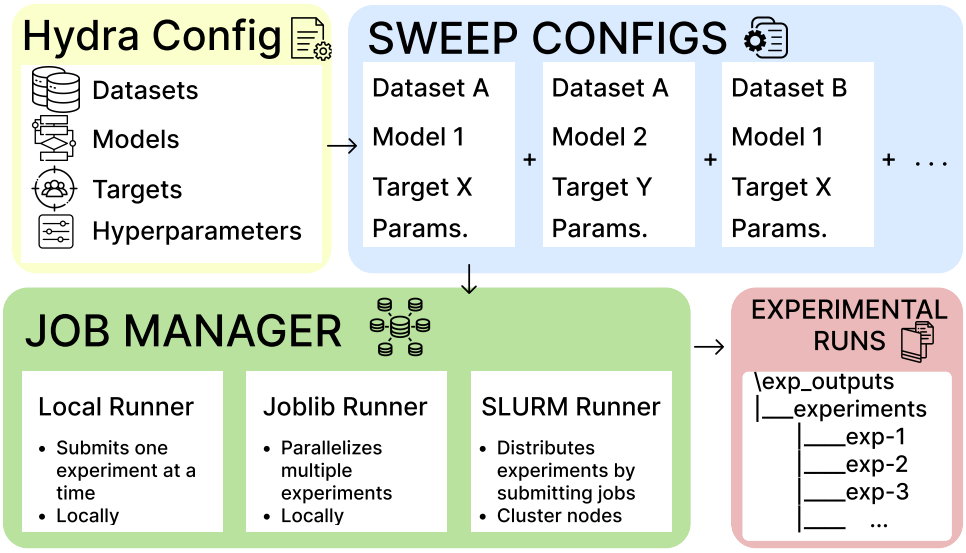



Figure 2.5: **Scalability and job management across execution environments.** The pipeline scales from local execution to joblib parallelism and SLURM-based clusters through a unified job manager that sweeps parameter configurations and caches completed runs; new datasets, metrics, representations, and models can be added through the same system without changing its core logic.

cution, joblib parallelism, and SLURM-based distributed runs through a unified job manager. New datasets can be integrated by specifying dataset configuration entries (paths, diffusion metadata, and preprocessing descriptors) without altering the core pipeline logic. New microstructure metrics can be added by registering additional metric computation functions, and new representation pathways can be introduced by subclassing the abstract preparation pipeline. The model interface is backend-agnostic, so novel classical, deep, or graph-based models can be plugged in with minimal interface changes. This extensibility supports continuous expansion of the benchmark scope without compromising reproducibility or architectural clarity.

In Figure 2.5, we illustrate the scalability of the pipeline and how the job manager can use one of the different execution modes. When the configurations of the experiments are defined, a sweep across the different parameters is performed and we create a set of runs for each of the possible combinations of the parameters. Then, for each of these runs, we check if the results are already cached and if not, we submit the job to be executed in one of the execution modes and store the results.

### 2.4.4 Design and Benchmark Philosophy

Multiple representations are used in the framework to account for the diverse spatial scales and anatomical contexts of diffusion microstructure. Surface-based representations provide anatomically faithful mapping of cortical features, skeleton-based

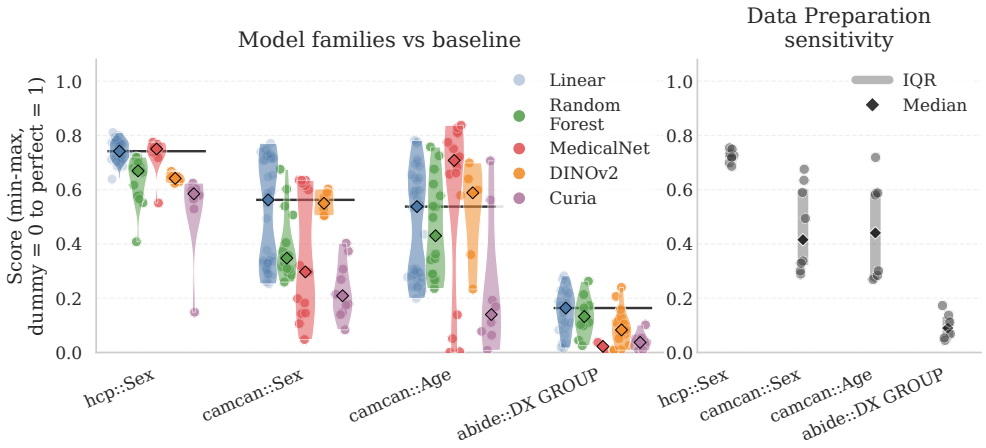



Figure 2.6: **Model performance depends more on representation than model family.** *(Left)* Model-family performance normalized across metrics and folds (dummy = 0, perfect = 1) reveals no consistent advantage of any model class. *(Right)* Preprocessing sensitivity across (feature, tissue) configurations demonstrates that performance strongly depends on representation and tissue selection.

projections reduce white-matter complexity while preserving tract specificity, and volumetric representations preserve spatial continuity for CNNs. The modular architecture ensures that each step (preprocessing, representation learning, and evaluation) is conceptually isolated, enabling fair benchmarking across heterogeneous datasets and modeling paradigms. By standardizing microstructure computation, spatial registration, and evaluation metrics, the framework mitigates methodological confounds and promotes scientific validity. At the same time, its extensible design allows researchers to explore novel representations, harmonization strategies, and predictive models while maintaining a transparent and reproducible benchmark foundation.

## 2.5 Results

We presented DiffBench, a standardised benchmarking framework for evaluating diffusion MRI microstructural representations across datasets, tasks, tissue types, and model families. Using DiffBench, we assessed the predictive performance stability, tissue-specific contributions, and the added value of diffusion-derived microstructural features beyond non-diffusion baselines.

Figure 2.6 highlights two central findings. First we evaluated the relative performance of different model families across datasets and configurations. To enable direct comparison across tasks, performance metrics were rescaled using min–max normalization relative to a dummy baseline that predicts the training-set mean (for regression) or majority class (for classification). DiffBench provides a set of reference results, for different model families, that can serve as a reliable baseline for future dMRI prediction studies.

The deep learning models considered here are not fully fine-tuned for each task and therefore do not reflect the full potential of deep learning in dMRI; it highlights the value of diffusion-specific pretraining and task-adapted fine-tuning as directions for future work. In the context of dMRI, limited sample sizes combined with high-dimensional signals further constrain the effective use of deep architectures – e.g. ABIDE II includes only 218 samples. Indeed, Curia and DINOv2 were employed as frozen feature extractors without fine-tuning. This transfer-learning regime typically requires larger cohorts to reach competitive performance, and the learned representations may not capture the microstructural properties most relevant to dMRI-based prediction. These results therefore do not reflect the ceiling of deep learning for dMRI, but rather highlight the potential value of diffusion-specific pretraining and task-adapted fine-tuning as directions for future work.

Direct numerical comparison with prior work is complicated by differences in preprocessing pipelines, feature extraction strategies, evaluation protocols, and the absence of publicly available code in most studies. We attempted to reproduce results reported in recent literature for sex and ASD classification. For sex classification, Chen et al. [2024] report accuracies of 86–88% using a 3D-CNN with MedicalNet features. Although no official implementation is available, we obtain comparable performance, with MedicalNet achieving a mean accuracy of 87% in our experiments. Using the Cam-CAN dataset, Dibaji et al. [2024] report an accuracy of 92% with a 3D-CNN; in our setting, MedicalNet is the closest comparable architecture, achieving a best accuracy of 80.50%. While this performance is lower, the discrepancy is expected given that we do not use the same model configuration or input modality, as their method is based on T1-weighted images. In contrast, for ASD classification, ElNakieb et al. [2021] report a mean accuracy of 99.9% with a linear SVM 5-fold cross validationusing a feature representation of DTI metrics. None of our presented results come close to this performance. However, they also evaluate results for RF and RBF-SVM models with mean accuracies of 59.1% and 57.3%; our closest comparable configurations for RF and RBF-SVM achieve mean accuracies around 56.5%, comparable to their results. Even thought they provide the hyperparameters used for their models, we were not able to find the exact code to reproduce their results. A more recent study of Cardenas-Hernandez et al. [2025] reports a peak accuracy of 97.37% using a Random Forest model; however, we were unable to replicate these results under our evaluation framework, with the best-performing configuration reaching 61.50% accuracy across folds.

In both cases, harmonization was not considered and the differences in preprocessing and evaluation pipelines makes it dificult to assess the sources of discrepancy. Furthermore, the lack of publicly available code makes it difficult to reproduce their results and to directly compare them with DiffBench. Table 2.1 summarises the available results from prior studies including dataset, prediction target, microstructural representation, and code availability.

Second, preprocessing and microstructural choices introduce substantial variability. In several tasks, differences across feature and tissue configurations exceed those

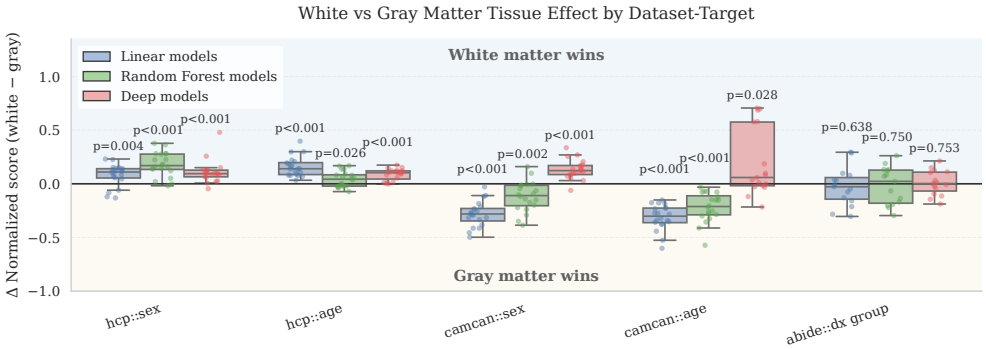



Figure 2.7: **The predictive value of white and gray matter is dataset-dependent.** Fold-level paired performance differences ($\Delta = \text{white} - \text{gray}$) were computed for identical model configurations after removing baseline differences between microstructural representations. Statistical significance was assessed using paired bootstrap tests with Benjamini–Hochberg FDR correction ($\alpha = 0.05$). White matter consistently outperforms gray matter in HCP, gray matter is favored in CamCAN. The optimal tissue compartment depends on the dataset and prediction task.

between model families. This pattern is also seen by Chen et al. [2024], who evaluated three deep learning architectures across FA, MD, and MK on HCP sex classification and found that performance differences across microstructural metrics within the same model were comparable in magnitude to differences across models within the same metric. For instance, the best-performing ViT variant achieved 0.87–0.90 across metrics, while the same architecture trained from scratch dropped to 0.66–0.69 regardless of the input metric, suggesting that training regime and feature choice interact as strongly as architecture selection. This indicates that performance cannot be attributed to model choice alone. Preprocessing is a major source of variance, and conclusions from a single configuration may be unreliable, emphasizing the need for robustness analyses.

**Impact of tissue type.**

In Figure 2.7, we quantify the impact of tissue type by comparing the performance of model families with white and gray matter for each (dataset, task) configuration. As before, performance is min-max normalized relative to a dummy baseline and we compute $\Delta = \text{white} - \text{gray}$. We remove microstructural representation baseline effects to isolate tissue differences within each (dataset, task) configuration. Statistical inference is performed on fold-level paired differences for each configuration. Significance is assessed using a nonparametric paired bootstrap test to assess whether the mean tissue effect differs from zero. Resulting p-values are corrected for multiple comparisons using the Benjamini–Hochberg false discovery rate (FDR) procedure across all dataset, (task, model family) comparisons ($\alpha = 0.05$).

Figure 2.7 shows that tissue effects are strongly dataset-dependent. In HCP, positive effect distributions indicate consistent white-matter superiority across con-

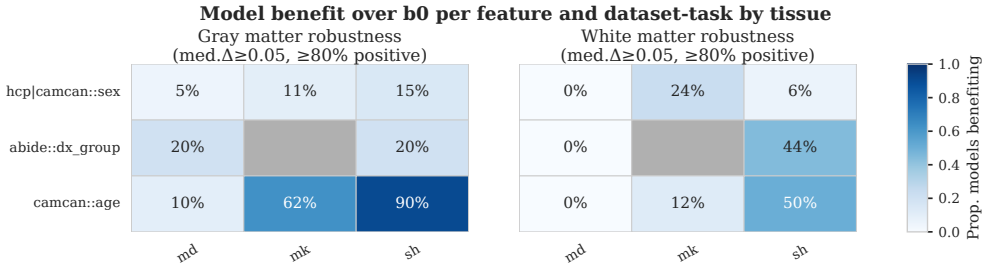



Figure 2.8: **Diffusion microstructure provides task-specific benefits beyond the non-diffusion baseline.** Proportion of model configurations outperforming the $b_0$ baseline, by task and microstructural representation; diffusion-derived features help substantially for tasks tied to microstructure (age, ASD) but offer little advantage for sex classification, showing that dMRI's value depends on the target's biological relevance rather than being universal. Given the large number of configurations and multiple comparisons, we adopt a robustness-based criterion (median $\Delta \geq 0.05$ and $\geq$80% positive folds) rather than formal hypothesis testing.

figurations. In contrast, CamCAN exhibits the opposite pattern, with gray matter showing a stable advantage. For ABIDE, the distribution of differences is centered around zero, suggesting no clear advantage of either tissue type.

### Impact of diffusion microstructure features

To assess whether diffusion-derived microstructural features provide predictive value beyond non-diffusion signal, we quantified model-wise improvements relative to b0. Given the large number of configurations (datasets, tasks, model families, and features), formal hypothesis testing would require extensive multiple-comparison correction and provide limited interpretability. Instead, we adopt a robustness-based criterion, defining consistent improvement as a median fold-level gain $\Delta \geq 0.05$ with at least 80% of folds positive. Under this criterion, microstructural information does not uniformly enhance performance across tasks or datasets. For tasks such as sex classification in HCP and CamCAN, diffusion microstructural features do not confer a systematic advantage over b0. It suggests that non-diffusion signal, likely reflecting global anatomical properties such as brain size, already captures most of the discriminative information. In contrast, tasks more directly linked to microstructural variation show measurable gains from diffusion metrics. Age prediction in CamCAN and diagnostic classification (DX_GROUP) in ABIDE exhibit consistent improvements for several diffusion representations, particularly under linear models operating on array-based features.

Deep learning models demonstrate positive gains in certain sex classification settings, likely reflecting sensitivity to morphology encoded in image-based representations rather than diffusion-specific contrast. Overall, these findings indicate that the added value of dMRI depends strongly on the biological relevance of the task, the dataset characteristics, and the modeling framework, rather than representing a

universal performance boost.

This finding is difficult to contextualise against the broader literature because, to our knowledge, no prior dMRI prediction study systematically compares diffusion-derived features against a matched non-diffusion baseline under the same experimental conditions. The closest reference point is Hwang et al. [2021], who predicted brain age from T2-weighted images, a contrast closely approximating the $b_0$ signal, achieving an MAE of 4.22 years on an internal test set and 9.96 years on an external dataset (IXI). While this demonstrates that meaningful age prediction is achievable from non-diffusion data alone, their results cannot be directly reproduced or compared, as neither the code nor the internal dataset is publicly available. More broadly, the absence of non-diffusion baselines in the dMRI prediction literature means that reported performance figures cannot be straightforwardly attributed to microstructural specificity. A model trained on $b_0$ alone may achieve comparable accuracy to one trained on FA or MD, yet the two findings carry very different biological interpretations. This limitation becomes especially consequential for prediction targets where microstructural specificity is the primary scientific question, such as cognitive phenotypes or clinical outcomes, where it is essential to establish not just whether a prediction is possible, but whether it requires diffusion-specific information at all. DiffBench directly addresses this gap by including $b_0$ as an explicit baseline under identical preprocessing, feature extraction, and evaluation conditions, enabling a principled assessment of the marginal contribution of diffusion microstructure across tasks, tissues, and model families.

## 2.6 Conclusion

In this work, we addressed a central limitation in dMRI prediction: the lack of controlled, comparable evaluation across preprocessing choices, microstructural representations, tissue types, model families, and datasets. By introducing DiffBench, a unified and open benchmark spanning three cohorts and multiple prediction tasks, we separate configuration-driven variability from modelling effects and provide a reproducible foundation for progress in the field.

Our analysis reveals three consistent findings. First, under the evaluated configurations, no single model family consistently outperforms the others: regularised linear models remain competitive with deep learning architectures tested here across tasks and datasets. This does not reflect the ceiling of deep learning for dMRI, the foundation models were employed without diffusion-specific pretraining or fine-tuning, and sample sizes across all three cohorts fall below what deep architectures typically require to reach their representational potential. A fair evaluation of deep learning for dMRI remains an open question. Second, preprocessing and microstructural representation choices introduce variability that frequently exceeds the differences between model families, confirming that performance results from single-configuration studies are unlikely to generalise and should be interpreted with caution. Third, the added value of diffusion-derived features over the non-diffusion $b_0$ baseline is task-

Table 2.3: Performance comparison across datasets, tasks, and model families considered in this study. Metrics are reported as balanced accuracy for classification tasks and $R^2$ for regression.

| Dataset | Task | Microstructure | Tissue | Metric | Model | Score |
|---|---|---|---|---|---|---|
| HCP | Sex | MD, MK, SH, b0 | GM, WM | Balanced acc. | Linear | 0.875 |
| | | | | | RF | 0.850 |
| | | | | | MedNet | 0.880 |
| | | | | | Dino | 0.825 |
| | | | | | Curia | 0.795 |
| Cam-CAN | Sex | MD, MK, SH, b0 | GM, WM | Balanced acc. | Linear | 0.785 |
| | | | | | RF | 0.675 |
| | | | | | MedNet | 0.650 |
| | | | | | Dino | 0.780 |
| | | | | | Curia | 0.605 |
| Cam-CAN | Age | MD, MK, SH, b0 | GM, WM | $R^2$ | Linear | 0.510 |
| | | | | | RF | 0.415 |
| | | | | | MedNet | 0.710 |
| | | | | | Dino | 0.590 |
| | | | | | Curia | 0.120 |
| ABIDE II | ASD | MD, MK, b0 | GM, WM | Balanced acc. | Linear | 0.575 |
| | | | | | RF | 0.565 |
| | | | | | MedNet | 0.505 |
| | | | | | Dino | 0.545 |
| | | | | | Curia | 0.520 |

dependent rather than universal. Sex classification shows little systematic gain, while age prediction and ASD diagnosis exhibit more consistent improvements under specific configurations.

These findings are difficult to contextualise against prior work, and that difficulty is itself a finding. We attempted to reproduce the methodology of several studies cited in this benchmark [Chen et al., 2024; Cardenas-Hernandez et al., 2025; Hwang et al., 2021], but in each case encountered the same barrier: neither the code nor, in some instances, the processed data are publicly available. The performance gaps we observe (DiffBench obtains 0.75 normalised accuracy on HCP sex classification versus 0.83–0.91 reported by [Chen et al., 2024], and 0.15 on ABIDE II ASD classification versus 97.37% reported by [Cardenas-Hernandez et al., 2025]) cannot be reliably attributed to differences in model capacity, preprocessing, or feature design, because the exact configurations that produced those results cannot be examined or replicated. This reproducibility gap reflects a structural problem in the field where performance claims accumulate without the infrastructure needed to verify, compare, or build upon them. DiffBench is a direct response to this problem. By releasing the full pipeline, configurations, and evaluation protocol as open-source software, it provides a reference point that can be interrogated, extended, and contested.

Beyond the specific results presented here, DiffBench establishes that advancing dMRI-based prediction requires more than proposing new models or representations in

isolation. It requires a shared evaluation infrastructure that makes variability visible, baselines explicit, and comparisons reproducible. The inclusion of $b_0$ as a non-diffusion baseline is one illustration of this principle. Without it, the microstructural specificity of any reported result remains ambiguous, as demonstrated by the fact that no prior study in this area systematically reports such a comparison.

Future work will extend DiffBench to additional datasets, predictive tasks (including cognitive phenotypes and clinical outcomes where diffusion microstructure is expected to provide unique discriminative value) and microstructural representations beyond those evaluated here. Another beneficial approach is to incorporate harmonisation methods to better assess cross-dataset generalisation, and to enable region-level analyses that localise predictive signal within tissue compartments. We see DiffBench not as a final benchmark but as an evolving infrastructure. A shared reference point that makes it possible to ask, and begin to answer, which aspects of dMRI carry genuine predictive value, under what conditions, and for which targets.

# Chapter 3

# Stability Selection for Interpretable Diffusion MRI

## 3.1 Introduction

Chapter 2 analysed machine learning methods for predicting demographic, cognitive, and clinical variables from high-dimensional diffusion MRI data. It showed that a range of predictive models can achieve strong performance on tasks such as sex classification and age regression. However, predictive accuracy alone is insufficient if neuroscientists cannot understand which brain features drive these predictions [Tonekaboni et al., 2019; Rosenbacke et al., 2024; Thibeau-Sutre et al., 2023]. In neuroimaging, models are valued not only for their predictive performance but also for the extent to which their decisions can be understood, questioned, and trusted [Thibeau-Sutre et al., 2023]. Consequently, interpretability has become essential for translating predictive models into meaningful neuroscientific knowledge.

Among the different forms of interpretability, spatial interpretability is particularly important because it identifies the brain regions and networks that contribute to model predictions. By linking model decisions to anatomical structures, it provides a bridge between machine learning and neuroscience. Spatial interpretation is commonly based on the weights of fitted models, which quantify the contribution of individual features [Haufe et al., 2014]. However, interpreting these weights is challenging in high-dimensional neuroimaging data, where features are strongly correlated. This is particularly true in diffusion MRI, where the information relevant to prediction tends to be distributed across several anatomically related brain regions rather than concentrated in a single one. As a result, different subsets of these regions can carry largely overlapping information and achieve very similar predictive performance. Because of this redundancy, a model trained multiple times on the same data, for instance using different random splits or initialisations, may assign high importance to a different subset of regions each time, even though the overall predictive signal has not changed. This makes it difficult to distinguish which highlighted re-

gions reflect a genuine, biologically meaningful contribution to prediction, and which simply happened to be selected over an equally informative, correlated alternative [Panigrahi et al., 2025; Poldrack et al., 2020].

Furthermore, interpreting a single model fitted on a single feature representation does not guarantee that the identified brain regions capture robust neurobiological information [Botvinik-Nezer et al., 2020; Poldrack et al., 2020]. Predictive models may rely on representation-specific characteristics, dataset-specific biases, or other signals that do not generalize. In Chapter 2, we showed that several diffusion MRI embeddings and machine learning models achieved comparable predictive performance across multiple tasks, suggesting that diffusion-derived features contain robust predictive information. However, this leaves an important question unanswered: when different models reach similar strong predictive performance, do they consistently rely on the same brain regions to do so, or can comparable performance be achieved by relying on different regions each time? If the latter is the case, it would suggest that the regions identified as important are not being reliably or consistently selected, and that the predictive information extracted from diffusion MRI may not reflect a stable neurobiological signal.

To address this problem, we use stability selection [Meinshausen and Bühlmann, 2010] as an interpretability framework for sparse predictive models applied to diffusion MRI. Rather than estimating feature importance from a single fitted model, stability selection measures how consistently features are selected across repeated perturbations of the learning process. By aggregating results across multiple sparse estimators and experimental settings, it identifies brain regions that contribute robustly to prediction while reducing sensitivity to model-specific variability. This approach is particularly well suited to diffusion MRI because it explicitly accounts for the correlations inherent in brain data and shifts the focus from individual model solutions to consensus patterns that persist across many successful models.

The objective of this chapter is therefore not to develop a new predictive algorithm, but to determine whether the predictive information extracted from diffusion MRI exhibits a stable and biologically meaningful anatomical organization. Specifically, we address three complementary questions. First, we ask whether different embedding representations converge toward similar patterns of regional and subnetwork importance despite their distinct feature constructions. Second, we assess whether these patterns replicate across two independent cohorts, providing evidence of robustness beyond a single dataset. Third, we compare the resulting stability maps with existing neuroimaging literature to evaluate their biological plausibility. Together, these analyses provide a framework for moving beyond predictive performance toward robust interpretation, allowing us to assess whether the anatomical patterns identified by machine learning models reflect reproducible neurobiological organization rather than model-specific statistical artefacts.

## 3.2 Related Work

### 3.2.1 Region based analysis in neuroimaging

Region-based analysis is a widely used strategy for studying structural and functional variation in neuroimaging data [Kim and Ye, 2020; Shanmugan et al., 2022; Serio et al., 2024; Weis et al., 2020]. Rather than analysing individual voxels independently, these approaches aggregate measurements within anatomically meaningful brain regions, reducing dimensionality while facilitating biological interpretation. The specific methodology depends on the representation of the neuroimaging data and the level of anatomical detail required for the analysis [Snook et al., 2007].

Early region-based studies relied on manually defined regions of interest (ROIs). In diffusion tensor imaging (DTI), Mielke et al. [2009] developed standardised protocols for defining major white matter tracts, including the fornix, cingulum bundle, splenium, and cerebral peduncles. Using visible brain anatomy and diffusion images to define regions, the authors found that repeated manual delineations of the same tracts led to highly consistent measurements, showing that this approach can be reliable when performed carefully. However, while manual delineation allows precise anatomical targeting, it is labour-intensive, difficult to scale to large samples, and susceptible to operator-dependent variability.

To address these limitations, automated atlas-based approaches have been developed to provide consistent anatomical parcellations across subjects. Rather than manually defining regions for each subject, brains are registered to a common anatomical atlas and automatically segmented into predefined parcels. Faria et al. [2010], proposed an atlas-based framework that partitions the brain into 130 anatomical regions from which regional diffusion and volumetric measurements were extracted. By reducing the number of variables relative to voxel-wise analyses, atlas-based methods provide a more tractable representation of neuroimaging data while preserving anatomical specificity [Snook et al., 2007]. Atlas-based methods have been successfully applied in numerous clinical settings to investigate neurological disorders. Lin et al. [2015] employed an atlas-based framework to study vascular cognitive impairment without dementia (VCIND). By extracting diffusion measures from major white matter tracts, the authors identified widespread reductions in FA and increases in MD across multiple white matter pathways. These regional abnormalities were significantly correlated with cognitive performance, illustrating how region-based analysis can relate microstructural changes to outcomes of direct clinical relevance.

These clinical findings illustrate a broader methodological point: aggregating measurements within predefined anatomical regions, rather than evaluating at voxel-level, offers a particular advantage when relating brain structure to behaviour or disease. Voxel-wise approaches evaluate statistical relationships independently at each voxel. It provides fine spatial detail but, because so many statistical tests are performed at once, they require correction for multiple comparisons and often produce results that are challenging to interpret anatomically [Snook et al., 2007]. Region-based methods

address both of these limitations directly. By aggregating information within anatomically defined structures before testing, they reduce the number of comparisons required and obtain findings that are more directly linked to established neuroanatomy. For this reason, voxel-wise and region-based analyses are generally viewed as complementary, trading fine spatial detail for anatomical clarity [Snook et al., 2007].

Instead of serving only as units for statistical comparison, anatomical parcels can be used as predictive features in classification and regression models [Serio et al., 2024]. This development has created a growing interest in methods capable of identifying which regions contribute most strongly to model predictions. One common strategy is perturbation-based importance analysis. Chen et al. [2024] performed an occlusion analysis in which brain regions were systematically masked and the resulting degradation in predictive performance was measured. Regions whose removal produced the largest performance loss were interpreted as the most important for the prediction task.

Overall, region-based analysis represents a compromise between the high spatial detail of voxel-wise approaches and the interpretability required for clinical and neuroscientific applications. Building on this idea, the present chapter uses anatomically defined cortical regions to estimate the importance of different brain areas in diffusion MRI-based prediction tasks, and to assess whether the regions identified by predictive models are consistent with existing neurobiological knowledge.

### 3.2.2 Interpretability and explainability in ML for neurosciences

The need for interpretable and explainable models has grown substantially as machine learning methods have been adopted in healthcare and neuroscience [Tonekaboni et al., 2019; Rosenbacke et al., 2024]. While predictive performance remains an important objective, clinical applications often require more than accurate predictions [Thibeau-Sutre et al., 2023]. In many cases, neurologists, radiologists, and other healthcare professionals must understand the factors driving a model's decision before incorporating its output into their diagnostic or treatment planning processes. Consequently, a model that achieves high classification accuracy may still have limited practical utility if its predictions cannot be justified or linked to meaningful neurobiological mechanisms [Thibeau-Sutre et al., 2023].

Interpretability in machine learning is commonly divided into two broad categories: intrinsic interpretability (or transparency) and post-hoc explainability [Thibeau-Sutre et al., 2023; Arrieta et al., 2020; Burkart and Huber, 2021]. Transparent models are inherently understandable because the relationship between input features and predictions can be directly studied [Rudin, 2019]. We can consider as examples of transparent models linear models with sparsity constraints, such as Lasso, Elastic Net, and Group Lasso, where model coefficients can be interpreted as measures of feature importance. In contrast, post-hoc explainability methods are applied after model training to explain the behaviour of complex or opaque models, such as deep neural networks. These approaches attempt to provide insight into the model's predictions

without necessarily revealing the complete internal decision-making process.

The distinction between these two categories is particularly relevant in neuroimaging applications. Thibeau-Sutre et al. [2023] argue that evaluating machine learning models solely through predictive metrics may be insufficient in clinical settings. The authors further highlight that strong predictive performance on a given dataset does not necessarily guarantee robust generalization across independent cohorts, emphasizing the importance of understanding the learned representations and decision mechanisms. We could relate this to the findings of the previous chapter, where we showed the variability in performance across datasets and tissue types.

The growing popularity of deep learning has led to the development of numerous explainable artificial intelligence (XAI) methods. Qian et al. [2023] provide a comprehensive review of interpretability techniques for deep learning models in neuroimaging. Their review distinguishes between model-specific approaches, such as Class Activation Mapping (CAM) [Zhou et al., 2016], Gradient-weighted Class Activation Mapping (Grad-CAM) [Selvaraju et al., 2017], Layer-wise Relevance Propagation (LRP) [Bach et al., 2015], and attention mechanisms, and model-agnostic approaches, including SHAP [Lundberg and Lee, 2017], LIME [Ribeiro et al., 2016], and occlusion-based methods. These techniques typically generate saliency maps or feature attribution scores that identify image regions contributing to a prediction. While such methods offer valuable insights into deep learning models, they remain post-hoc explanations and may not always faithfully reflect the model's true reasoning process.

Among model-agnostic methods, SHAP (SHapley Additive exPlanations) has become one of the most widely adopted approaches for quantifying feature contributions [Lundberg and Lee, 2017; Arrieta et al., 2020]. SHAP assigns an importance value to each feature according to its marginal contribution to the model's prediction. In neuroimaging studies, SHAP has been used to identify anatomical regions, white matter tracts, and clinical variables associated with neurological disorders and clinical outcomes [Rosenbacke et al., 2024; Saglam et al., 2023; Lee et al., 2024]. Saglam et al. [2023] employed SHAP to interpret an XGBoost model that differentiated early-onset schizophrenia from early-onset bipolar disorder using diffusion tensor imaging features. Similarly, Lee et al. [2024] combined random forests and SHAP values to identify clinically relevant predictors of balance recovery after stroke, demonstrating the contribution of both clinical assessments and diffusion-based biomarkers.

Despite their popularity, feature attribution methods such as SHAP present important limitations [Arrieta et al., 2020; Chen et al., 2023b]. In particular, highly correlated features can lead to ambiguous attribution scores because the method must distribute predictive credit among variables that contain similar information. This issue is especially relevant in neuroimaging data, where neighbouring brain regions and connectivity measures often exhibit strong correlations [Haufe et al., 2014]. As a result, feature importance estimates may become unstable across datasets or model configurations, potentially complicating biological interpretation.

Beyond post-hoc explanations, several studies have explored intrinsically interpretable approaches for neuroimaging analysis. Ryali et al. [2012] applied stability

selection to estimate sparse partial correlation networks from functional MRI data. Stability selection [Meinshausen and Bühlmann, 2010] combines repeated subsampling with sparse model fitting to identify features that are consistently selected across multiple data perturbations, thereby improving robustness while controlling false discoveries. Such approaches provide a direct link between model parameters and neurobiological features, making them attractive for scientific interpretation.

The present work follows this transparency-oriented perspective. Rather than relying exclusively on post-hoc explanation techniques, region importance is derived directly from sparse linear models based on Elastic Net and Group Lasso regularization. These models promote the selection of informative brain regions while maintaining a clear relationship between model coefficients and predictive outcomes. Consequently, the resulting region importance estimates can be interpreted as part of the model itself, facilitating both neuroscientific interpretation and clinical trust.

Despite substantial progress in region-based neuroimaging analysis and machine learning interpretability, a key limitation of existing approaches is the reliance on single-model or single-configuration analyses for identifying relevant brain regions. In many studies, feature importance is derived from a single trained model or a single feature representation, making it difficult to assess whether the resulting region importance estimates reflect stable properties of the data or are influenced by specific analytical choices.

To address this limitation, the present work adopts a stability-based framework for estimating region importance in diffusion MRI. Instead of relying on a single model, region importance is derived by aggregating sparse feature selections across multiple preprocessing strategies, embedding representations, subject partitions, and sparse estimators.

The resulting stability maps are then evaluated in relation to existing neuroimaging literature to assess their biological plausibility. In the following sections, we describe the proposed stability selection framework and present the corresponding region-level results.

## 3.3 Methods

We now introduce the methodology used to estimate region importance from diffusion MRI data through stability-based sparse modelling. The proposed analysis builds upon Chapter 2, extending it from predictive performance evaluation to the identification of anatomically meaningful predictors.

The analysis consists of repeatedly fitting sparse predictive models under multiple perturbations of the data preparation pipeline. These perturbations include variations in the feature representations of the diffusion MRI data, subject partitions, and model estimators. For each fitted model, selected regions are cached and subsequently aggregated using stability selection [Meinshausen and Bühlmann, 2010] to estimate the consistency with which each region contributes to prediction.

The analysis combines the following sources of variability:

(i) multiple embedding representations of the diffusion signal, (ii) multiple subject partitions (train/test splits), (iii) multiple sparse estimators, and (iv) stability selection across subsamples.

The resulting stability scores quantify how frequently each region is selected across different modelling configurations, reflecting the robustness of its contribution to prediction under variations in the analytical pipeline. Regions consistently selected across perturbations are interpreted as robust contributors to prediction and are subsequently examined in relation to existing neuroimaging findings.

### 3.3.1 Data Representation

The analysis uses the MD diffusion MRI microstructural representations introduced in the DiffBench framework (Chapter 2). As for datasets, the analysis focuses on the sex classification tasks for the HCP and CamCAN datasets, which were selected for their large sample sizes and high-quality diffusion MRI data. We consider multiple embedding representations in order to group the vertex-level features into cortical parcels and therefore, capture complementary aspects of the diffusion signal. These embeddings, include representations based on the mean value of the microstructural features within each parcel, principal component analysis (PCA) of the vertex-level features, a quantile-based representation, and summary statistics (mean, standard deviation, maximum and minimum value of the region and skewness) of the features within each parcel. Note that, although several of these embeddings describe a region using more than one value, for instance several principal components, or several summary statistics, each region is nonetheless treated as a single unit when predictive models are fitted, as described in the following section. The analysis is restricted to the cerebral cortex, using the Schaefer atlas at a resolution of 100 parcels [Schaefer et al., 2018].

By evaluating region importance across several data representations rather than a single representation space, the analysis reduces the risk that identified regions reflect representation-specific characteristics. Regions that remain important across multiple embedding representations are therefore more likely to correspond to stable biological signal.

Let $X \in \mathbb{R}^{n \times (p \times d_i)}$ denote the feature matrix composed of $n$ subjects, $p$ brain regions and representation of dimensiono $d_i$, and let $y \in \mathbb{R}^n$ represent the prediction target for a given task. For each dataset-task pair, several alternative representations of the same underlying biological data are constructed to probe the sensitivity of region selection to representation choices.

Specifically, let $\mathcal{E} = \{e_1, \ldots, e_E\}$ the set of representations, and $\mathcal{S} = \{s_1, \ldots, s_S\}$ the set of subject partitions. Each experimental configuration yields a transformed feature matrix $X^{(e,s)}$.

### 3.3.2 Sparse Predictive Models

Neuroimaging feature matrices are characterised by a high-dimensional regime in which the number of brain regions $p$ can substantially exceed the number of subjects $n$. In this setting, unregularised estimators are ill-posed and tend to overfit, producing dense coefficient vectors that assign non-zero weights to many regions simultaneously. Such dense solutions are difficult to interpret anatomically, it is not clear which regions drive the prediction and which are merely correlated bystanders.

Sparse estimators address this challenge by incorporating a regularisation term that constrains the complexity of the coefficient vector while fitting the predictive model. For a linear model, the prediction is expressed as $\hat{y} = X\beta$, where $X \in \mathbb{R}^{n \times p}$ is the feature matrix, $y$ is the target variable, and $\beta \in \mathbb{R}^p$ is the vector of model coefficients. A region may be described by several values given an embedding representation (for example, several principal components or several summary statistics); in this case, these values are combined by the sparse estimator into a single coefficient per region, so that $\beta$ always has exactly $p$ entries, one for each anatomical region, regardless of how that region was originally represented. This coefficient vector is obtained by solving an optimisation problem that balances predictive accuracy against model complexity:

$$\hat{\beta} = \arg\min_{\beta} \left( \frac{1}{2n} \|y - X\beta\|_2^2 + \Omega(\beta) \right), \tag{3.1}$$

where the first term measures prediction error and $\Omega(\beta)$ is a regularisation function that imposes structural constraints on the coefficient vector. The estimated solution $\hat{\beta}$ therefore corresponds to the coefficient values that minimise this objective. Different choices of $\Omega(\beta)$ define different assumptions about how predictive information is distributed across features.

In this work, sparse predictive models are used to estimate regional importance across multiple experimental configurations. For each configuration, a sparse model is fitted and the resulting coefficient vector $\hat{\beta}^{(r)} \in \mathbb{R}^p$ is cached, where run $r = (e, s, a)$ indexes the combination of embedding $e$, subject partition $s$, and sparse estimator $a$. Two complementary estimators are considered, each imposing a different form of structured sparsity.

*Group Lasso.* The Group Lasso introduces sparsity at the level of predefined groups rather than individual features. In this work, features are also grouped according to Schaefer atlas parcels, such that all features belonging to the same parcel are considered as one. The penalty is defined as

$$\Omega_{\mathrm{GL}}(\beta) = \alpha \sum_{g=1}^{G} \|\beta_g\|_2, \tag{3.2}$$

where $G = 100$ corresponds to the number of Schaefer parcels and $\beta_g$ denotes the coefficients associated with parcel $g$. This formulation represents a mixed $\ell_1/\ell_2$ penalty; the $\ell_2$ norm measures the combined contribution of features within each parcel, while the summation across parcels acts as an $\ell_1$ penalty on parcel-level contributions.

Consequently, entire parcels can be driven to zero, while all features within selected parcels are retained. This produces solutions that are directly interpretable in terms of anatomical regions.

*Sparse Group Elastic Net.* The Sparse Group Elastic Net combines group-level and feature-level sparsity by incorporating both an $\ell_1$ penalty and a Group Lasso penalty:

$$\Omega_{\text{SGEN}}(\beta) = \alpha\rho\|\beta\|_1 + \alpha(1-\rho)\sum_{g=1}^{G}\|\beta_g\|_2. \tag{3.3}$$

Here, $\alpha > 0$ controls the overall regularisation strength and $\rho \in [0, 1]$ determines the relative contribution of feature-level and parcel-level sparsity. The $\ell_1$ term encourages individual coefficients to shrink to zero, allowing selective elimination of features within an active parcel, whereas the group penalty encourages entire parcels to be excluded. As a result, this estimator allows more flexible sparsity patterns than Group Lasso alone. Both $\alpha$ and $\rho$ are treated as hyperparameters and selected through cross-validation.

Together, these two estimators represent complementary assumptions about the organisation of predictive information in the brain. Group Lasso assumes that predictive signal is primarily structured at the parcel level, whereas Sparse Group Elastic Net allows additional sparsity within selected parcels. Using both estimators reduces dependence on a single regularisation strategy and enables the identification of regions that remain consistently associated with prediction across different sparsity assumptions.

The final region importance estimates are obtained by considering the ensemble of runs spanning all combinations of data representation, subject partition, and estimator family. Regions selected consistently across this ensemble are interpreted as regions whose predictive contribution is robust across analytical choices, rather than being specific to a single modelling configuration.

**Hyperparameter optimisation.** Regularisation hyperparameters are selected, as in Chapter 2, independently for each run $r$ via grid search cross-validation, implemented using *GridSearchCV* from `scikit-learn` [Pedregosa et al., 2011].

For each subject partition $s$, the training fold is further divided into $K = 5$ inner folds. The hyperparameter configuration achieving the best mean cross-validated performance across inner folds is selected, using mean absolute error for regression tasks and balanced accuracy for classification tasks. The model is then refitted on the complete training fold using the selected hyperparameters before the coefficient vector $\hat{\beta}^{(r)}$ is recorded.

Hyperparameter optimisation is performed independently for every analytical configuration, allowing the level of regularisation to adapt to the specific representation, estimator, and training sample under consideration. Hyperparameter selection is performed entirely within the training data, ensuring that the held-out observations remain unseen until final evaluation.




Figure 3.1: **Stability selection pipeline for region importance.** Sparse models are fitted across many combinations of embedding representation, subject partition, and estimator; low-performing runs are filtered out, region coefficients are thresholded into binary region-selection indicators, and selection frequencies are aggregated into a stability score per region.

**Limitations of single-run sparse estimation.** Although sparse estimators provide explicit feature selection, the regions selected by any individual model may depend on the particular analytical configuration being considered. In neuroimaging data, neighbouring brain regions often exhibit substantial correlation due to shared anatomical organisation and common microstructural properties. E.g., in diffusion MRI, features from spatially proximate parcels are often highly correlated due to shared white matter connectivity and proximity on the cortical surface. Under these conditions, multiple correlated regions may provide similar predictive information, making the selection process sensitive to the training sample, microstructural representation, and embedding.

Consequently, the regions identified by a single fitted model should not necessarily be interpreted as uniquely important. Rather, they represent one possible sparse solution among several alternatives that may achieve comparable predictive performance. This limitation is particularly relevant when one of the goals is biological interpretation and relevance, where conclusions based on a single model may be influenced by modelling choices rather than stable underlying signal.

To address this issue, the present work estimates region importance by aggregating selections across a large ensemble of sparse models fitted under multiple perturbations of the analysis pipeline. The resulting stability scores provide a measure of how consistently each region contributes to prediction across alternative analytical conditions.

### 3.3.3 Stability-Based Region Importance

The objective of the stability analysis [Meinshausen and Bühlmann, 2010] is to estimate how consistently each brain region contributes to prediction across the ensemble of sparse models described above. Figure 3.1 summarises this pipeline, from the generation of the model ensemble to the computation of the final per-region stability scores; each of these steps is detailed in the subsections below.

#### Ensemble Generation

For each dataset-task pair, sparse models are fitted across all combinations of microstructural configurations, embedding representations, and subject partitions. Each fitted model corresponds to a run $r = (e, s, a)$, where (e) denotes the embedding representation, (s) the subject partition, and (a) the sparse estimator. Each anatomical region is represented by a single aggregated feature prior to model fitting. As a result, each run produces one coefficient per region rather than per voxel, yielding a coefficient vector $\hat{\beta}^{(r)} \in \mathbb{R}^p$, where $p$ is the number of anatomical regions and each entry quantifies the contribution assigned by the model to a single, specific region. Because the ensemble spans a broad range of analytical configurations, it provides many largely independent perspectives on the same underlying prediction problem.

#### Performance-Based Filtering and Coefficient Thresholding

Models with poor predictive performance may produce unstable coefficient patterns that are less likely to reflect meaningful predictive signal. To reduce the influence of such models, runs whose test performance score $s_r$ falls below the $q$-th quantile ($q = 10$) of the score distribution for the given dataset-task pair are discarded. The retained ensemble is therefore defined as

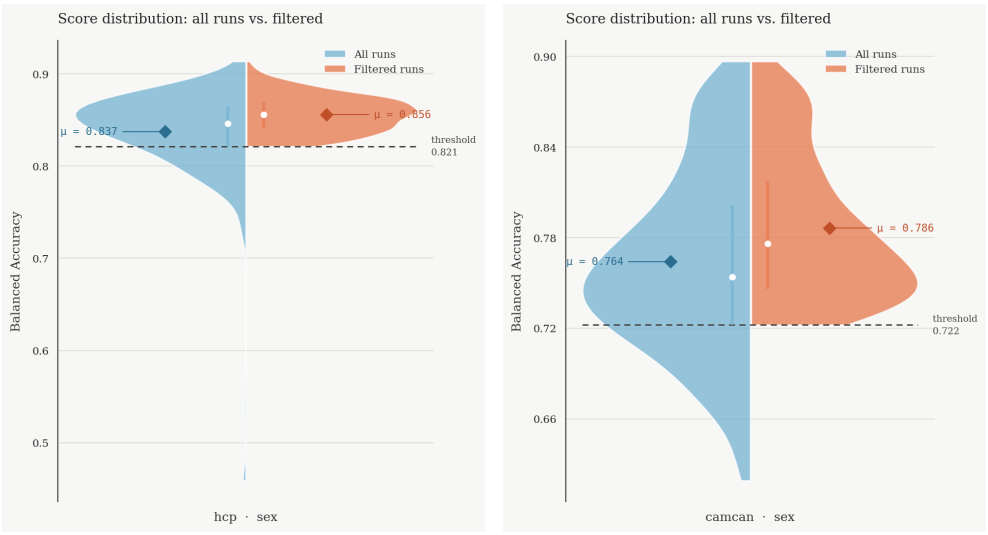
$$\mathcal{R}^* = \{r \in \mathcal{R} : s_r \geq Q_q(s)\}, \tag{3.4}$$

where $Q_q(s)$ denotes the $q$-th quantile of test scores across all runs. Only models belonging to ($\mathcal{R}^*$) and exhibiting sufficient predictive validity contribute to the final stability map.

To illustrate the effect of this step, figure 3.2 shows the distribution of balanced accuracy scores before and after filtering for the HCP and CamCAN sex-classification experiments. This filtering step allows to discard poorly performing runs in the region importance analysis.

#### Stability Estimation

The coefficient vectors produced by different estimators and feature representations are not directly comparable because their scales vary substantially across analytical configurations. To obtain a common selection criterion, each continuous coefficient vector is converted into a binary region-selection indicator.




(a) (b)

Figure 3.2: **Performance-based filtering removes unstable low-performing models.** Balanced accuracy distributions for HCP (a) and CamCAN (b) before (blue) and after (orange) excluding runs below the 10th performance percentile; this removes noisy, poorly fitted models before stability analysis, yielding more reliable coefficient estimates.

For each retained run ($r \in \mathcal{R}^*$), a run-specific threshold is defined as $90^{\text{th}}$ percentile of absolute coefficient magnitudes within that run,

$$T_r = Q_{0.9}\left(|\hat{\beta}^{(r)}|\right), \tag{3.5}$$

and each region $j$ is marked as selected or not selected according to whether its coefficient exceeds this threshold. Being the binary selection indicators assigned as,

$$Z_j^{(r)} = \mathbf{1}\left(|\hat{\beta}_j^{(r)}| \geq T_r\right), \quad Z_j^{(r)} \in \{0,1\}. \tag{3.6}$$

In other words, only regions belonging to the top decile of coefficient magnitudes within each run are therefore considered selected. The 90th percentile was chosen as a compromise between retaining enough regions to capture distributed predictive signal while maintaining sufficient sparsity to highlight the most influential regions within each model. Defining the threshold relative to each run separately, rather than using one fixed value across all runs, allows regions to be identified consistently even though absolute coefficient magnitudes vary considerably between different estimators and embedding representations.

Applying this procedure to every retained run produces a binary selection vector $Z^{(r)} \in \{0,1\}^p$, indicating which of the $p$ anatomical regions were selected in that particular run. Region-wise stability scores are then obtained by calculating how

frequently each region was selected across the filtered ensemble of runs,

$$\hat{\Pi}_j = \frac{1}{|\mathcal{R}^*|} \sum_{r \in \mathcal{R}^*} Z_j^{(r)}. \tag{3.7}$$

The resulting stability map, $\hat{\Pi} = (\hat{\Pi}_1, \ldots, \hat{\Pi}_p)$, assigns each anatomical region $j$ a score between 0 and 1, corresponding to the proportion of retained runs in which that region was selected. A region with a stability score close to 1 was therefore selected almost universally across the ensemble, while a score close to 0 indicates that it was rarely, if ever, identified as important.

This procedure is closely related to classical stability selection [Meinshausen and Bühlmann, 2010], where variable importance is estimated through repeated sparse estimation under perturbations of the data. In the present analysis, perturbations are obtained not only from variations in the training data but also from alternative embedding strategies, and sparse estimators. Regions with high stability scores therefore correspond to predictors whose contribution remains consistent across a broad range of analytical conditions.

### 3.3.4 Interpretation of Stability Maps

The final output of the analysis is a stability map assigning a stability score to each anatomical region. These scores do not quantify effect size or causal relevance. Instead, they measure the consistency with which a region contributes to prediction across the analysed ensemble.

The stability score $\hat{\Pi}_j$ can be interpreted as an empirical estimate of the frequency with which region $j$ is selected among models that achieve sufficient predictive performance. Formally,

$$\hat{\Pi}_j \approx P(j \text{ selected} \mid s_r \geq Q_q(s)), \tag{3.8}$$

under the perturbation distribution induced by the subject partitions, microstructural configurations, embedding representations, and sparse estimators considered in this study. Regions with high stability scores therefore correspond to predictive signals that remain consistently identifiable across heterogeneous analytical conditions, and represent strong candidates for robust and reproducible neurobiological biomarkers.

In the context of diffusion MRI, these stability maps should be interpreted as identifying cortical regions whose microstructural properties carry consistent discriminative information for the prediction target, regardless of the specific, embedding strategy, or sparse estimator used. Such regions are appropriate candidates for comparison with prior neuroimaging literature on microstructural correlates of target variables. Agreement between stability maps and previously reported anatomical correlates increases confidence that the identified regions reflect meaningful neurobiological signal rather than modelling artefacts.

Contrarily, regions with low stability scores should be interpreted with caution. Their selection may reflect characteristics of a particular embedding or preprocessing

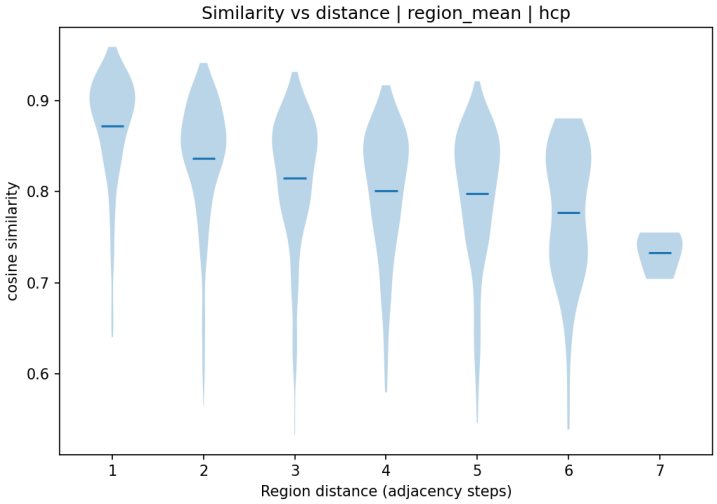



Figure 3.3: **Pairwise cosine similarity between cortical parcel representations as a function of spatial distance.** Cosine similarity between the mean diffusion-derived feature vectors of all pairs of the 100 cortical parcels, averaged across subjects. Plotted as a function of cortical region distance (measured in adjacency steps). It shows a high inter-parcel correlations across both short and long distances.

pipeline rather than a genuine biological signal. The stability map thus provides a principled criterion for distinguishing analytically robust findings from those that are contingent on methodological choices.

### 3.3.5 Limitations

Several limitations should be considered when interpreting the resulting stability maps.

First, stability does not imply causality. Regions with high selection frequencies are predictive of the target under a broad range of analytical conditions, but this does not establish that they are causally involved in the underlying biological process. Causal interpretation would require experimental or longitudinal evidence beyond the scope of the present framework.

Second, under high inter-regional correlation, common in diffusion MRI data due to shared white matter pathways and spatial proximity, predictive information may be distributed across several neighbouring parcels. In such cases, the group penalty may select the entire correlated neighbourhood rather than a single driving region, and selection probability may spread across anatomically adjacent structures. The stability map should therefore be interpreted at the level of cortical territories rather than individual parcels when inter-regional correlations are high.

Third, the coefficient thresholding strategy based on the $90^{\text{th}}$ percentile of absolute magnitudes within each run is heuristic. Although this approach normalises selection

across heterogeneous model classes and representations, alternative percentile thresholds would yield slightly different stability maps. Linked to this, stability estimates depend on the set of representations, preprocessing strategies, and model families included in the analysis. Different analytical choices may lead to different stability distributions.

Finally, performance-based filtering may introduce a subtle selection bias towards preprocessing strategies or embedding representations that are inherently easier to fit, rather than those that are most biologically informative. If certain analytical configurations consistently produce higher test scores, independently of biological signal, they will be over-represented in $\mathcal{R}^*$ and contribute disproportionately to the final stability map. This effect is partially mitigated by the diversity of configurations included in the ensemble, but cannot be fully eliminated by design.

## 3.4 Results

To interpret the predictive patterns learned by the sparse models, we analyse region importance through stability selection. Substantial correlations between cortical regions can make the selection of individual features unstable despite the presence of a consistent underlying signal. Therefore, multiple equivalent model solutions may achieve comparable predictive performance while differing in the specific regions retained. Consequently, the importance of a region is better characterized by its reproducibility across an ensemble of models than by its presence in any single fitted solution.

We first evaluated the correlation of the cortical regions. Figure 3.3 shows the relationship between pairwise parcel similarity and spatial distance. As expected, nearby regions tend to exhibit greater similarity, consistent with known spatial continuity in cortical microstructure. However, substantial correlations persist across anatomically distant regions, indicating that predictive information is distributed across the cortex rather than confined to localized clusters. This correlation structure motivates interpreting feature importance at the ensemble level. Rather than focusing on a single sparse solution, stability selection identifies regions that are consistently retained across repeated perturbations of the learning pipeline, including variations in training data, sparse estimators, and embedding representations.

Before interpreting stability maps, we filtered the runs to remove outliers that performed poorly on the test set (Figure 3.2). The resulting ensemble therefore represent independently successful predictive solutions from which stable feature-selection patterns can be reliably estimated.

For each retained run, we cached the set of cortical parcels selected by the corresponding sparse model. Across both datasets and the four microstructural representations considered in the DiffBench framework (Chapter 2), that is $b_0$, mean diffusivity (MD), mean kurtosis (MK), and spherical harmonics (SH), a total of 2,026 runs were retained after performance-based filtering: 1,075 for HCP and 951 for CamCAN, distributed across two sparse estimators (Group Lasso and Sparse Group Elastic Net) and

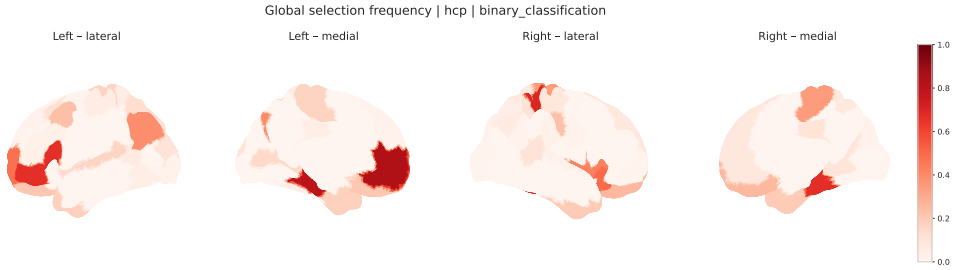



Figure 3.4: **Stable cortical regions for HCP sex classification.** Cortical parcels are coloured by their stability score, defined as the proportion of retained runs in which each parcel was selected by the sparse model. Regions with higher scores represent consistently predictive anatomical locations across embeddings, estimators, and data splits.

five embedding representations (mean and standard deviation, flattened vertex values, principal component analysis, percentile-based summaries, and summary statistics). The present analysis focuses specifically on mean diffusivity, having therefore 321 runs for HCP and 259 for CamCAN. Selection frequencies were then aggregated separately for each dataset across these mean-diffusivity-specific runs, including the five embedding representations and two sparse estimators described above. The stability score assigned to a parcel therefore reflects the proportion of successful runs in which that parcel was retained as informative. Parcels with high stability values are consistently selected across diverse analytical configurations, whereas parcels with low stability values are selected only sporadically.

The global stability map (Figure 3.4) visualizes these selection frequencies on the cortical surface. Each parcel is coloured according to its stability score, providing a representation of regional importance across the entire analysis framework. Regions with higher values correspond to cortical areas that repeatedly contribute to prediction regardless of the specific embedding representation, sparse estimator, or train-test partition employed.

**Embeddings are coherent at subnetwork level.**

Although multiple embeddings and models achieved comparable predictive performance (Chapter 2), equivalent performance alone does not imply that they exploit the same underlying biological signal. Different embeddings could, technically, achieve similar accuracies while relying on different subsets of features, raising the possibility that predictive performance is driven by none informative features in microstructure.

To address this possibility, we examined whether different embeddings converged on similar stability-selection patterns. Figure 3.7 summarizes the selection frequencies of the most stable regions and subnetworks across the different embeddings. To define the region aggregation into subnetworks, we rely on the Schaefer 100-parcel atlas, which assigns each parcel to one of the seventeen subnetworks (e.g. Default_A, Default_B, DorsAttn_A). For each run, a subnetwork was considered selected if at least one of its parcels was selected by the model. Subnetwork stability was then

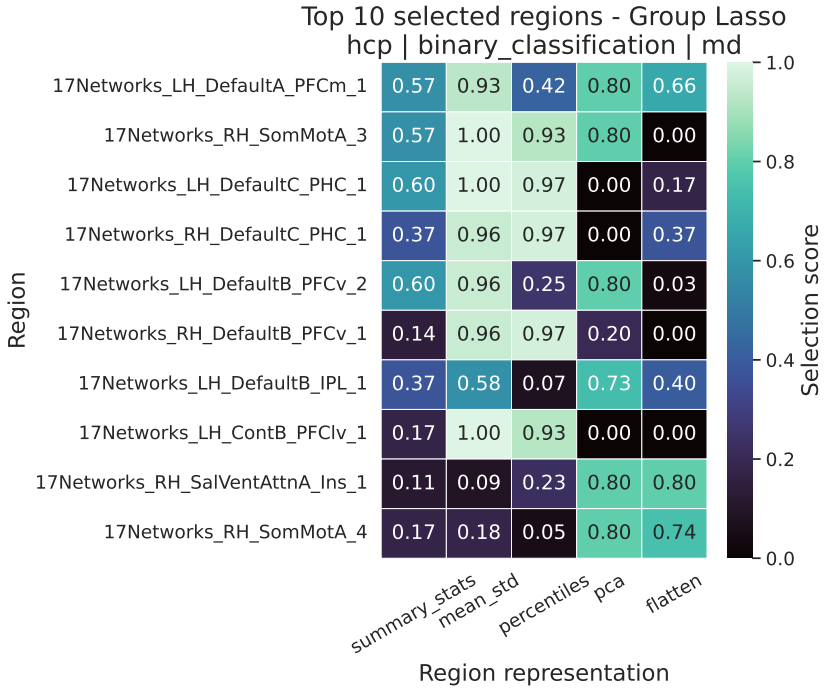



Figure 3.5:

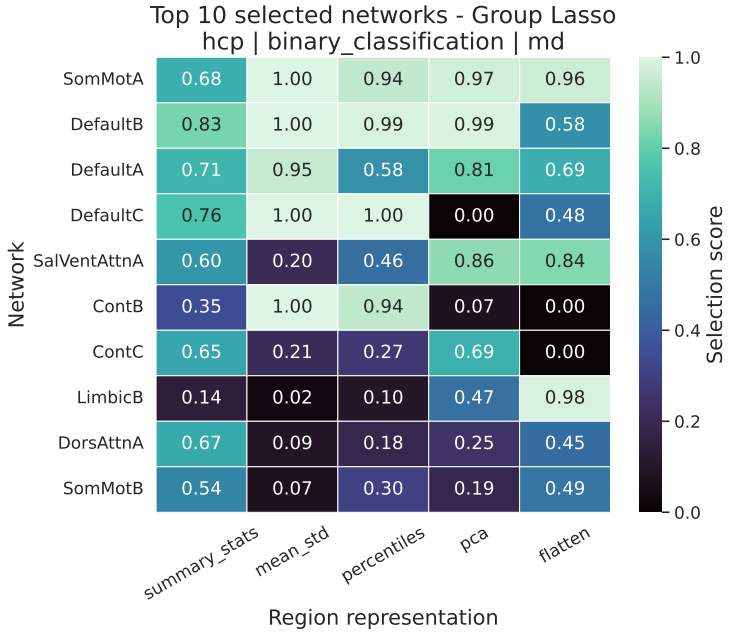



Figure 3.6:

Figure 3.7: **Embedding representations converge on stable predictive subnetworks.** Heatmaps show stability scores of the 10 most consistently selected cortical parcels (a) and functional subnetworks (b) across embedding representations for the HCP sex-classification task with group lasso. Although individual parcels vary between embeddings, similar subnetwork-level patterns indicate that different representations capture common predictive structures.

defined as the proportion of runs in which this condition was satisfied.

Overall, agreement across embeddings is substantially greater at the subnetwork level than at the level of individual regions. Part of this increase is expected simply because grouping regions into a much smaller number of subnetworks makes it easier for two embeddings to agree by chance, since there are fewer possible categories to disagree on. However, the level of agreement observed at the subnetwork level exceeds 0.5 for nearly all pairs of embeddings, corresponding to a level of agreement well above what would be expected by chance alone, and indicating that this convergence reflects genuine agreement between embeddings on which subnetworks are most informative, rather than a consequence of using fewer categories [Meinshausen and Bühlmann, 2010]. At the regional level, the `flatten` and `summary_stats` representations show lower consistency with the remaining embeddings, whereas the other representations display similar selection patterns. Once regions are aggregated into subnetworks, this agreement increases across all embeddings. The five highest-ranked subnetworks are consistently assigned high selection frequencies, while the five lowest-ranked subnetworks receive uniformly low scores across representations. The only notable exception is the `summary_stats` embedding, which assigns comparatively high selection frequencies to the lower five ranked subnetworks. Nevertheless, the overall ordering of the most and least informative subnetworks remains highly consistent across embeddings.

To quantify this convergence more systematically, we summarised the stability scores obtained for each embedding representation as a single vector, referred to here as its importance profile, with one entry per anatomical region (or, in a second analysis, per functional subnetwork). We then computed the pairwise cosine similarity between these importance profiles for every pair of embedding representations, separately for each dataset. Cosine similarity ranges from 0, indicating that two embeddings assign importance to entirely different regions, to 1, indicating that they assign importance to the same regions in very similar proportions; it therefore provides a direct measure of how much two embeddings agree on which regions are most important, independent of the overall scale of their stability scores. The resulting similarity values are displayed in Figure 3.8 as a grid comparing every pair of embeddings, for both datasets, with light blue colour indicating higher agreement between the two embeddings being compared. At the level of individual regions, these similarities reproduced the pattern already observed from the raw selection frequencies: the `flatten` representation showed comparatively low agreement with the other embeddings, both in HCP and in CamCAN. However, when importance was instead aggregated at the level of functional subnetworks, agreement between embeddings increased substantially across the board, indicating that although embeddings can disagree on the precise regions they consider most important, they converge much more closely once these regions are grouped into broader functional systems.

This finding suggests that different embeddings capture a common large-scale organizational signal even when they disagree on the precise contribution of individual parcels. Importantly, if embeddings were driven by distinct predictive information, one would expect low agreement across representations and a lack of reproducible

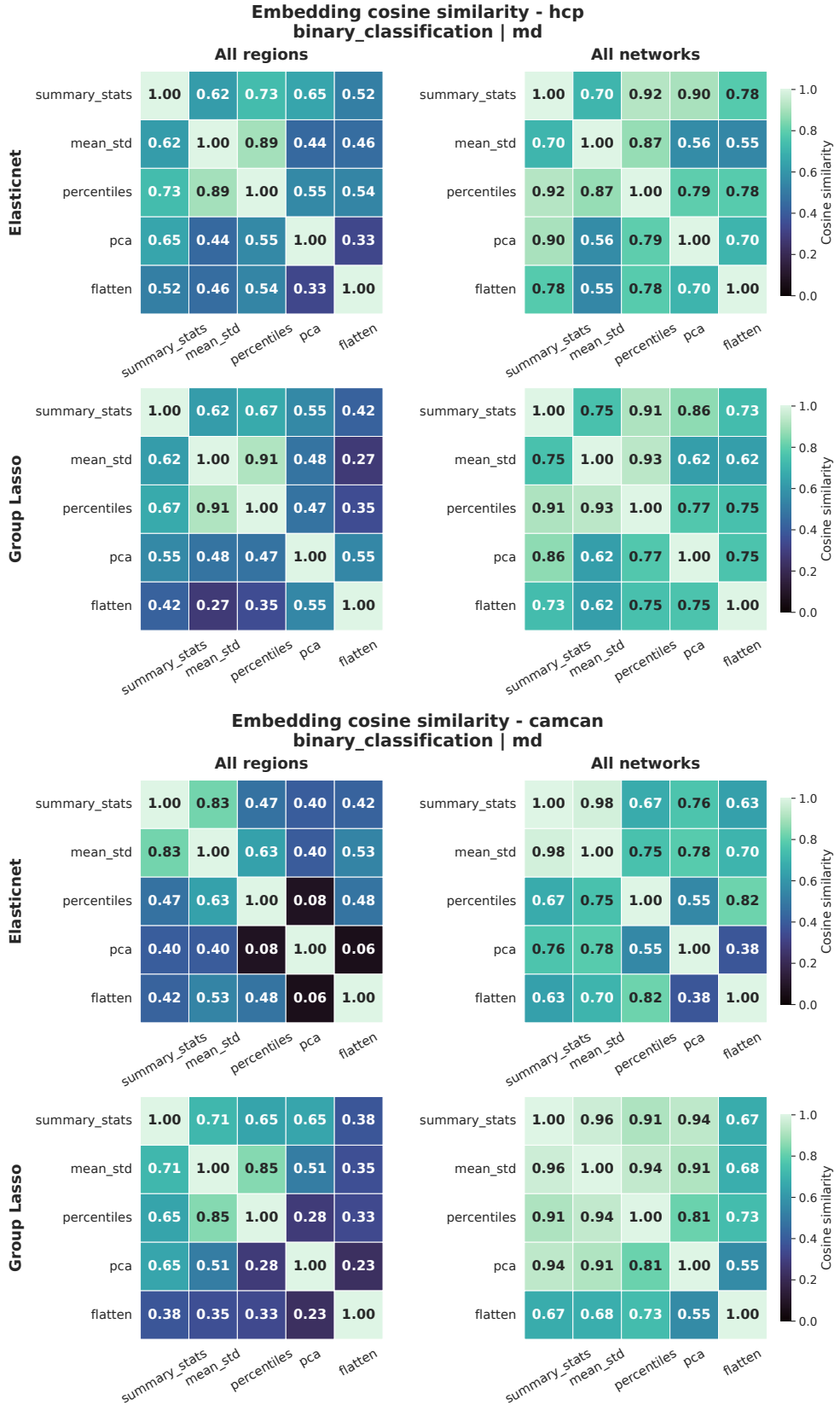



Figure 3.8: **Agreement between embedding-specific importance profiles.** Cosine similarity between stability-based importance profiles across embedding representations for HCP (top) and CamCAN (bottom) sex classification. Agreement is higher at the subnetwork level than at the individual-region level, showing that embeddings converge on similar large-scale predictive organization despite differences in regional selection.

anatomical organization. Instead, the observed convergence indicates that stability selection is recovering similar structural features regardless of the feature representation employed. This observation is consistent with the region correlation shown in Figure 3.3, where substantial similarity is observed across cortical parcels. Such regional correlations may lead different embeddings to select alternative but highly related parcels, thereby reducing agreement at the parcel level while preserving consistency at the subnetwork level.

Table 3.1: Global network selection ranking concordance between HCP and CamCAN (averaged across all models and embedding representations)

| Network | HCP score | HCP rank | CamCAN score | CamCAN rank | Δ rank |
|---|---|---|---|---|---|
| SomMotA | 0.8674 | **1** | 0.5461 | **2** | -1 |
| DefaultB | 0.8501 | **2** | 0.5441 | **3** | -1 |
| DefaultA | 0.7870 | **3** | 0.2460 | **13** | -10 |
| DefaultC | 0.6905 | **4** | 0.0952 | **17** | -13 |
| SalVentAttnA | 0.6291 | **5** | 0.4132 | **6** | -1 |
| ContB | 0.5659 | **6** | 0.3478 | **8** | -2 |
| ContC | 0.4639 | **7** | 0.3494 | **7** | +0 |
| DorsAttnB | 0.4156 | **8** | 0.2839 | **10** | -2 |
| SomMotB | 0.3902 | **9** | 0.4509 | **5** | +4 |
| LimbicB | 0.3864 | **10** | 0.4900 | **4** | +6 |
| DorsAttnA | 0.3765 | **11** | 0.2797 | **11** | +0 |
| VisCent | 0.3433 | **12** | 0.7051 | **1** | +11 |
| ContA | 0.3145 | **13** | 0.2861 | **9** | +4 |
| VisPeri | 0.3092 | **14** | 0.1868 | **15** | -1 |
| TempPar | 0.3091 | **15** | 0.2791 | **12** | +3 |
| LimbicA | 0.2617 | **16** | 0.2211 | **14** | +2 |
| SalVentAttnB | 0.1652 | **17** | 0.1732 | **16** | +1 |

**Independent cohorts show rank-level consistency.** Given the stronger convergence observed at the subnetwork level, we next evaluated whether the identified predictive information generalized across independent cohorts. For each dataset, subnetworks were ranked according to their global stability scores, and rank correspondence was assessed between HCP and CamCAN for the sex classification task. To compute this rank we first need to compute the consensus score. We recorded, for each of the five analysed embeddings, the selection proportion of every region or network for each unique combination of dataset and model. The consensus score for each region or network for a given dataset-model combination was then defined as the mean of its selection proportions across those five embeddings. For each subnetwork, its consensus score was obtained by averaging the parcel-level consensus scores across all parcels belonging to it, and separately across all models, obtaining one score per (dataset, subnetwork) pair. Subnetworks were then ranked within each dataset in descending order of this score, with rank 1 assigned to the most consistently selected subnetwork.

Despite the datasets differences, the overall ordering of subnetworks was largely

preserved across cohorts (Table 3.1). To quantify this, we computed, for each subnetwork, the absolute difference between its rank in CamCAN and its rank in HCP, referred to here as the $\Delta$ rank. The median $\Delta$ rank across subnetworks was only 2 positions, indicating that most subnetworks occupied a very similar position in both cohorts. The corresponding mean $\Delta$ rank was noticeably higher, at 3.64 positions; because the mean is sensitive to such outliers while the median is not, the median offers a more representative summary of how consistent subnetwork ranking was overall [Leys et al., 2013]. These two measures indicate that although a few subnetworks changed position substantially between cohorts, the ranking of most subnetworks remained highly consistent overall.

Most networks differed by fewer than three ranking positions, indicating that the relative importance of the major functional systems remains stable across datasets. Nevertheless, replication was not uniform across all subnetworks. A small number of subnetworks (VisCent, DefaultC, and DefaultA) displayed pronounced rank shifts of 10–13 positions. These deviations suggest that certain networks are more sensitive to cohort-specific factors, potentially reflecting differences in imaging characteristics, or population heterogeneity.

Taken together, the results suggest a meaningful level of cross-cohort reproducibility. The overall organization of predictive subnetwork importance is largely conserved, while a limited number of networks exhibit substantial dataset-dependent reweighting. This suggests that the main predictive signal is robust across cohorts, although some networks remain sensitive to differences in cohort-specific sources of variability.

**Stable predictive subnetworks align with established functional brain literature.** To further evaluate the biological plausibility of the identified predictive systems, we compared our findings with the existing literature on sex-related brain differences [Kim and Ye, 2020; Liu et al., 2026; Shanmugan et al., 2022; Serio et al., 2024; Weis et al., 2020]. Because most previous studies have reported results at the level of large-scale functional networks rather than individual cortical parcels, and because comparable region-level analyses remain scarce in diffusion MRI, we performed this comparison using the Schaefer functional-network framework. This choice was intentional. Although our models were trained on diffusion-derived microstructural representations, mapping the results onto a widely used functional atlas enables direct comparison with a broader body of neuroimaging evidence, particularly from functional MRI studies. For the network aggregation, the seventeen subnetworks were mapped to the seven canonical Yeo networks (Visual, SomatoMotor, Dorsal Attention, Ventral Attention/Salience, Limbic, Frontoparietal/Control, Default) by matching each subnetwork label to its network prefix.

To summarize evidence across studies, we defined an Evidence-Based Importance (EBI) metric representing the proportion of reviewed publications reporting a given network as relevant for sex classification (Equation 3.9).

$$\text{EBI} = \frac{\text{Number of studies reporting network as relevant}}{\text{Total number of studies reviewed}} \tag{3.9}$$

Table 3.2: Summary of network-level evidence reported in the literature regarding sex-related brain differences.

| **Network** | **EBI** | **Evidence** | **HCP score (rank)** | **CamCAN score (rank)** |
|---|---|---|---|---|
| Default Mode (DMN) | 1.00 | Strongly emphasized across all reviewed studies. DMN-related regions, including the posterior cingulate cortex (PCC), precuneus, and medial frontal cortex, are repeatedly identified as highly predictive of sex. | 0.9953 (**1**) | 0.7330 (**2**) |
| Frontoparietal / Control | 0.8 | Highlighted by Kim et al. (CCN), Shanmugan et al. (frontoparietal among the strongest contributors), Liu et al. (significant feature importance), and indirectly by Weis et al. through lateral frontal and temporo-parietal regions. | 0.8650 (**3**) | 0.6404 (**4**) |
| Ventral Attention / Salience | 0.6 | Reported in Kim et al., Shanmugan et al., and Liu et al., with additional support from Weis et al. through the involvement of the insula and cingulate regions. | 0.6678 (**4**) | 0.4913 (**6**) |
| Somatomotor | 0.3 | Identified as important in Kim et al. and Liu et al., but less consistently emphasized in the remaining studies. | 0.9251 (**2**) | 0.7217 (**3**) |
| Limbic | 0.3 | Present in Kim et al. and Liu et al. The orbitofrontal–limbic system is highlighted by Liu et al., although evidence is less consistent than for DMN or Control networks. | 0.4610 (**7**) | 0.5603 (**5**) |
| Dorsal Attention | 0.2 | Reported by Kim et al. and Liu et al., although not consistently highlighted as one of the strongest contributors across studies. | 0.6250 (**5**) | 0.4667 (**7**) |
| Visual | 0 | Reported in Kim et al. and Liu et al.; generally appears as a secondary contributor relative to association networks. | 0.5212 (**6**) | 0.7442 (**1**) |

We then compared literature-derived EBI scores with the network rankings obtained from stability selection (Table 3.2).

Overall, the network rankings derived from stability selection showed substantial agreement with the existing literature. The strongest correspondence was observed for the Default Mode and Frontoparietal/Control networks, which received the highest support across previous studies and were also among the most consistently selected systems in both HCP and CamCAN. It is worth noting that these two networks and the Ventral Attention/Salience network are, by construction, weakly correlated with one another, since the Schaefer functional network parcellation defines them as distinct large-scale systems based on differing patterns of functional connectivity [Yeo et al., 2011]. The same large-scale networks emerge as important both in prior work and in our analysis, suggesting that the predictive patterns are not due to a particular dataset or method, but instead reflect reproducible and biologically meaningful organization in the brain. Additional agreement was observed for the Limbic and Dorsal Attention networks, both of which occupy relatively modest positions in the literature and similarly showed limited importance in our analyses. The consistency of both highly ranked and weakly ranked systems strengthens the overall correspondence between our results and previous reports.

Notable discrepancies also emerged. The Somatomotor network was among the most consistently selected systems in both cohorts despite receiving comparatively limited attention in the literature. This observation suggests that sensorimotor systems may contribute more strongly to sex-related microstructural differences than previously appreciated and highlights a potentially underexplored line for future investigation. The largest discrepancy involved the Visual network. Consistent with previous studies, the network occupied a relatively low rank in HCP. In contrast, it emerged as the highest-ranked network in CamCAN despite receiving little support from the existing literature. Because this effect was not reproduced in HCP and deviates substantially from prior reports, it should be interpreted cautiously. Rather than evidence against the overall validity of the framework, this result likely reflects a cohort-specific phenomenon whose biological or methodological origins need further investigation.

Although the EBI metric provides only a simplified summary of the literature, the overall concordance between published findings and the stability-selection rankings supports the biological plausibility of the predictive results. Collectively, these results suggest that sex-related differences can be found across multiple interacting large-scale brain networks and that the structures identified by stability selection are not merely predictive, but also neurobiologically interpretable.

## 3.5 Conclusion

In this chapter, we evaluated the interpretability of array-based diffusion MRI models by studying whether the regions contributing to prediction followed a consistent pattern. We focused on sparse models, by assigning zero weight to many regions,

they reduce the number of features driving the prediction. However, cortical regions are highly correlated, meaning that different regions can carry similar predictive information and may be selected interchangeably by different models. For this reason, interpreting a single sparse solution is not enough. Instead, we used stability selection across sparse estimators, embedding representations, and train-test partitions to assess whether prediction was supported by reproducible cortical and subnetwork-level patterns rather than by model-specific feature choices.

Our analysis reveals three main findings. First, embedding representations were more coherent at the subnetwork level than at the individual-region level. Although different embeddings did not always select the same parcels, they showed stronger agreement when regions were grouped into larger functional subnetworks. This suggests that predictive information is more robustly captured at broader spatial scales. Second, the importance of subnetworks was largely consistent across the HCP and CamCAN cohorts. When comparing the rank of each subnetwork between datasets, we observed a median difference of only two positions, indicating that most subnetworks occupied similar positions in both cohorts. Third, the stable predictive networks were broadly aligned with previous neuroimaging literature on sex-related brain differences.

At the same time, some results require caution. Stability selection does not establish causality, and consistently selected regions should be interpreted as reproducible predictive markers. In addition, the strong importance of the Visual network in CamCAN, which was not reproduced in HCP, and the higher-than-expected contribution of the Somatomotor network show that some effects may depend on the cohort or remain underexplored in the current literature.

Future work should explore whether these findings generalise to other prediction tasks, parcellation schemes, and spatially informed sparse models. This chapter only applied the stability-selection framework to sex classification, but applying it to other tasks would help determine whether the identified patterns are consistent with task-specific findings reported in the literature. Moreover, using different numbers of parcels or parcellation schemes, especially those better suited to diffusion MRI, could provide a more detailed view of the relationship between brain structure and prediction. Finally, evaluating models such as the fused sparse group lasso, which explicitly accounts for the spatial organization of the brain, could lead to more biologically meaningful and spatially coherent interpretations.

# Chapter 4

# Predicting Cognition from Cortical Microstructure With Diffusion MRI

## 4.1 Introduction

A central question in cognitive neuroscience is how differences in brain structure between people give rise to differences in how they think and behave [Genon et al., 2022]. Most of what is known about this relationship comes from studying the brain's macrostructure. Gray matter volume, for instance, is known to vary alongside cognitive performance across the healthy adult lifespan [Zimmerman et al., 2006]. However, macrostructural measures such as volume only capture the large-scale outcome of processes that take place at a much smaller scale, as the density, and organisation of individual neurons and the connections between them. Understanding brain structure at this scale, cellular and microstructural level, offers the possibility of relating cognition more directly to the biological properties thought to actually support it, instead of a summary measure of them.

Progress toward this goal has been uneven. Most structural work relating brain measurements to behaviour has focused on white matter, typically studied through the reconstruction of large fibre pathways, known as tractography, connecting distant brain regions [Genon et al., 2022]. This focus is understandable, since white matter pathways are relatively well suited to current imaging and modelling techniques, but it leaves gray matter comparatively underexplored, despite gray matter being where the cell bodies, dendrites, and local connections most directly involved in cognitive processing are found [DeFelipe, 2011; Elston, 2003]. Gray matter has received less attention in part because it is genuinely harder to model. Unlike the long, coherently organised fibres of white matter, gray matter is composed of densely packed, more randomly oriented cellular structures, for which the physical models linking diffusion

measurements to underlying tissue properties are still being developed [Novikov et al., 2019].

Methodologically, most work in this area has relied on multivariate correlational techniques, Canonical Correlation Analysis (CCA) and Partial Least Squares (PLS), which identify the strongest linear association between brain and behavioural measures within a given sample [Genon et al., 2022; Helmer et al., 2024]. These methods have shaped much of what is currently known about brain-behaviour relationships, but the associations they report often do not hold up when tested on participants outside the original sample [Genon et al., 2022; Helmer et al., 2024]. This limitation has lead towards predictive modelling, in which a model's ability to predict behaviour is tested directly on data held-out from model fitting, offering a more reliable test of whether a brain measure is genuinely informative.

Addressing this gap requires both a measurement capable of capturing gray matter microstructure, and an analytical approach capable of testing whether that measurement is informative about individual people. For the first requirement, we use the return-to-origin probability (RTOP), a measure derived from diffusion MRI that indexes how restricted the movement of water molecules is within a small volume of tissue [Özarslan et al., 2013]. We focus on RTOP, rather than other microstructural measures, because it does not depend on assuming a single dominant fibre direction, an assumption that works well in white matter but is poorly suited to the cortex, where local cellular organisation is far less uniformly oriented. This makes RTOP particularly well suited to gray matter specifically. For the second requirement, we adopt a predictive modelling framework, training a model on one subset of participants and evaluating its predictions on a separate, held-out subset that played no role in fitting the model. This approach addresses the central limitation of CCA- and PLS-based analyses described above, and is becoming the standard for relating brain measurements to behaviour, particularly in work predicting cognition from brain patterns[He et al., 2020; Ooi et al., 2022].

Using this approach, we ask whether cortical RTOP can predict individual differences in cognitive performance in a large, healthy adult sample. We show that it can: models trained on cortical RTOP successfully predicted a range of cognitive abilities in participants not used to train the model, and did so considerably better than they predicted psychosocial or emotional measures, suggesting that cortical microstructure is more informative about certain kinds of behaviour than others. Finally, because a genuinely useful brain-based measure should not depend on the particular sample or scanner used to derive it, we tested whether these findings held in a second, independent dataset, acquired using different imaging parameters. Prediction generalised successfully to this dataset, supporting the conclusion that the relationship between cortical microstructure and cognition reflects a robust biological signal rather than a feature specific to a single study population.

## 4.2 Related Work

### 4.2.1 The brain-behaviour gap and the case for gray microstructure.

A goal of cognitive neuroscience is to understand why people differ from one another in how they think, remember, and reason, and how these differences relate to differences in brain structure [Genon et al., 2022]. Macrostructural measurements that capture the brain's large-scale features, such as gray matter volume or cortical thickness, have already been linked to cognitive performance; for instance, gray matter volume changes alongside cognitive ability across the healthy adult lifespan [Zimmerman et al., 2006]. However, most of the evidence supporting this idea comes from studies that examine associations across groups of participants rather than within individuals, typically asking whether a brain measure and a behavioural measure are related on average across a sample, and reporting a single association describing the group as a whole [Genon et al., 2022; Helmer et al., 2024]. This is a meaningful first step, but it leaves open whether the same association would hold in a different group of people: an association that is clear and consistent within one sample does not automatically mean it will replicate in an independent sample of individuals not used to derive it. Distinguishing between brain-behaviour relationships that have only been described within a single sample and those shown to generalise to new individuals is a question considered in this chapter.

A further related gap concerns, not just how brain-behaviour relationships are tested, but which aspect of brain structure is being tested in the first place. Existing work relating brain structure to behaviour has concentrated primarily on the organisation of white matter pathways, or on macroscopic properties of gray matter such as volume and thickness [Genon et al., 2022], leaving the cellular organisation of the cortex itself, its microstructure, underexplored. This is a notable gap, since cognitive processing depends directly on the properties of cortical neurons and their connections, including how densely they are packed, how their dendrites and axons are arranged, and how these structures are organised relative to one another [DeFelipe, 2011; Elston, 2003]. A measurement sensitive to this fine cellular organisation should carry information about cognitive ability that structural measures such as volume or thickness cannot capture.

In practice, however, this kind of measurement has been difficult to obtain. Cortical microstructure is considerably harder to characterise non-invasively than the large white matter tracts studied using tractography, both because the underlying biophysical models describing gray matter are less well established, and because the cortex itself poses specific measurement challenges [Novikov et al., 2019].

### 4.2.2 Diffusion MRI and RTOP: a candidate biomarker of cortical microstructure

Diffusion MRI can measure cortical microstructure indirectly, by tracking how water molecules move within tissue. Because this movement is constrained by the surrounding cellular structures, its statistical pattern carries information about the geometry of the tissue itself [Basser et al., 1994; Palombo et al., 2020; Zhang et al., 2012; Özarslan et al., 2013].

Many standard diffusion MRI models describe tissue in terms of a dominant fibre direction, an assumption that works well in the organised fibre bundles of white matter but breaks down in the cortex, where fibres are locally dispersed and no single direction dominates. In this setting, direction-based models tend to produce unstable and unreliable estimates.

The return-to-origin probability (RTOP) avoids this problem. RTOP is derived from the mean apparent propagator representation of the diffusion signal [Özarslan et al., 2013] and captures, for a given diffusion time, how likely a water molecule is to end up back where it started after a fixed diffusion time $t$:

$$\text{RTOP} = P(\mathbf{0}, t) = \int_{\mathbb{R}^3} \widehat{P}(\mathbf{q}, t)\, \mathrm{d}\mathbf{q}, \tag{4.1}$$

where $\widehat{P}(\mathbf{q}, t)$ is the diffusion signal measured in $q$-space, and the integral corresponds to the zeroth-order MAP-MRI coefficient. A molecule moving through open, isotropic tissue will travel far during the interval $t$ and rarely return to its starting point, giving a low RTOP value. By contrast, a molecule constrained within a small, enclosed space (such as the interior of a cell body) is pushed back toward its origin far more often, giving a high RTOP value. RTOP therefore provides a direct, model-free index of how restricted water movement is at the microscopic scale, without relying on the directional assumptions that tensor-based models require. Because it integrates the diffusion propagator over all directions, RTOP is rotationally invariant, which makes it particularly robust in the cortical ribbon, where fibre orientations are locally dispersed and tensor-based estimates tend to be unreliable. Robust estimation of RTOP from large-cohort imaging datasets, is possible to the Laplacian-regularised MAP-MRI framework [Fick et al., 2016], which improves signal fidelity and propagator reconstruction relative to standard MAP-MRI in HCP-quality data.

Biologically, a high RTOP value is thought to reflect densely packed cortical tissue, in which neurons and glial cells leave little surrounding space for water to move freely. Consistent with this interpretation, a large-scale comparison of diffusion MRI measures across cortical tissue found that RTOP, along with closely related measures RTAP and RTPP, is most strongly associated with tissue density and the amount of freely moving water, rather than with fibre orientation [Sadikov et al., 2025]. RTOP can therefore be understood as an index of how tightly packed and dense cortical tissue is, which makes it a plausible candidate for capturing individual differences in the cellular basis of cognition.

### 4.2.3 Into predictive modelling

The measurements described above make it possible to ask whether cortical microstructure relates to cognition. Methodologically, this question has traditionally been addressed using correlational techniques, such as Canonical Correlation Analysis (CCA) and Partial Least Squares (PLS), which identify the combination of brain and behavioural variables most strongly associated within a given sample [Genon et al., 2022; Helmer et al., 2024]. These methods have shaped much of the current understanding of brain-behaviour relationships, but they share an important limitation; since the same data are used both to build and to evaluate the model, the resulting association can appear strong even when it does not hold for participants outside the original sample [Genon et al., 2022; Helmer et al., 2024]. This makes harder to establish whether a brain measure could be used as a proper biomarker of cognitive ability, since a useful biomarker must generalise to subjects it was not derived from, rather than simply describing structure already present in the sample used to characterise it.

This limitation has motivated a shift toward predictive modelling, evaluated through cross-validation, which addresses the problem directly by separating the participants used to build a model from those used to test it. Rather than describing an association within a single sample, this approach trains a model on one group of participants and evaluates its predictions on a separate, held-out group who contributed nothing to model fitting. We adopt Kernel Ridge Regression (KRR) for this purpose, an approach already established as a benchmark method for predicting individual behaviour from brain measurements, most notably from patterns of functional connectivity [He et al., 2020; Ooi et al., 2022]. We apply this framework to whole-cortex RTOP.

Within this framework, prediction accuracy refers specifically to how well a model's predictions match the true scores of participants it has never seen during training. This is a different, and considerably stricter, test than an in-sample correlation, since it directly evaluates whether the brain measure generalises to new individuals rather than simply describing structure already present in the training data. It is this predictive, held-out framework that allows us to test whether cortical RTOP constitutes an informative signal of cognitive ability, and it forms the theoretical basis for the modelling approach detailed in the Methods section of this chapter.

Building on these gaps, the present chapter makes three main analyses. First, we show that whole-cortex RTOP carries information about individual cognitive performance, with a correlation of $\rho \approx 0.27$ between predicted and observed scores in held-out participants. Second, we show that cognitive measures are substantially more predictable than psychosocial or emotional traits, and that this difference is reflected in distinct patterns of how prediction accuracy scales with sample size. Third, we find that anatomically grounded parcellations of the cortex (Schaefer [Schaefer et al., 2018]) outperform parcellations derived purely from patterns of brain activity (DiFuMo [Dadi et al., 2020]), offering practical guidance for atlas selection in future microstructure-based prediction studies.

## 4.3 Methodology

### 4.3.1 Datasets

**HCP Young Adult (HCP-YA).** We used data from the HCP-YA S1200 release [Van Essen et al., 2013]. Multi-shell diffusion MRI was acquired on a customised Siemens 3T Connectome scanner with a maximum gradient strength of 100 mT/m, using three shells ($b$ = 1000, 2000, 3000 s/mm$^2$) at 1.25 mm isotropic resolution, with 90 diffusion directions per shell plus 18 additional $b = 0$ volumes. We retained all participants with complete diffusion MRI, structural MRI, and all 58 behavioural measures ($N = 884$). Before any analysis was carried out, a fixed held-out test set of 88 participants was set aside; the remaining 796 participants formed the pool used for model training and cross-validation.

**WAND.** The WAND dataset [McNabb et al., 2025] contains approximately 120 participants scanned with high-resolution multi-shell diffusion MRI at a higher maximum gradient strength than HCP-YA, allowing us to assess whether our findings generalise to data acquired under different, and in some respects more favourable, imaging conditions. The same processing pipeline was applied to WAND without modification, so that results from the two datasets could be compared directly. The behavioural measures available in WAND were WASI Vocabulary, Logical Memory (immediate recall, LM1B; delayed recall, LM2B; recognition, LM2C), and a category-based Verbal Fluency task (naming friends).

### 4.3.2 Preprocessing and data preparation

Diffusion data from both datasets were preprocessed using the HCP minimal preprocessing pipeline [Glasser et al., 2013], which corrects for susceptibility-induced distortion (TOPUP), eddy currents and motion-related signal dropout (EDDY), and gradient non-linearities. Diffusion signal was then sampled onto the cortical midthickness surface at each vertex, using the ribbon-constrained volume-to-surface mapping method provided by the HCP pipeline.

At each cortical vertex, we fitted the mean apparent propagator (MAP-MRI) model [Özarslan et al., 2013] using the regularised estimation scheme implemented in Dipy [Fick et al., 2016; Garyfallidis et al., 2014]. The model was fitted with a radial order of 6, Laplacian regularisation strength $\lambda = 0.05$, and a non-negativity constraint on the estimated propagator. From each fitted model, we computed the return-to-origin probability (RTOP) as the zeroth-order moment of the diffusion signal in $q$-space (Equation 4.1). Vertex-wise RTOP values were then averaged within each region of a given parcellation to obtain one RTOP value per cortical region, per participant.

We repeated this region-averaging step across several cortical parcellation schemes, chosen to span a range of spatial resolutions and parcellation philosophies; the anatomically derived Schaefer parcellation [Schaefer et al., 2018] at ten resolutions from 100 to 1000 parcels, in steps of 100, the functionally derived DiFuMo parcellation [Dadi

et al., 2020] at five resolutions (64, 128, 256, 512, and 1024 components) and the multimodal Brainnetome atlas [Fan et al., 2016], restricted to its 246 cortical regions. For surface visualization, the DiFuMo maps were converted into a discrete cortical parcellation by assigning each surface vertex to its most probable component (argmax) and retaining only left- and right-hemisphere cortical components, yielding 2, 26, 73, 196, and 431 cortical regions for the five nominal resolutions, respectively.

### 4.3.3 Behavioural measurements

For the analysis we used all 58 behavioural scores available in the HCP-YA dataset, spanning cognitive task performance, psychosocial self-report measures, and emotion-recognition tasks. Each measure was standardised using the mean and standard deviation computed from the training set only, so that no information from held-out participants influenced the standardisation.

To visualise how prediction accuracy varied across broad behavioural domains, we grouped the 58 behavioural measures into three categories, Cognition, Psychosocial, and Emotion, using a standard statistical technique for identifying groups of related variables within a larger set of measurements. Specifically, we applied principal component analysis (PCA), a method that summarises a large set of correlated variables using a small number of underlying components, each capturing a distinct pattern of shared variation across the original measures. Because the components produced directly by PCA can be difficult to interpret, since each one may combine contributions from many variables at once, we applied an additional step known as Varimax rotation [Kaiser, 1958], which adjusts the components so that each behavioural measure is more clearly associated with one component in particular. This makes the resulting components easier to interpret as distinct behavioural domains, following the approach used in prior work relating brain measures to behaviour [Ooi et al., 2022]. We retained three rotated components, corresponding to the three domains of interest. Each behavioural measure has a loading on each component, indicating how strongly it is associated with that component. We assigned each measure to the component on which it had the highest loading when the corresponding loading was at least 0.2. The resulting assignments defined the three behavioural domains.

### 4.3.4 Model

We used Kernel Ridge Regression (KRR) with a linear kernel to predict each behavioural measure from region-wise RTOP features, following the approach established for behavioural prediction from neuroimaging data [He et al., 2020]. The model estimates the coefficients $\alpha$ by finding a solution to the following optimisation problem:

$$\min_{\alpha} \|K\alpha - y\|^2 + \lambda\, \alpha^{\mathsf{T}} K \alpha, \qquad K = XX^{\mathsf{T}}, \tag{4.2}$$

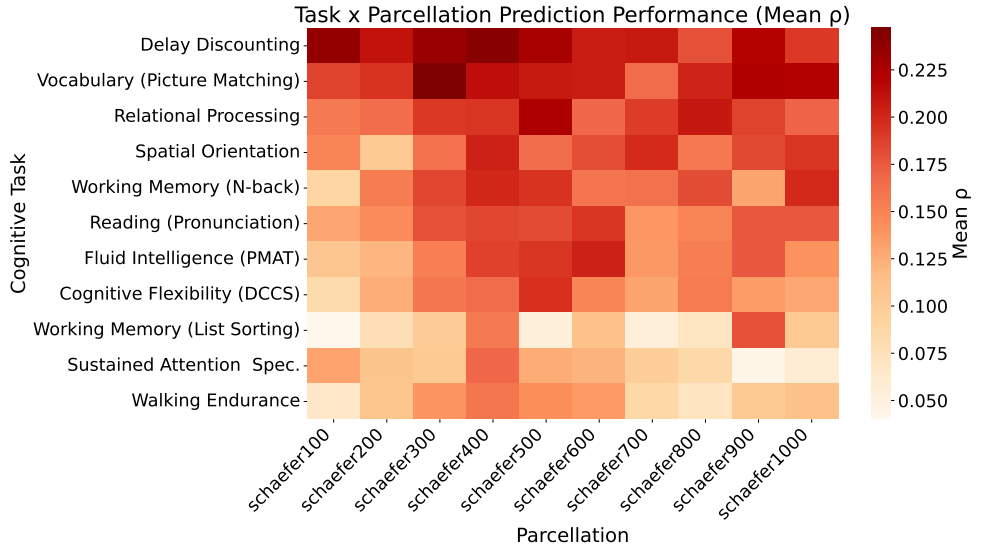



Figure 4.1: **Task-by-parcellation prediction heatmap.** Prediction accuracy, as the mean absolute Pearson correlation, ($\rho$) between predicted and observed scores, for 11 cognitive tasks that reached $\rho \geq 0.15$ at one or more resolutions across ten Schaefer resolutions. Top-performing tasks (e.g. Delay Discounting, $\rho \approx 0.27$) remain stable across resolutions, while weaker tasks are more resolution-dependent.

where $X$ denotes the matrix of region-wise RTOP features, $y$ the behavioural measure being predicted, and $\lambda$ a regularisation parameter controlling model complexity.

Model performance was evaluated using nested cross-validation. In the outer loop, data were split into 10 folds, and this procedure was repeated 10 times with different random fold assignments to improve the stability of the resulting accuracy estimates. Within each outer training fold, an inner 5-fold cross-validation loop was used to select the regularisation parameter $\lambda$ from the range $\{10^{-3}, \ldots, 10^{3}\}$. Prediction accuracy, defined as the Pearson correlation between predicted and observed behavioural scores, was computed only on the outer-fold test participants, ensuring that no participant contributed to both model fitting and evaluation for the same fold. All models were implemented using Scikit-learn [Pedregosa et al., 2011].

## 4.4 Results

**Cortical RTOP carries information about cognition with task-specific magnitude and consistency across parcellation resolutions.** Figure 4.1 shows, for each cognitive task and each resolution of the Schaefer parcellation (100 to 1000 parcels, in steps of 100), the mean absolute Pearson correlation ($\rho$) between predicted and observed scores, computed exclusively on held-out test participants across the 10 outer cross-validation folds. Of the 58 behavioural variables available in HCP-YA, 11 reached $\rho \geq 0.15$ at one or more resolutions and are shown here. The strongest

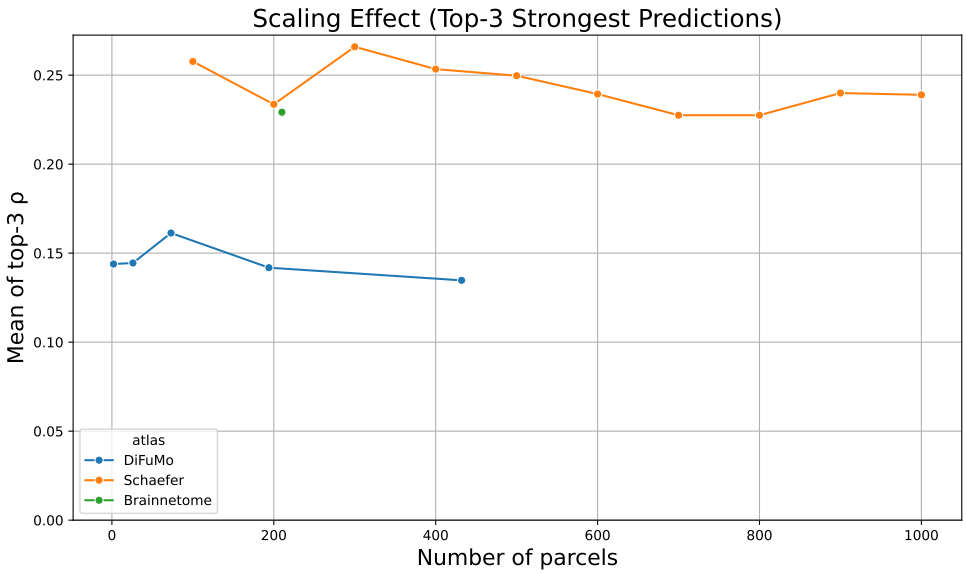



Figure 4.2: **Prediction performance across parcellation atlases and resolutions.** Comparison of predictive performance across three cortical parcellation atlases, Schaefer, DiFuMo, and Brainnetome, by plotting the mean Pearson correlation ($\rho$) of the top three cognitive predictions against the number of parcels. Schaefer outperforms DiFuMo by $\Delta\rho \approx 0.10$ on average, with Brainnetome intermediate; accuracy peaks quickly for both (around 300 Schaefer and 73 DiFuMo parcels) and declines at finer resolutions.

prediction was obtained for Delay Discounting, reaching $\rho \approx 0.27$ and remaining close to this value across most parcellation resolutions. Vocabulary (Picture Matching) and Relational Processing showed the next strongest predictions, and were similarly consistent across resolutions. A further group of tasks, including Spatial Orientation, Working Memory (N-back and List Sorting), Reading (Pronunciation), Fluid Intelligence (PMAT), and Cognitive Flexibility (DCCS), showed intermediate prediction accuracy ($0.15 \leq \rho \leq 0.20$). Sustained Attention and Walking Endurance showed the weakest predictions among the tasks displayed, which is consistent with these measures depending substantially on subcortical and cerebellar circuitry that is not captured by cortical RTOP alone.

A further pattern visible in Figure 4.1 concerns how stable each task's prediction is across parcellation resolutions. The best-performing tasks, Delay Discounting, Vocabulary, and Relational Processing, maintained high $\rho$ values continuously from 200 to 900 parcels, suggesting that the information supporting these predictions is present robustly across a wide range of spatial scales. Lower-ranked tasks, by contrast, showed greater sensitivity to parcellation resolution, which may indicate a noisier or more spatially localised microstructural signal.

**Anatomically grounded parcellations substantially outperform functional atlas decompositions.** Figure 4.2 compares the three parcellation families by plot-

ting the mean of the top three cognitive predictions against the number of parcels. Schaefer parcellations outperformed DiFuMo consistently across all tested resolutions, by approximately $\Delta\rho \approx 0.10$, with Brainnetome performing at an intermediate level between the two.

Prediction accuracy did not increase monotonically with the number of Schaefer parcels. Instead, accuracy reached its highest point quickly, around 300 parcels, and declined at finer resolutions, a pattern also observed for DiFuMo reaching its peak at 73 parcels. This indicates that increasing the number of parcels beyond a relatively low resolution does not improve prediction.

**Cognition is substantially more predictable than psychosocial or emotional traits, with a qualitatively different learning-curve profile.** Using PCA followed by Varimax rotation on the 58 HCP-YA behavioural measures, we defined three data-driven domains (Cognition, Psychosocial, and Emotion) and examined how prediction accuracy scaled with training-set size.

*Cognition.* As shown in Figure 4.3, prediction accuracy for the Cognition domain reached its highest value at the full training sample size, with $\rho = 0.287$ for Vocabulary (Picture Matching), and showed no indication of plateauing even at 800 training participants. Among individual measures within this domain, Walking Endurance showed the flattest learning curve, while Vocabulary, Fluid Intelligence (PMAT), and Spatial Orientation showed the steepest improvement with increasing sample size. The absence of a clear plateau suggests that the predictive ceiling for cortical RTOP-based cognitive prediction has not yet been reached within the present sample, and that larger cohorts would likely output better accuracy results.

*Psychosocial.* Figure 4.4 shows a markedly different pattern for the Psychosocial domain, which reached a substantially lower peak accuracy of $\rho = 0.158$ (Meaning and Purpose) and plateaued at approximately 300 training participants. This early plateau at a low ceiling contrasts qualitatively with the pattern observed for the cognitive measures, and cannot be attributed to insufficient training data alone, since cognitive predictions continued to improve well beyond this sample size using the same modelling approach.

*Emotion.* Figure 4.5 reveals a distinctive pattern within the Emotion domain. The aggregate Emotion Recognition Total score reached $\rho = 0.217$, considerably higher than any of its individual sub-scores (Angry, Fear, Neutral, and Sad recognition), each of which remained at or below $\rho = 0.10$. This substantial advantage of the composed score over its individual components illustrates a broader principle, combining several noisy individual measurements into a single composite score improves reliability and reduces measurement noise, thereby increasing the signal available for structure-behaviour prediction. While this principle is well established in psychometrics [Spearman, 1910], it is rarely demonstrated explicitly in imaging-based prediction studies. As with the psychosocial measures, individual emotion sub-scores plateaued at approximately 300 training participants.

**Cross-dataset replication in WAND, with enhanced sample efficiency**

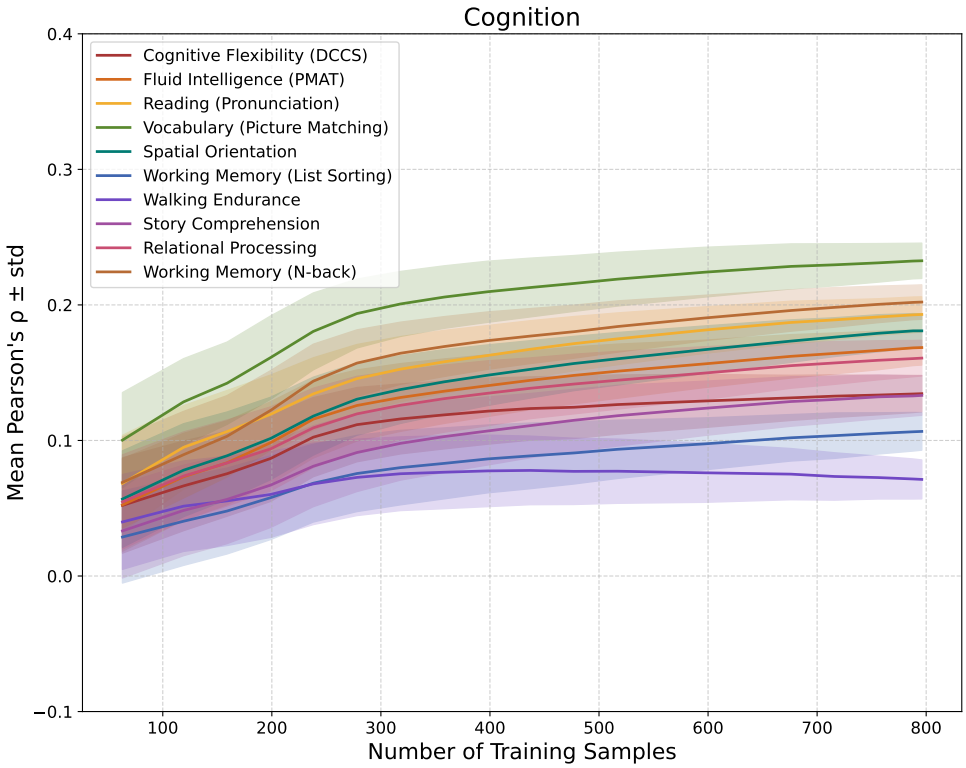



Figure 4.3: **Learning curves stratified by behavioural domain (HCP-YA).** Mean Pearson correlation ($\rho\pm$ s.d.) between predicted and observed scores as a function of training-set size, with Schaefer-800 parcels and KRR. Behavioural measures were grouped into Cognition, Psychosocial, and Emotion domains using PCA with Varimax rotation. **(A) Cognition:** peak $\rho = 0.287$ (Vocabulary), with no sign of plateauing through 800 participants.

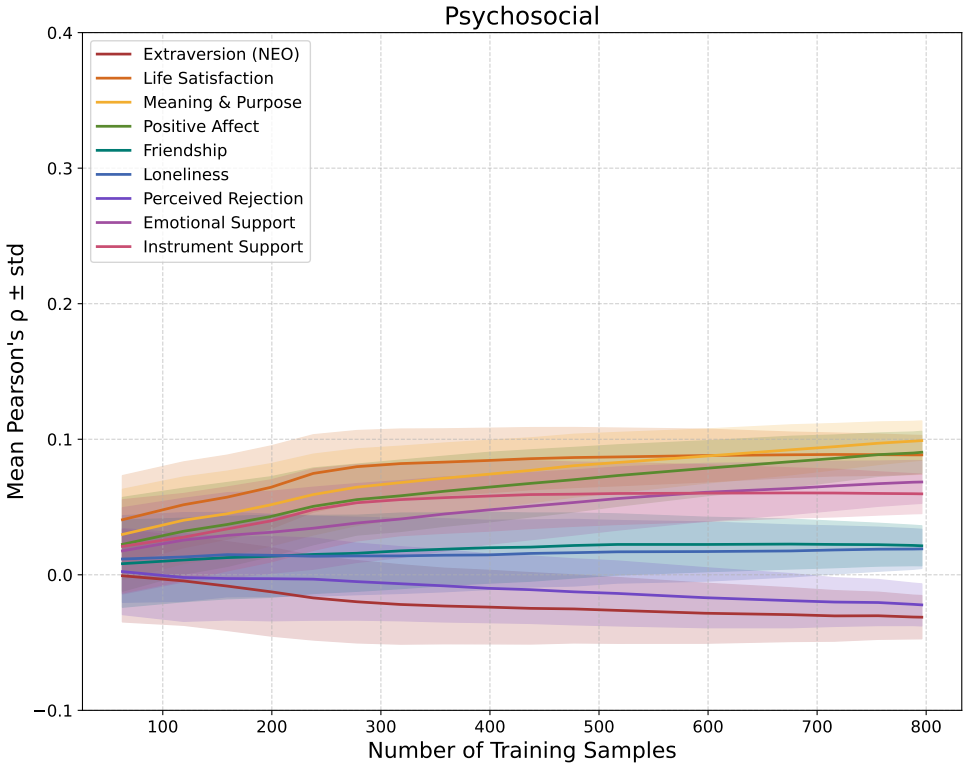



Figure 4.4: **(B) Psychosocial:** peak $\rho = 0.158$ (Meaning and Purpose), plateauing at approximately 300 training participants.

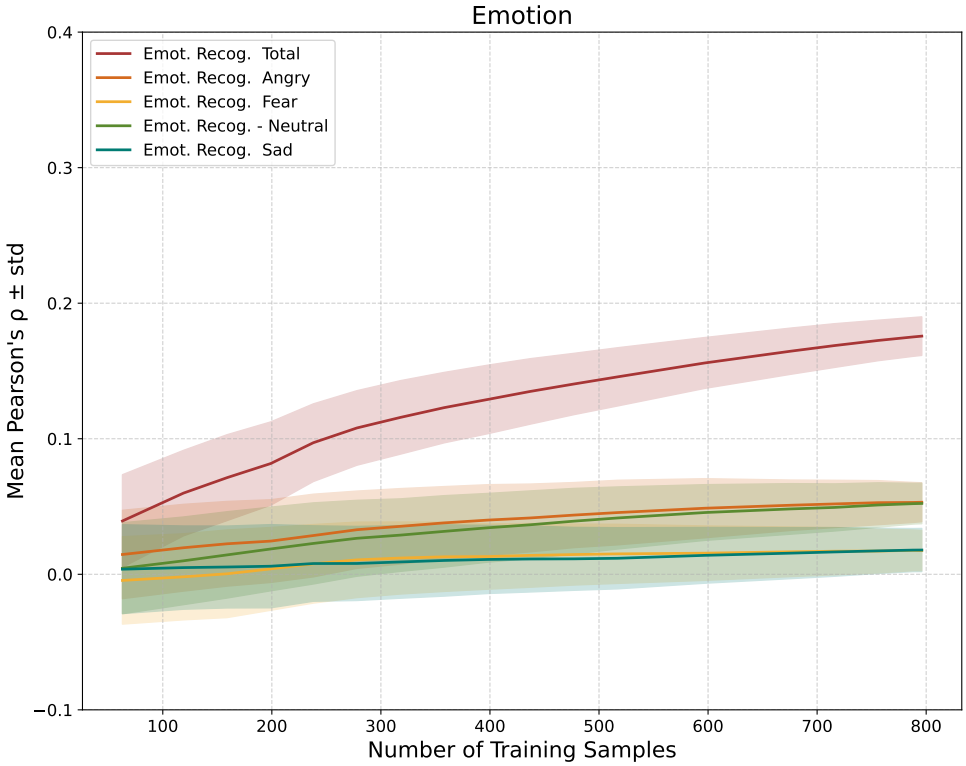



Figure 4.5: **(C) Emotion:** peak $\rho = 0.217$ (Emotion Recognition Total), considerably exceeding its individual sub-scores, illustrating the benefit of psychometric aggregation for microstructure-based prediction.

**at higher gradient strength.** Applying the identical processing and prediction pipeline to the WAND dataset confirmed that RTOP-based cognitive prediction generalises across datasets and acquisition protocols (Figure 4.6). Verbal Fluency in WAND reached $\rho = 0.185$ using only approximately 107 training participants, exceeding the corresponding prediction accuracy in HCP-YA at a matched training sample size ($\rho_{\text{HCP}} = 0.159$). This higher sample efficiency in WAND is consistent with its stronger maximum gradient strength [McNabb et al., 2025], which is known to improve the quality of diffusion signal available for microstructural estimation. This finding suggests that the ceiling on prediction accuracy is not a fixed property of the underlying brain-behaviour relationship, but instead depends, at least in part, on the quality of the acquisition used to measure it.

## 4.5 Conclusion

In this work, we evaluated whether cortical microstructure measured with diffusion MRI carries information about cognition in a predictive modelling framework. Previous brain-behaviour studies have often relied on association-based methods such

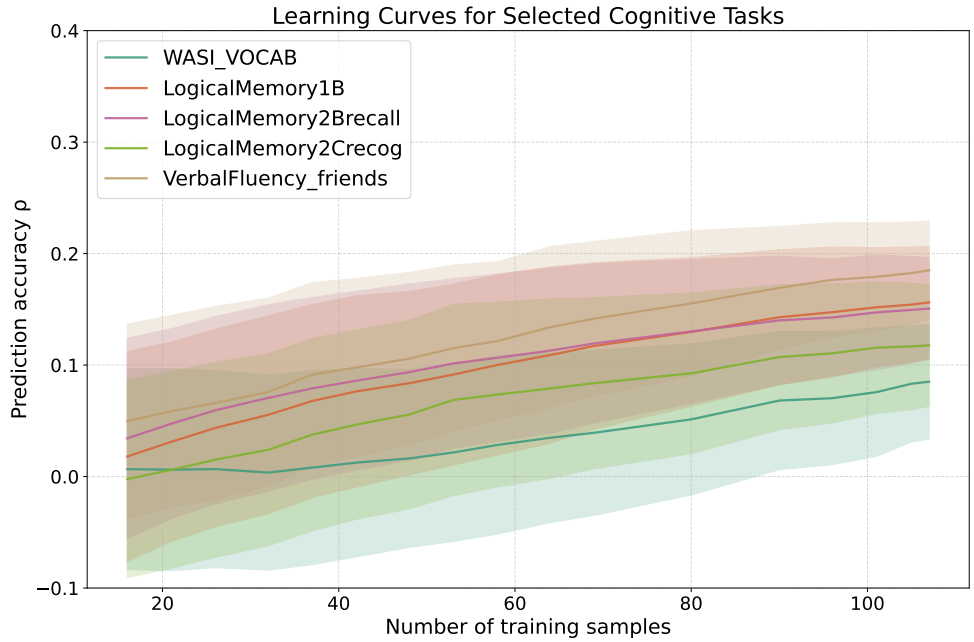



Figure 4.6: **Cross-dataset replication: learning curves in the independent WAND dataset.** Mean Pearson correlation ($\rho\pm$ s.d.) between predicted and observed scores as a function of training-set size, using Schaefer cortical RTOP features (800 parcels) and Kernel Ridge Regression.Verbal Fluency reaches $\rho = 0.185$ with only 107 training participants, exceeding the corresponding accuracy in HCP-YA at a matched training-set size ($\rho_{\text{HCP}} = 0.159$).

as CCA or PLS, which are useful for describing shared patterns between imaging and behavioural variables [Genon et al., 2022] but do not directly test how well cognition can be predicted in unseen participants. By combining the return-to-origin probability (RTOP), a diffusion-derived measure related to cellular restriction, with cross-validated prediction, we assessed whether cortical microstructural organisation contains information that can support cognitive prediction.

Three findings, taken together, support this conclusion. First, RTOP measured across the cortex reliably carries information about individual differences in cognitive performance in participants not seen during model training. Second, this predictive relationship was substantially stronger for cognitive measures than for psychosocial or emotional traits. Moreover, cognitive and non-cognitive domains showed different relationships between training sample size and predictive accuracy, suggesting that the advantage observed for cognition, does not only reflect model performance, but also a stronger link between cortical microstructure and cognitive variability. Third, this pattern generalised across datasets and acquisition protocols, with prediction generalising to an independent cohort acquired with different scanning parameters [McNabb et al., 2025]. This indicates that the relationship between cortical organisation and cognition reflects a genuine biological signal rather than an artefact of a particular

sample or imaging setup.

Together, these results show that cortical RTOP provides a biologically informative marker for studying cognitive variability from diffusion MRI. This extends previous work, mainly focused on white matter organisation or macroscopic gray matter measures such as cortical thickness and volume, and shows that diffusion-derived cortical microstructure can contribute meaningfully to cognitive prediction. At the same time, this work opens several directions for future research. Further work is needed to clarify what aspects of cortical biology RTOP captures at the level of individual variation, for example by relating RTOP to independent markers of neuronal density, cytoarchitecture, or other microstructural features where such data are available [DeFelipe, 2011; Elston, 2003]. In addition, cognitive prediction accuracy did not clearly plateau at the largest sample sizes tested here, suggesting that larger cohorts may help establish a more precise upper bound on how much cognitive variability can be explained by cortical microstructure. Finally, replication in older adults, clinical populations, and longitudinal samples would help determine whether cortical RTOP is informative for cognitive differences at a single time point and over time, which would strengthen its potential relevance as a biomarker.

# Conclusions and Perspectives

In this work, we developed a set of methodological and applied contributions with the objective of improving and understanding how diffusion MRI is used to study the relationship between brain structure and behaviour. Chapter 2 introduced an open-source benchmarking framework for standardized, reproducible comparison of predictive models across diffusion MRI modalities and methodological choices, revealing substantial variability in performance depending on feature choice and models, and underscoring the need for common evaluation protocols in the field. Chapter 3 introduced stability selection as an approach for improving the reliability and interpretability of feature selection across variable experimental configurations, showing that this way consistently identifies a stable subset of predictive variables across resampling schemes. Chapter 4 showed that cortical RTOP, a diffusion-derived index of cellular restriction [Özarslan et al., 2013], carries information about individual cognitive performance more reliably than psychosocial or emotional measures. This relationship also generalises across independent datasets [McNabb et al., 2025], establishing RTOP as a biologically grounded correlate of cognition. Overall, this work contributes a set of tools and findings that support more reliable, reproducible, and theoretically grounded uses of diffusion MRI for studying brain-behaviour relationships. This work also opens several directions for future work, considering methodological improvements within each chapter as well as broader considerations for the field.

**Integrating harmonization into the benchmarking framework is a key step toward improving cross-dataset generalisability.** Building on the benchmarking framework introduced in Chapter 2, an important next step is to integrate harmonization directly into the evaluation pipeline. The predictive models used have been trained and evaluated separately within each dataset, and while this design was necessary to establish reliable baseline performance for each cohort individually, it leaves an open question: how much of the variability observed between datasets reflects genuine differences between the populations studied, and how much reflects differences in scanner hardware, acquisition protocol, or preprocessing choices. Harmonization methods are designed specifically to address this second source of variability, by statistically adjusting imaging features so that differences given by scanner or site are reduced, while differences of biological interest are preserved [Fortin et al., 2018]. Incorporating harmonization directly into the benchmark, would allow differ-

ent harmonization strategies to be compared systematically, and would help clarify which combinations of harmonization method and acquisition protocol most reliably improve predictive generalisability across datasets. A related, and currently unanswered, question is at which stage harmonization should ideally be applied.

Beyond harmonization, domain adaptation offers a complementary strategy. Rather than adjusting the input data itself, domain adaptation methods modify the learning process so that a model trained on one dataset can perform reliably when applied to another, a scenario that mirrors closely how these methods would likely be used in practice, for instance when a model developed on a large research cohort is later applied to data collected in a clinical setting [Wachinger et al., 2021]. Testing whether such methods improve cross-dataset performance beyond what harmonization alone can achieve would help determine whether the two approaches are best used together or independently. Finally, and more broadly, the specific preprocessing pipeline used to generate imaging features, is itself a source of variability that has not been systematically examined here; comparing how different, commonly used preprocessing pipelines affect predictive performance and stability would help identify which processing decisions matter most for obtaining results that are both reliable and reproducible.

**Three complementary directions would extend the benchmark's reach: modelling, task coverage, and alternative data representations.** The benchmark developed compares a defined set of models across several microstructural representations and datasets; this range, however, leaves several natural extensions underexplored. On the modelling side, deep learning architectures were not fully explored, despite their growing use in neuroimaging prediction tasks; a systematic evaluation of such architectures, including appropriate strategies for fine-tuning them to a given task and dataset, would clarify whether their added complexity translates into a meaningful predictive advantage over the simpler, more interpretable models considered, or whether, simpler models remain competitive. Testing this broader set of models across a varied and wider range of datasets, beyond those used here, would further strengthen confidence that the benchmark's conclusions reflect general properties of these methods, rather than characteristics specific to the particular cohorts studied.

On the task side, building further on the autism prediction task in order to improve the results and achieve current state-of-the-art performance is a natural next step. Extending the benchmark to additional tasks of interest to the neuroscience community, for instance by incorporating additional clinical cohorts or related neurodevelopmental conditions, would increase the practical relevance of the benchmark's conclusions beyond the specific tasks considered. Finally, the benchmark in its current form is restricted to representations of brain data defined on arrays or volumes; extending it to alternative representations, such as graphs or cortical meshes that explicitly encode the spatial relationships between brain regions, would allow a wider range of modelling approaches, including methods drawn from geometric deep learning, to be evaluated within the same standardized framework, and would help determine whether representing this spatial structure explicitly offers a genuine predictive advantage over the

representations used [Bronstein et al., 2017].

**Spatial awareness may lead to more reliable feature selection.** The stability selection approach introduced in Chapter 3 could be extended by incorporating spatial structure more explicitly into the sparse models. The Group Lasso and Sparse Group Elastic Net estimators used, treat brain regions as independent units, so two regions that are spatially adjacent, and therefore often strongly correlated, may still be selected inconsistently across runs, since the model has no built-in notion of which regions are neighbours. A natural extension is the fused sparse group lasso, which augments the standard group penalty with an additional term that encourages neighbouring regions to be assigned similar coefficient values:

$$\Omega_{\text{FSGL}}(\beta) = \lambda_1 \sum_{g=1}^{G} \|\beta_g\|_2 + \lambda_2 \|\beta\|_1 + \lambda_3 \sum_{(g,g')\in\mathcal{N}} w_{gg'} \|\beta_g - \beta_{g'}\|_1, \tag{4.3}$$

where the first two terms correspond to the region-level and within-region sparsity already used by the estimators considered in this thesis, and the third, additional term penalises differences in coefficients between spatially neighbouring regions, with $\mathcal{N}$ denoting the set of region pairs considered neighbours and $w_{gg'}$ reflecting how close each pair is in space. In practice, this last term acts as a smoothing constraint; instead of letting the model select one region from a correlated neighbouring pair arbitrarily, it encourages both to be treated similarly, reducing the risk that selection switches unpredictably between them across different training runs. Evaluating whether this additional structure improves regional stability would help determine whether explicitly accounting for spatial proximity leads to more consistent and reliable feature selection than the region-independent estimators used here. This question is particularly relevant given the difficulty, in distinguishing between strongly correlated neighbouring regions using models that consider each region independently. Beyond its effect on regional stability, it would also be valuable to assess whether this spatial structure improves interpretability specifically at the level of individual regions, potentially obtaining results that are more anatomically precise than the broader subnetwork-level patterns.

**Larger datasets and more expressive models could strengthen the predictive utility of cortical RTOP.** The work relating cortical RTOP to cognition in Chapter 4 could be extended in several ways. Applying this approach to larger cohorts, such as the UK Biobank, would allow the predictive power of RTOP to be tested across a substantially larger and more diverse sample of participants [Miller et al., 2016], while testing alternative models to Kernel Ridge Regression may help improve prediction for psychosocial and emotional measures, which performed comparatively poorly in our results. It would also be worth to further investigate the geometric properties underlying RTOP itself, and how the cortex is represented as a mesh, for instance through models based on spheres, graph-based, or other geometric representations of cortical diffusion, given that the true geometry of cortical tissue is considerably more complex than simple restriction models assume. In particular,

representing the diffusion signal as a quantity defined directly on the cortical surface, rather than on a regular volumetric grid, could allow more accurate modelling of the relationship between tissue microstructure and cognitive outcomes, and would open the door to geometric deep learning methods specifically designed for this kind of non-Euclidean, surface-based data [Bronstein et al., 2017].

**Clinical impact depends on balancing predictive performance with interpretability.** Beyond these chapter-specific directions, clinical translation remains an important broader objective. Achieving this will require carefully balancing interpretability against predictive accuracy, so that these methods can be responsibly adopted and trusted by clinical experts. Relatedly, closer collaboration with neuroscientists and psychologists would help refine how cognitive, psychosocial, and emotional measures are grouped and assessed, supporting the development of behavioural assessments that more faithfully capture the constructs these models aim to predict. More broadly, continued methodological work on gray matter microstructural modelling is needed to characterise which aspects of tissue microstructure are most informative across different demographic and behavioural outcomes, and to better understand the relationship between diffusion MRI metrics and cognitive outcomes at a finer level. Applying these approaches to longitudinal datasets would further allow cognitive change to be tracked over time, offering a more complete picture of how microstructural properties relate to cognitive trajectories rather than to a single point in time.

# Bibliography

A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gramfort, B. Thirion, and G. Varoquaux. Machine learning for neuroimaging with scikit-learn. *Frontiers in neuroinformatics*, 8:14, 2014.

H. Adeli, S. Minni, and N. Kriegeskorte. Transformer brain encoders explain human high-level visual responses. *Advances in neural information processing systems*, 38: 56840–56871, 2026.

D. C. Alexander, T. B. Dyrby, M. Nilsson, and H. Zhang. Imaging brain microstructure with diffusion mri: practicality and applications. *NMR in Biomedicine*, 32(4): e3841, 2019.

A. Alkemade, R. Großmann, P.-L. Bazin, and B. U. Forstmann. Mixed methodology in human brain research: integrating mri and histology. *Brain Structure and Function*, 228(6):1399–1410, 2023.

J. L. Andersson and S. N. Sotiropoulos. An integrated approach to correction for off-resonance effects and subject movement in diffusion mr imaging. *Neuroimage*, 125:1063–1078, 2016a.

J. L. R. Andersson and S. N. Sotiropoulos. An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. *NeuroImage*, 125:1063–1078, Jan. 2016b. ISSN 1053-8119.

J. Annese. The importance of combining mri and large-scale digital histology in neuroimaging studies of brain connectivity and disease. *Frontiers in Neuroinformatics*, 6:13, 2012.

A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

Y. Assaf and O. Pasternak. Diffusion tensor imaging (dti)-based white matter mapping in brain research: a review. *Journal of molecular neuroscience*, 34(1):51–61, 2008.

B. B. Avants, N. Tustison, and H. Johnson. Advanced Normalization Tools (ANTS). 2(365):1–35, 2009.

S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

P. J. Basser and D. K. Jones. Diffusion-tensor mri: theory, experimental design and data analysis–a technical review. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo*, 15 (7-8):456–467, 2002.

P. J. Basser and C. Pierpaoli. Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor mri. *Journal of magnetic resonance*, 213 (2):560–570, 2011.

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

C. Beaulieu. The basis of anisotropic water diffusion in the nervous system–a technical review. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo*, 15(7-8):435–455, 2002.

B. J. Bedell. Diffusion mri and the diffusion tensor model. https://imaging-cro.biospective.com/our-innovations/diffusion-mri-ftd/presentation/introduction, 2026. Accessed: 2026-07-09.

C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

R. Botvinik-Nezer, F. Holzmeister, C. F. Camerer, A. Dreber, J. Huber, M. Johannesson, M. Kirchler, R. Iwanir, J. A. Mumford, R. A. Adcock, et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810): 84–88, 2020.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

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

R. W. Brown, Y.-C. N. Cheng, E. M. Haacke, M. R. Thompson, and R. Venkatesan. *Magnetic resonance imaging: physical principles and sequence design*. John Wiley & Sons, 2014.

N. Burkart and M. F. Huber. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317, 2021.

N. A. Cardenas-Hernandez, M. Perez-Diaz, K. B. García-Ramó, and M. d. C. Valdés Hernández. Autism spectrum disorder detection using diffusion tensor imaging and machine learning. *PLOS Digital Health*, 4(12):e0001155, 2025.

C.-L. Chen, Y.-C. Hsu, L.-Y. Yang, Y.-H. Tung, W.-B. Luo, C.-M. Liu, T.-J. Hwang, H.-G. Hwu, and W.-Y. I. Tseng. Generalization of diffusion magnetic resonance imaging–based brain age prediction model through transfer learning. *NeuroImage*, 217:116831, 2020a.

G. Chen, Y. Hong, K. M. Huynh, and P.-T. Yap. Deep learning prediction of diffusion mri data with microstructure-sensitive loss functions. *Medical image analysis*, 85: 102742, 2023a.

H. Chen, I. C. Covert, S. M. Lundberg, and S.-I. Lee. Algorithms to estimate shapley value feature attributions. *Nature Machine Intelligence*, 5(6):590–601, 2023b.

J. Chen, V. L. Bayanagari, S. Chung, Y. Wang, and Y. W. Lui. Deep learning with diffusion mri as in vivo microscope reveals sex-related differences in human white matter microstructure. *Scientific reports*, 14(1):9835, 2024.

S. Chen, K. Ma, and Y. Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020b.

M. Cieslak, P. A. Cook, X. He, F.-C. Yeh, T. Dhollander, A. Adebimpe, G. K. Aguirre, D. S. Bassett, R. F. Betzel, J. Bourque, et al. Qsiprep: an integrative platform for preprocessing and reconstructing diffusion mri data. *Nature methods*, 18(7): 775–778, 2021.

L. Cordero-Grande, D. Christiaens, J. Hutter, A. N. Price, and J. V. Hajnal. Complex diffusion-weighted image estimation via matrix recovery under general noise models. *NeuroImage*, 200:391–404, Oct. 2019. ISSN 1053-8119.

C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

R. W. Cox. AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages. *Computers and Biomedical Research*, 29(3):162–173, June 1996. ISSN 0010-4809.

K. Dadi, G. Varoquaux, A. Machlouzarides-Shalit, K. J. Gorgolewski, D. Wassermann, B. Thirion, and A. Mensch. Fine-grain atlases of functional modes for fmri analysis. *NeuroImage*, 221:117126, 2020.

A. M. Dale, B. Fischl, and M. I. Sereno. Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. *NeuroImage*, 9(2):179–194, Feb. 1999. ISSN 1053-8119.

C. Dancette, J. Khlaut, A. Saporta, H. Philippe, E. Ferreres, B. Callard, T. Danielou, L. Alberge, L. Machado, D. Tordjman, et al. Curia: A multi-modal foundation model for radiology. *arXiv preprint arXiv:2509.06830*, 2025.

S. M. de Mooij, R. N. Henson, L. J. Waldorp, and R. A. Kievit. Age differentiation within gray matter, white matter, and between memory and white matter in an adult life span cohort. *Journal of Neuroscience*, 38(25):5826–5836, 2018.

J. DeFelipe. The evolution of the brain, the human nature of cortical circuits, and intellectual creativity. *Frontiers in neuroanatomy*, 5:11068, 2011.

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

A. Di Martino, D. O'connor, B. Chen, K. Alaerts, J. S. Anderson, M. Assaf, J. H. Balsters, L. Baxter, A. Beggiato, S. Bernaerts, et al. Enhancing studies of the connectome in autism using the autism brain imaging data exchange ii. *Scientific data*, 4(1):1–15, 2017.

M. Dibaji, J. Ospel, R. Souza, and M. Bento. Sex differences in brain mri using deep learning toward fairer healthcare outcomes. *Frontiers in Computational Neuroscience*, 18:1452457, 2024.

G. A. G. Diez. *Inferring and comparing structural parcellations of the human brain using diffusion MRI*. PhD thesis, COMUE Université Côte d'Azur (2015-2019), 2018.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

M. Egle, S. Hilal, A. M. Tuladhar, L. Pirpamer, E. Hofer, M. Duering, J. Wason, R. G. Morris, M. Dichgans, R. Schmidt, et al. Prediction of dementia using diffusion tensor mri measures: the optimal collaboration. *Journal of Neurology, Neurosurgery & Psychiatry*, 93(1):14–23, 2022.

A. Einstein. On the theory of the brownian movement. *Ann. Phys*, 19(4):371–381, 1906.

Y. ElNakieb, M. T. Ali, A. Elnakib, A. Shalaby, A. Soliman, A. Mahmoud, M. Ghazal, G. N. Barnes, and A. El-Baz. The role of diffusion tensor mr imaging (dti) of the

brain in diagnosing autism spectrum disorder: promising results. *Sensors*, 21(24): 8171, 2021.

A. D. Elster. What is diffusion? https://mriquestions.com/what-is-diffusion.html, 2026. Accessed: 2026-07-09.

G. N. Elston. Cortex, cognition and the cell: new insights into the pyramidal neuron and prefrontal function. *Cerebral cortex*, 13(11):1124–1138, 2003.

B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline. Machine learning for medical imaging. *radiographics*, 37(2):505–515, 2017.

O. Esteban, C. J. Markiewicz, R. W. Blair, C. A. Moodie, A. I. Isik, A. Erramuzpe, J. D. Kent, M. Goncalves, E. DuPre, M. Snyder, et al. fmriprep: a robust preprocessing pipeline for functional mri. *Nature methods*, 16(1):111–116, 2019.

L. Fan, H. Li, J. Zhuo, Y. Zhang, J. Wang, L. Chen, Z. Yang, C. Chu, S. Xie, A. R. Laird, et al. The human brainnetome atlas: a new brain atlas based on connectional architecture. *Cerebral cortex*, 26(8):3508–3526, 2016.

A. V. Faria, J. Zhang, K. Oishi, X. Li, H. Jiang, K. Akhter, L. Hermoye, S.-K. Lee, A. Hoon, E. Stashinko, et al. Atlas-based analysis of neurodevelopment from infancy to adulthood using diffusion tensor imaging and applications for automated abnormality detection. *Neuroimage*, 52(2):415–428, 2010.

R. H. Fick, D. Wassermann, E. Caruyer, and R. Deriche. Mapl: Tissue microstructure estimation using laplacian-regularized map-mri and its application to hcp data. *NeuroImage*, 134:365–385, 2016.

R. D. Fields. White matter in learning, cognition and psychiatric disorders. *Trends in neurosciences*, 31(7):361–370, 2008.

C. M. Filley and R. D. Fields. White matter and cognition: making the connection. *Journal of Neurophysiology*, 116(5):2093–2104, Nov. 2016. ISSN 1522-1598. doi: 10.1152/jn.00221.2016. URL http://dx.doi.org/10.1152/jn.00221.2016.

A. Fornito, A. Zalesky, and E. Bullmore. *Fundamentals of brain network analysis*. Academic press, 2016.

J.-P. Fortin, N. Cullen, Y. I. Sheline, W. D. Taylor, I. Aselcioglu, P. A. Cook, P. Adams, C. Cooper, M. Fava, P. J. McGrath, et al. Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage*, 167:104–120, 2018.

P. Gao and S. Ganguli. On simplicity and complexity in the brave new world of large-scale neuroscience. *Current opinion in neurobiology*, 32:148–155, 2015.

E. Garyfallidis, M. Brett, B. Amirbekian, A. Rokem, S. Van Der Walt, M. Descoteaux, and I. Nimmo-Smith. Dipy, a library for the analysis of diffusion MRI data. *Frontiers in Neuroinformatics*, 8, Feb. 2014. ISSN 1662-5196.

S. Genon, S. B. Eickhoff, and S. Kharabian. Linking interindividual variability in brain structure to behaviour. *Nature Reviews Neuroscience*, 23(5):307–318, 2022.

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*, 80:105–124, 2013.

D. N. Greve and B. Fischl. Accurate and robust brain image alignment using boundary-based registration. *Neuroimage*, 48(1):63–72, 2009.

T. Hastie, R. Tibshirani, J. Friedman, et al. The elements of statistical learning, 2009.

S. Haufe, F. Meinecke, K. Görgen, S. Dähne, J.-D. Haynes, B. Blankertz, and F. Bießmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*, 87:96–110, 2014.

B. He and Z. Liu. Multimodal functional neuroimaging: integrating functional mri and eeg/meg. *IEEE reviews in biomedical engineering*, 1:23–40, 2008.

H. He, F. Zhang, S. Pieper, N. Makris, Y. Rathi, W. Wells, and L. J. O'Donnell. Model and predict age and sex in healthy subjects using brain white matter features: a deep learning approach. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2022.

T. He, R. Kong, A. J. Holmes, M. Nguyen, M. R. Sabuncu, S. B. Eickhoff, D. Bzdok, J. Feng, and B. T. Yeo. Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *NeuroImage*, 206:116276, 2020.

M. Helmer, S. Warrington, A.-R. Mohammadi-Nejad, J. L. Ji, A. Howell, B. Rosand, A. Anticevic, S. N. Sotiropoulos, and J. D. Murray. On the stability of canonical correlation analysis and partial least squares with application to brain-behavior associations. *Communications biology*, 7(1):217, 2024.

I. Hwang, E. K. Yeon, J. Y. Lee, R.-E. Yoo, K. M. Kang, T. J. Yun, S. H. Choi, C.-H. Sohn, H. Kim, and J.-h. Kim. Prediction of brain age from routine t2-weighted spin-echo brain magnetic resonance images with a deep convolutional neural network. *Neurobiology of Aging*, 105:78–85, 2021.

I. O. Jelescu and M. D. Budde. Design and validation of diffusion mri models of white matter. *Frontiers in physics*, 5:61, 2017.

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

J. H. Jensen, J. A. Helpern, A. Ramani, H. Lu, and K. Kaczynski. Diffusional kurtosis imaging: the quantification of non-gaussian water diffusion by means of magnetic

resonance imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 53(6):1432–1440, 2005.

H. Johansen-Berg and T. E. Behrens. *Diffusion MRI: from quantitative measurement to in vivo neuroanatomy*. Academic Press, 2013.

Johns Hopkins Medicine. Anatomy of the brain. https://www.hopkinsmedicine.org/health/conditions-and-diseases/anatomy-of-the-brain, 2024. Accessed: 2026-07-09.

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*, 73:239–254, 2013.

H. F. Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, 1958.

E. Kellner, B. Dhital, V. G. Kiselev, and M. Reisert. Gibbs-ringing artifact removal based on local subvoxel-shifts. *Magnetic resonance in medicine*, 76(5):1574–1581, 2016.

B.-H. Kim and J. C. Ye. Understanding graph isomorphism network for rs-fmri functional connectivity analysis. *Frontiers in neuroscience*, 14:630, 2020.

T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

R. Kotikalapudi, B. Kincses, G. Gallitto, R. Englert, K. Hofffschlag, J. Li, C. Büchel, U. Bingel, and T. Spisak. On the replicability of diffusion weighted mri-based brain-behavior models. *Communications Biology*, 8(1):1512, 2025.

J. Kruper, A. Richie-Halford, J. Qiao, A. Gilmore, K. Chang, M. Grotheer, E. Roy, S. Caffarra, T. Gomez, S. Chou, et al. A software ecosystem for brain tractometry processing, analysis, and insight. *PLoS computational biology*, 21(8):e1013323, 2025.

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. In *International conference on medical image computing and computer-assisted intervention*, pages 469–477. Springer, 2017.

D. Le Bihan. Looking into the functional architecture of the brain with diffusion mri. *Nature reviews neuroscience*, 4(6):469–480, 2003.

D. Le Bihan. Diffusion mri: what water tells us about the brain. *EMBO molecular medicine*, 6(5):569–573, 2014.

D. Le Bihan and E. Breton. Imagerie de diffusion in-vivo par résonance magnétique nucléaire. *Comptes-Rendus de l'Académie des Sciences*, 93(5):27–34, 1985.

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

S. Lee, E. Lee, K.-S. Lee, and S.-B. Pyun. Explainable artificial intelligence on safe balance and its major determinants in stroke patients. *Scientific reports*, 14(1): 23735, 2024.

J. P. Lerch, A. J. Van Der Kouwe, A. Raznahan, T. Paus, H. Johansen-Berg, K. L. Miller, S. M. Smith, B. Fischl, and S. N. Sotiropoulos. Studying neuroanatomy using mri. *Nature neuroscience*, 20(3):314–326, 2017.

C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of experimental social psychology*, 49(4):764–766, 2013.

L. Lin, Y. Xue, Q. Duan, B. Sun, H. Lin, X. Chen, L. Luo, X. Wei, and Z. Zhang. Microstructural white matter abnormalities and cognitive dysfunction in subcortical ischemic vascular disease: an atlas-based diffusion tensor analysis study. *Journal of Molecular Neuroscience*, 56(2):363–370, 2015.

G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

T. Liu, J. Dey, B. Xu, E. W. Bridgeford, S. Alldritt, K.-H. Nenning, K. Byeon, T. Xu, and J. T. Vogelstein. Statistically valid explainable black-box machine learning: applications in sex classification across species using brain imaging. *PloS one*, 21 (4):e0346575, 2026.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

S. Marek, B. Tervo-Clemmens, F. J. Calabro, D. F. Montez, B. P. Kay, A. S. Hatoum, M. R. Donohue, W. Foran, R. L. Miller, T. J. Hendrickson, et al. Reproducible brain-wide association studies require thousands of individuals. *Nature*, 603(7902): 654–660, 2022.

C. B. McNabb, I. D. Driver, V. Hyde, G. Hughes, H. L. Chandler, H. Thomas, C. Allen, E. Messaritaki, C. J. Hodgetts, C. Hedge, et al. Wand: A multi-modal dataset integrating advanced mri, meg, and tms for multi-scale brain analysis. *Scientific data*, 12(1):220, 2025.

N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(4):417–473, 2010.

S. L. Mendes, W. H. L. Pinaya, P. Pan, and J. R. Sato. Estimating gender and age from brain structural mri of children and adolescents: A 3d convolutional neural network multitask learning model. *Computational intelligence and neuroscience*, 2021(1):5550914, 2021.

V. Menon, G. Gallardo, M. A. Pinsk, V.-D. Nguyen, J.-R. Li, W. Cai, and D. Wassermann. Microstructural organization of human insula is linked to its macrofunctional circuitry and predicts cognitive control. *elife*, 9:e53470, 2020.

M. M. Mielke, N. A. Kozauer, K. C. G. Chan, M. George, J. Toroney, M. Zerrate, K. Bandeen-Roche, M.-C. Wang, J. Pekar, S. Mori, et al. Regionally-specific diffusion tensor imaging in mild cognitive impairment and alzheimer's disease. *Neuroimage*, 46(1):47–55, 2009.

K. L. Miller, F. Alfaro-Almagro, N. K. Bangerter, D. L. Thomas, E. Yacoub, J. Xu, A. J. Bartsch, S. Jbabdi, S. N. Sotiropoulos, J. L. Andersson, et al. Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nature neuroscience*, 19(11):1523–1536, 2016.

M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

S. Mori and P. C. Van Zijl. Fiber tracking: principles and strategies–a technical review. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo*, 15(7-8):468–480, 2002.

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

S. Mori, K. Oishi, H. Jiang, L. Jiang, X. Li, K. Akhter, K. Hua, A. V. Faria, A. Mahmood, R. Woods, et al. Stereotaxic white matter atlas based on diffusion tensor imaging in an icbm template. *Neuroimage*, 40(2):570–582, 2008.

A. N. Nielsen, D. M. Barch, S. E. Petersen, B. L. Schlaggar, and D. J. Greene. Machine learning with neuroimaging: evaluating its applications in psychiatry. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 5(8):791–798, 2020.

D. S. Novikov. The present and the future of microstructure mri: From a paradigm shift to normal science. *Journal of Neuroscience Methods*, 351:108947, 2021.

D. S. Novikov, E. Fieremans, S. N. Jespersen, and V. G. Kiselev. Quantifying brain microstructure with diffusion mri: Theory and parameter estimation. *NMR in Biomedicine*, 32(4):e3998, 2019.

L. Q. R. Ooi, J. Chen, S. Zhang, R. Kong, A. Tam, J. Li, E. Dhamala, J. H. Zhou, A. J. Holmes, and B. T. Yeo. Comparison of individualized behavioral predictions across anatomical, diffusion and functional connectivity mri. *NeuroImage*, 263: 119636, 2022.

M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

E. Özarslan, C. G. Koay, T. M. Shepherd, M. E. Komlosh, M. O. İrfanoğlu, C. Pierpaoli, and P. J. Basser. Mean apparent propagator (map) mri: a novel diffusion imaging method for mapping tissue microstructure. *NeuroImage*, 78:16–32, 2013.

C. Paillette, P. Griset, and Y. Agid. Aux origines des neurosciences-l'héritage et les fondations, du xixe siècle aux années 1940, 2021.

M. Palombo, A. Ianus, M. Guerreri, D. Nunes, D. C. Alexander, N. Shemesh, and H. Zhang. Sandi: A compartment-based model for non-invasive apparent soma and neurite imaging by diffusion mri. *Neuroimage*, 215:116835, 2020.

B. Panigrahi, S. Razavi, L. E. Doig, B. Cordell, H. V. Gupta, and K. Liber. On robustness of the explanatory power of machine learning models: Insights from a new explainable ai approach using sensitivity analysis. *Water Resources Research*, 61(3):e2024WR037398, 2025.

S. Parisot, S. I. Ktena, E. Ferrante, M. Lee, R. Guerrero, B. Glocker, and D. Rueckert. Disease prediction using graph convolutional networks: application to autism spectrum disorder and alzheimer's disease. *Medical image analysis*, 48:117–130, 2018.

O. Pasternak, N. Sochen, Y. Gur, N. Intrator, and Y. Assaf. Free water elimination and mapping from diffusion mri. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 62(3): 717–730, 2009.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

H. Peng, W. Gong, C. F. Beckmann, A. Vedaldi, and S. M. Smith. Accurate brain age prediction with lightweight deep neural networks. *Medical image analysis*, 68: 101871, 2021.

L. Penke, S. M. Maniega, M. E. Bastin, M. C. Valdés Hernández, C. Murray, N. A. Royle, J. M. Starr, J. M. Wardlaw, and I. J. Deary. Brain white matter tract integrity as a neural foundation for general intelligence. *Molecular Psychiatry*, 17(10):1026–1030, May 2012. ISSN 1476-5578. doi: 10.1038/mp.2012.66. URL http://dx.doi.org/10.1038/mp.2012.66.

R. A. Poldrack, G. Huckins, and G. Varoquaux. Establishment of best practices for evidence for prediction: a review. *JAMA psychiatry*, 77(5), 2020.

J. Qian, H. Li, J. Wang, and L. He. Recent advances in explainable artificial intelligence for magnetic resonance imaging. *Diagnostics*, 13(9):1571, 2023.

M. E. Raichle. Behind the scenes of functional brain imaging: a historical and physiological perspective. *Proceedings of the National Academy of Sciences*, 95(3):765–772, 1998.

N. Raz and K. M. Rodrigue. Differential aging of the brain: patterns, cognitive correlates and modifiers. *Neuroscience & Biobehavioral Reviews*, 30(6):730–748, 2006.

M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

R. Rosenbacke, Å. Melhus, M. McKee, and D. Stuckler. How explainable artificial intelligence can increase or decrease clinicians' trust in ai applications in health care: systematic review. *Jmir Ai*, 3:e53207, 2024.

M. Rosenblatt, L. Tejavibulya, H. Sun, C. C. Camp, M. Khaitova, B. D. Adkinson, R. Jiang, M. L. Westwater, S. Noble, and D. Scheinost. Power and reproducibility in the external validation of brain-phenotype predictions. *Nature Human Behaviour*, 8(10):2018–2033, 2024.

C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

S. Ryali, T. Chen, K. Supekar, and V. Menon. Estimation of functional connectivity in fmri data using stability selection-based sparse partial correlation with elastic net penalty. *NeuroImage*, 59(4):3852–3861, 2012.

A. Sadikov, H. L. Choi, J. Xiao, L. T. Cai, and P. Mukherjee. Mapping the microstructure of human cerebral cortex in vivo with diffusion mri. *Communications Biology*, 8(1):1088, 2025.

Y. Saglam, A. Oz, G. Yildiz, C. Ermis, O. A. Kargin, S. Arslan, and G. Karacetin. Can diffusion tensor imaging have a diagnostic utility to differentiate early-onset forms of bipolar disorder and schizophrenia: A neuroimaging study with explainable machine learning algorithms. *Psychiatry Research: Neuroimaging*, 335:111696, 2023.

A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, and B. T. Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral cortex*, 28(9):3095–3114, 2018.

K. G. Schilling, K. Ramadass, V. Sairanen, M. E. Kim, F. Rheault, N. Newlin, T. Nguyen, L. Barquero, M. D'archangel, C. Gao, et al. Head motion in diffusion

magnetic resonance imaging: Quantification, mitigation, and structural associations in large, cross-sectional datasets across the lifespan. *Human Brain Mapping*, 46(3):e70143, 2025.

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

B. Serio, M. D. Hettwer, L. Wiersch, G. Bignardi, J. Sacher, S. Weis, S. B. Eickhoff, and S. L. Valk. Sex differences in functional cortical organization reflect differences in network topology rather than cortical morphometry. *Nature Communications*, 15(1):7714, 2024.

M. A. Shafto, L. K. Tyler, M. Dixon, J. R. Taylor, J. B. Rowe, R. Cusack, A. J. Calder, W. D. Marslen-Wilson, J. Duncan, T. Dalgleish, et al. The cambridge centre for ageing and neuroscience (cam-can) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. *BMC neurology*, 14(1): 204, 2014.

S. Shanmugan, J. Seidlitz, Z. Cui, A. Adebimpe, D. S. Bassett, M. A. Bertolero, C. Davatzikos, D. A. Fair, R. E. Gur, R. C. Gur, et al. Sex differences in the functional topography of association networks in youth. *Proceedings of the National Academy of Sciences*, 119(33):e2110416119, 2022.

R. Smith. Essay review: Origins of neuroscience: Nineteenth-century origins of neuroscientific concepts, medicine, mind and the double brain: A study in nineteenth-century thought, 1988.

S. M. Smith, M. Jenkinson, M. W. Woolrich, C. F. Beckmann, T. E. J. Behrens, H. Johansen-Berg, P. R. Bannister, M. De Luca, I. Drobnjak, D. E. Flitney, R. K. Niazy, J. Saunders, J. Vickers, Y. Zhang, N. De Stefano, J. M. Brady, and P. M. Matthews. Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23:S208–S219, Jan. 2004. ISSN 1053-8119.

S. M. Smith, M. Jenkinson, H. Johansen-Berg, D. Rueckert, T. E. Nichols, C. E. Mackay, K. E. Watkins, O. Ciccarelli, M. Z. Cader, P. M. Matthews, et al. Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. *Neuroimage*, 31(4):1487–1505, 2006.

L. Snook, C. Plewes, and C. Beaulieu. Voxel based versus region of interest analysis in diffusion tensor imaging of neurodevelopment. *Neuroimage*, 34(1):243–252, 2007.

C. Spearman. Correlation calculated from faulty data. *British journal of psychology*, 3(3):271, 1910.

O. Sporns. *Networks of the Brain*. MIT press, 2016.

O. Sporns, G. Tononi, and R. Kötter. The human connectome: a structural description of the human brain. *PLoS computational biology*, 1(4):e42, 2005.

E. O. Stejskal and J. E. Tanner. Spin diffusion measurements: spin echoes in the presence of a time-dependent field gradient. *The journal of chemical physics*, 42 (1):288–292, 1965.

C. M. Tax, F. Grussu, E. Kaden, L. Ning, U. Rudrapatna, C. J. Evans, S. St-Jean, A. Leemans, S. Koppers, D. Merhof, et al. Cross-scanner and cross-protocol diffusion mri data harmonisation: A benchmark database and evaluation of algorithms. *NeuroImage*, 195:285–299, 2019.

E. Thibeau-Sutre, S. Collin, N. Burgos, and O. Colliot. Interpretability of machine learning methods applied to neuroimaging. *Machine Learning for Brain Disorders*, pages 655–704, 2023.

S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*, pages 359–380. PMLR, 2019.

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

J.-D. Tournier, R. Smith, D. Raffelt, R. Tabbara, T. Dhollander, M. Pietsch, D. Christiaens, B. Jeurissen, C.-H. Yeh, and A. Connelly. Mrtrix3: A fast, flexible and open software framework for medical image processing and visualisation. *Neuroimage*, 202:116137, 2019.

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*, 80:62–79, 2013.

G. Varoquaux and V. Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ digital medicine*, 5(1): 48, 2022.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

J. Veraart, E. Fieremans, I. O. Jelescu, F. Knoll, and D. S. Novikov. Gibbs ringing in diffusion MRI. *Magnetic Resonance in Medicine*, 76(1):301–314, 2016a. ISSN 1522-2594.

J. Veraart, D. S. Novikov, D. Christiaens, B. Ades-aron, J. Sijbers, and E. Fieremans. Denoising of diffusion MRI using random matrix theory. *NeuroImage*, 142:394–406, Nov. 2016b. ISSN 1053-8119.

J. Veraart, D. S. Novikov, D. Christiaens, B. Ades-Aron, J. Sijbers, and E. Fieremans. Denoising of diffusion mri using random matrix theory. *Neuroimage*, 142:394–406, 2016c.

C. Wachinger, A. Rieckmann, S. Pölsterl, A. D. N. Initiative, et al. Detect and correct bias in multi-site neuroimaging datasets. *Medical Image Analysis*, 67:101879, 2021.

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(2):824–835, 2020.

J. Wen, J. Samper-González, S. Bottani, A. Routier, N. Burgos, T. Jacquemont, S. Fontanella, S. Durrleman, S. Epelbaum, A. Bertrand, et al. Reproducible evaluation of diffusion mri features for automatic classification of patients with alzheimer's disease. *Neuroinformatics*, 19(1):57–78, 2021.

C.-W. Woo, L. J. Chang, M. A. Lindquist, and T. D. Wager. Building better biomarkers: brain models in translational neuroimaging. *Nature neuroscience*, 20(3):365–377, 2017.

J. Wu, J. Li, S. B. Eickhoff, D. Scheinost, and S. Genon. The challenges and prospects of brain-based prediction of behaviour. *Nature Human Behaviour*, 7(8):1255–1264, 2023.

B. T. Yeo, F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari, M. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zöllei, J. R. Polimeni, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*, 2011.

R. Yuste. From the neuron doctrine to neural networks. *Nature reviews neuroscience*, 16(8):487–497, 2015.

S. A. Zergeroğlu and E. Nalçacı. Santiago ramon y cajal and neuron doctrine. *Turkish Journal of Neurology*, 21(3):081–084, 2015.

H. Zhang, T. Schneider, C. A. Wheeler-Kingshott, and D. C. Alexander. Noddi: practical in vivo neurite orientation dispersion and density imaging of the human brain. *Neuroimage*, 61(4):1000–1016, 2012.

B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

M. E. Zimmerman, A. M. Brickman, R. H. Paul, S. M. Grieve, D. F. Tate, J. Gunstad, R. A. Cohen, M. S. Aloia, L. M. Williams, C. R. Clark, et al. The relationship between frontal gray matter volume and cognition varies across the healthy adult lifespan. *The American journal of geriatric psychiatry*, 14(10):823–833, 2006.