Brain MRI Graphical Networks and Subgraphs: Developing a novel methodology for identifying small-scale underlying patterns responsible for large-scale functional differences between brains

Lindsay Wright

A Thesis

in

The Department

of

Physics

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Science (Physics) at

Concordia University

Montréal, Québec, Canada

August 2024

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Lindsay Wright**

Entitled: **Brain MRI Graphical Networks and Subgraphs: Developing a novel methodology for identifying small-scale underlying patterns responsible for large-scale functional differences between brains**

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Physics)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Claudine Gauthier

_____ External Examiner
Dr. Yiming Xiao

_____ Examiner
Dr. Mariana Frank

_____ Supervisor
Dr. Claudine Gauthier

_____ Co-supervisor
Dr. Rachael (Ré) Mansbach

Approved by _____
Dr. Valter Zazubovits, Chair
Department of Physics

_____ 2024

_____
Dr. Pascale Sicotte, Dean
Faculty of Arts and Science

Abstract

Brain MRI Graphical Networks and Subgraphs:

Developing a novel methodology for identifying small-scale underlying patterns responsible for large-scale functional differences between brains

Lindsay Wright

Magnetic resonance imaging (MRI) provides an unprecedented ability to investigate brain health and function. However, the high cost and high variability in the population limits its use in understanding complex diseases, which requires new methodologies. As such, the focus of this project was to define and identify a novel reorganisation of long Covid brain MRI data into network graphs and subgraphs based on functional, rather than spatial, connections between voxels. We define a *physiological connectome*: a graph in which the nodes are voxels, MRI metrics are node attributes, and edges are formed according to physiological similarity. From these graphs we define local neighborhood subgraphs containing each voxel's set of nearest neighbours, which we examine for their properties. MRI features most strongly associated with high connectivity included low values of axial diffusivity (AD), mean diffusivity (MD), radial diffusivity (RD), and isotropic volume fraction (ISOVF); and mean values of intracellular volume fraction (ICVF). Orientation dispersion index (ODI) and fractional anisotropy (FA) values were highly variable. The general data trends outside of highly connected voxels include an inverse relationship between FA and ODI, RD and FA, AD and ODI, ICVF and RD, and ICVF and MD; a positive correlation between AD and MD, ODI and RD, and RD and MD. RD and AD, and ISOVF and ICVF do not demonstrate a clear trend. In future, these subgraphs will form the basis for a generalisable data augmentation and analysis method, to identify underlying patterns responsible for large-scale functional differences between brains.

Acknowledgments

I would like to thank my supervisors for their tireless patience with me, my never-ending questions, and the many unforeseen speed bumps in the process of this thesis and my life. They have been kind and supportive, mentors both in academe and outside of it. Thank you for the opportunity to work with your labs, and to learn from your experience and expertise.

Thank you to my committee members, Drs Frank and Xiao for your interest in my work and your input into its evolution.

A huge thank you to my partner, Luca. Without your attentive support, I would not have made it through this. Thanks also to William. My degree might have been easier had you not entered my life, but it was so much richer an experience to have you with me through it. I love you both so much.

I am grateful for the support of my family and friends, and all the time they spent listening while I practiced my presentations or explained my latest figure to them. I am also very thankful for my labmates. They have taught me so much, and I had a lot of fun with them while I learned.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AD – axial diffusion

ASL – arterial spin labelling

ATT - arterial blood transit time

BOLD - blood oxygen level dependent

CÉRDNT - Comité d'éthique de la recherche et du développement des nouvelles technologies

CBF – cerebral blood flow

CO2 – carbon dioxide

CSF – cerebrospinal fluid

CVR - cerebrovascular reactivity

CT - Computed Tomography

DL – deep learning

DTI – diffusion tensor imaging

DWI – diffusion-weighted imaging

FA – fractional anisotropy

FES - free energy surfaces

FID – free induction decay

GAN - generative adversarial network

GM – grey matter

GRE – gradient recalled echo

ICA - independent component analysis

ICVF – intracellular volume fraction

ISOVF – isotropic volume fraction

k-d tree – k-dimensional treee

MD – mean diffusivity

ML – machine learning

MRI – magnetic resonance imaging

MS – multiple sclerosis

MTI – magnetisation transfer imaging

MTR - magnetic transfer ratio

MTsat - magnetic transfer saturation

ND – neurite density

NODDI - neurite orientation dispersion and density imaging

O2 - oxygen

ODI – orientation dispersion index

OEF – oxygen extraction fraction

PCA - principal component analysis

pCASL - pseudocontinuous arterial spin labelling

PDw - proton density-weighted images

PET - positron emission tomography

PLD – post-labelling delay

QSM - quantitative susceptibility mapping

RD – radial diffusivity

RF – radiofrequency

SMCs - smooth muscle cells

WM – white matter

1. Introduction

Meaningful quantification and modeling of living systems poses many challenges owing, in part, to biological variability and small datasets, a problem that is never more acute than in medical imaging of the brain. As large-scale studies involving imaging are costly, it is not uncommon to reach experimental conclusions from a limited sample size. Additionally, in rare diseases where there is a small pool of patients from which to recruit participants, it may only be possible to acquire a small dataset, limiting insights into the pathophysiology of these diseases. To address these challenges, one possible avenue is to develop a new methodology for identifying underlying patterns responsible for large-scale functional differences between brains. In this research, the first step in creating this methodology is to optimize downsampling of MRI data for subsequent analysis. This was done by creating clusters of physiologically similar voxels (i.e. voxels with similar combinations of imaging parameters), organised as network graphs.

To develop this technique, the population selected for study is women who have experienced long Covid. As is widely known at this time, Covid-19 is a viral illness affecting multiple organ systems including the respiratory, cardiovascular and neurologic systems (Raveendran et al., 2021). A subset of acute Covid patients, estimated at as high as 87% of hospitalised patients and 35% of outpatients, go on to develop long Covid (Carfì et al., 2020). Long Covid is described as persisting or new onset symptoms occurring weeks or months after an acute infection with Covid. Neurological symptoms may include headache, fatigue, “brain fog”, and loss of taste or smell (Raveendran et al., 2021). Long Covid is a complex process that affects many aspects of brain health and is a relatively new topic of research. Neuroimaging, including magnetic resonance imaging, can lead to a better understanding of the nature of the changes in the brain underlying long Covid (Douaud et al., 2022; Zhao et al., 2024); due to this and the timeliness of the subject, the study of a small cohort of patients self-reporting with long Covid makes an ideal case study for this work on methodological development.

Magnetic resonance imaging (MRI) can characterise physiologic functional impairment of brain tissue related to pathological conditions including long Covid. Through varying imaging parameters, MRI can provide insights into several different characteristics of brain tissue including its structure and organisation, (Huisman, 2010), perfusion (Haller et al., 2016), oxygen use (Biondetti et al., 2023), and vascular health (Liu et al., 2019). Each of these metrics can offer further understanding of the changes in the brain that may be responsible for the sequelae of long Covid. Interpreting these characteristics collectively across individuals in a meaningful way poses a challenge. To this end, organising MRI data in network graphs of labelled voxels can enable helpful comparison between sets of MRI properties.

A network graph is a framework for data organisation consisting of nodes, representing data points, interconnected by edges, representing connections between data (Betzel and

Bassett, 2017). In this thesis, we define a *physiological connectome* in the following manner, using the concept: a graph in which the nodes are voxels, MRI metrics are node attributes, and edges are formed according to the similarity between their physiological properties. This thesis tackles the challenge of organizing the data into physiologically meaningful subgraphs that group together voxels with a similar set of properties. Brain MRIs contain hundreds of thousands of voxels; as such, clustering helps divide data into manageable subsets. These clusters are based on network graphs, where the strength of the relationship between voxels is described by the network edges.

The small size of the dataset limits the strength of the conclusions that can be drawn using traditional analysis methods. One solution to address this limitation is to employ data augmentation (Pinaya et al., 2022). Data augmentation amplifies the important trends in the data to separate results from noise. Generative deep learning (DL) can be used to create synthetic data based on the features of the data it is trained on. For a DL model to learn, it requires large amounts of data to serve as examples. As such, the model cannot be trained on whole brain MRI data as there are insufficient samples. In future work, using the results of this thesis, we will train the DL model on the large number of subgraphs and reassemble the generated synthetic output into reconstructed brains mimicking the brain physiological connectome. Finally, these 'hypothetical brains' can then be analysed for significant trends in the data when compared to control which would suggest neurologic changes associated with long Covid. In identifying regions of disfunction in brains exhibiting long Covid, this work will demonstrate validity of this novel analytical approach for future applications.

2. Background

2.1 Covid

The ongoing COVID-19 (Covid) global pandemic began in 2019, causing more than 750 million cases and more than 7 million deaths worldwide as estimated by the WHO. As is widely known at this time, Covid is a viral illness affecting multiple organ systems including the respiratory, gastrointestinal, cardiovascular, and neurologic systems (Raveendran et al., 2021). Symptoms include cough, sore throat, fever, dyspnea, myalgia, joint pain, fatigue, headache, anosmia, dysgeusia, nausea, and chest pain (Carfì et al., 2020). Covid is well-known for its life-threatening respiratory complications. Less attention has been dedicated to the impact of Covid infections on brain health, though it has been associated with several neurological pathologies including stroke and encephalopathies (inflammation of the brain and spinal cord) (Paterson et al., 2020).

While acute Covid continues to be a public health concern, research as early as August 2020 reported ongoing symptoms after the acute phase of covid. A subset of acute Covid patients, estimated at as high as 35% of outpatients and 87% (with 55% reporting three or more symptoms) of hospitalised patients, go on to develop long Covid (Carfì et al., 2020), with risk factors including increasing age, greater than 5 symptoms in the acute phase of Covid, and being female, with the female to male ratio of long Covid being approximately 2:1. Long Covid is described as persisting or new onset symptoms occurring weeks or months after an acute infection with Covid (Raveendran et al., 2021). The symptoms commonly result from disfunction of several organ systems including the nervous system. Of particular concern is the significant effects on brain health. Reported neurological symptoms include headache, profound fatigue, “brain fog”, tremor, problems with attention and concentration, dysfunction in peripheral nerves, and loss of taste or smell (Raveendran et al., 2021).

The mechanism underlying long Covid is not certain; however, persistently elevated inflammatory markers suggest chronic inflammation may be involved (Raveendran et al., 2021). Other reasons for residual, persisting, or new onset symptoms after the acute phase of Covid could include organ damage and organ damage severity, resilience of each organ system, immunological response, autoimmune involvement, the rare occurrence of persisting Covid viral illness, complications from Covid, complications from co-morbidities or medications, deconditioning, and post-traumatic stress (Raveendran et al., 2021).

Recent work has shown that covid and long covid impact multiple aspects of brain structure, function and vascular and metabolic health (Ajčević et al., 2024; Giunta et al., 2024; Guillén et al., 2024; Horowitz et al., 2024; Hosp et al., 2024; Zhao et al., 2024), often for a long period after the disease. Furthermore, long covid has been associated with autonomic

dysfunction (Giunta et al., 2024) which may exacerbate other aspects of brain health dysfunctions. However, while the literature is now starting to provide insight into many of the brain health aspects that are affected by the disease, an integrated understanding is lacking. Because these patients are highly impacted in their daily lives, sample sizes remain limited, precluding multivariate and integrative approaches to studying the disease. To address some of these shortcomings of the currently available data, physiologically informed data augmentation would be instrumental in providing clinically relevant datasets for the study of the multi-faceted aspects of this disease. This project seeks to address this challenge.

2.2 Magnetic Resonance Imaging

2.2.1 General

Magnetic resonance imaging (MRI) is a non-invasive tool for examining internal structures (National Institute of Biomedical Imaging and Bioengineering, n.d.). To do this, it uses signals derived from the magnetic properties of hydrogen atoms in water. Hydrogen atoms each have an unpaired electron, giving them a quantum property called intrinsic angular momentum or spin which is responsible for the fourth quantum number $m_s = {}^{\pm 1}/_{2}$. This spin gives them a magnetic moment. In the absence of a magnetic field, the net magnetisation of a pool of hydrogens will be zero, as the spins will be randomly oriented. Body tissues such as the brain, the heart, and blood have high water content. Since water contains an abundance of hydrogens, the body essentially comprises many small bar magnets.

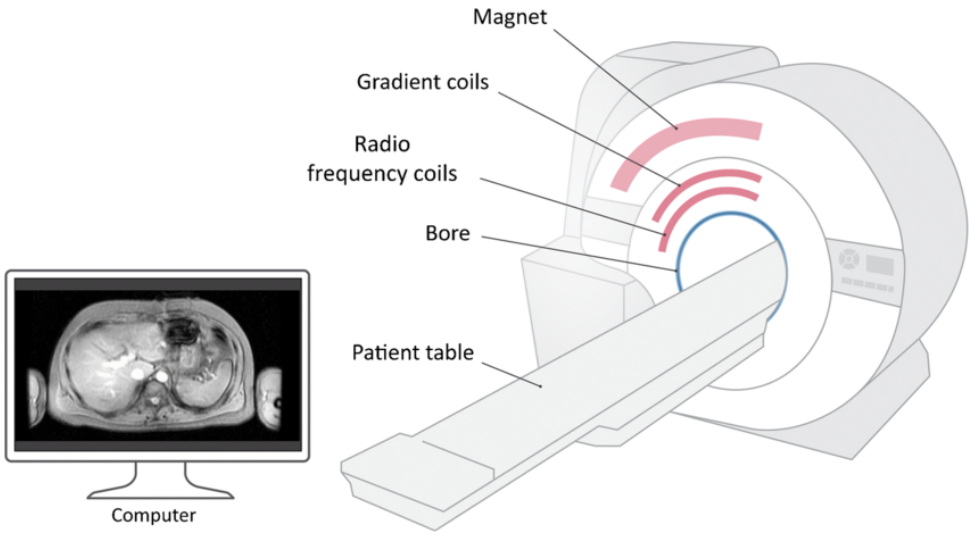

Figure 1 general schematic of an MRI. Within the coils, the magnetic field aligns with the axis of the patient table. (Serai, 2022)

MRIs are used to examine and study in-vivo organs such as the brain. For example, MR images can demonstrate the difference between areas of the brain comprising mostly of white matter (WM), where high concentrations of myelin-covered axons make the tissue appear white, and areas comprising mostly of grey matter, where the tissue is largely made up of neuron cell bodies (soma) and non-heavily myelinated axons (Mercadante and Tadi, 2023). An MRI machine generates a mostly uniform magnetic field (called B_0) in a configuration similar to a solenoid: parallel to the axis of the coil, through the bore of the machine (fig. 1). Exposure to this field causes a net shift in the magnetic moments of the hydrogens in corporeal water such that slightly more align parallel to B_0, in their lower energy state (Li et al., 2020) and resulting in a small net magnetisation (Jones et al., 2021). The direction of these moments is then disrupted by a radiofrequency (RF) field (called B_1) which reorients this spin to a higher energy state into the transverse plane. As the RF pulse ends, the hydrogen flips back to re-align with magnetic field of MR. The realignment releases electromagnetic energy which is detected and analysed. The time it takes for the spin to realign is tissue-dependent.

Important measurable quantities in an MRI experiment include T1 relaxation, T2 relaxation, and free induction decay. The process of realignment in the longitudinal direction is called T1 relaxation and follows an exponential trend with T1 as its first-order time constant (fig. 2 a., longitudinal magnetisation) (Elster, n.d. a). The realignment time constant T1 is a property of different materials and can be used to characterise tissues on imaging (T1-weighted imaging). It arises from loss of energy or heat to the lattice. T2 relaxation is a concept similar to T1 except it describes the relaxation of the transverse magnetisation components (the plane perpendicular to B_0) through spin-spin interactions (Elster, n.d. b). It can be used to generate T2-weighted images. Since there is no standing magnetic field in the perpendicular plane, any transverse magnetisation relaxes into incoherence (fig. 2 b., transverse magnetisation). T2 is the ideal time constant for this behaviour arising from only natural atomic and molecular interactions. A second constant called T2* is the observed transverse relaxation time constant which reflects dephasing due to magnetic field inhomogeneities (Elster, n.d. c). T2* is always less than or equal to T2. (Liang and Lauterbur, 2000). Free induction decay (FID) is a decaying sinusoidal signal that occurs post-RF excitation (Elster, n.d. d). It decays with time constant T2* and can be manipulated to generate useful imaging signals (see the description of gradient recalled echo in section 2.2.5). The ability to capture these different relaxation constants depends on the imaging parameters used. These parameters include, first, the repetition time, which determines the time between acquisition of two images; and second, the echo time, which determines the time between tipping the magnetization into the transverse plane and the acquisition of k-space.

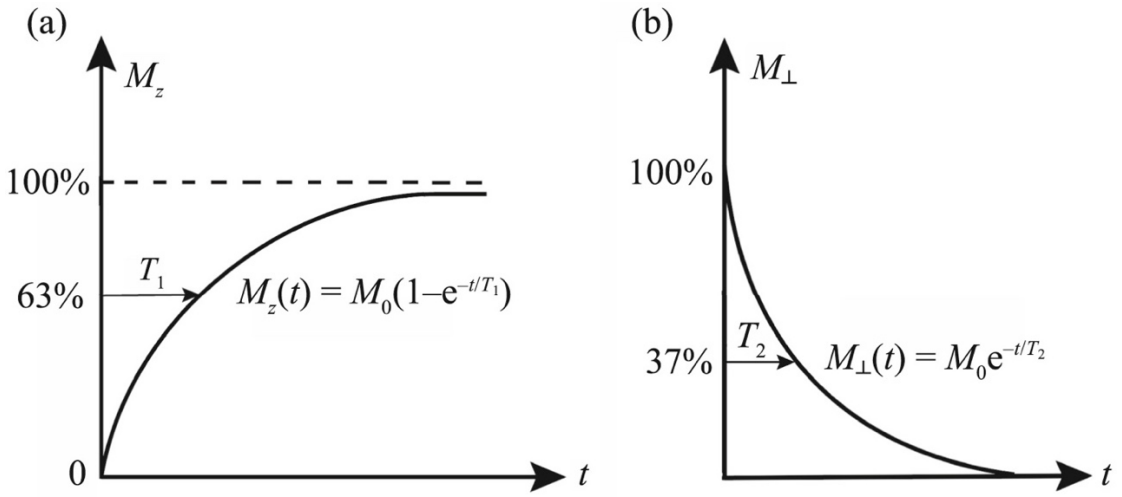

Figure 2: The relaxation process of the net magnetisation M_0 for (a) the longitudinal or M_z and (b) transverse or $M_\perp$ components, as calculated using the T1 and T2 relaxation times, respectively. (Xu et al., 2019)

In addition to general temporal information, spatial information can be extracted from an MRI, through the use of *gradients* for spatial encoding. Gradient fields can be applied during RF excitation by gradient coils (fig. 1); this alters the magnetic field in a known, predetermined pattern, leading to a position-dependent proton resonance frequency. This spatially encoded distortion of the signal allows for its localisation in three-dimensional space. The signal is collected in a space called k-space, in which the intensity of each element of the matrix being filled reflects the contribution of a given spatial frequency. The signal can then be converted to image space to obtain meaningful metrics using Fourier transforms, processing techniques, and mathematical models to construct two- or three-dimensional MR images that reflect a given property for each voxel (Hidalgo-Tobon, 2010).

In summary, an MRI sequence is a specific set of gradient and RF pulses which excite hydrogen electrons. As the electrons relax back into their low energy state, they release energy which is recorded to yield a signal. The signal is then processed according to an appropriate model into meaningful metrics which give information about tissue properties at every voxel. Different properties are obtained by varying the amount of time left for the three different relaxation types to evolve, as well as additional manipulations of the net magnetization (e.g. diffusion direction encoding).

MRI has many advantages over alternative neuroimaging methods, such as CT or positron emission tomography (PET). As MRI does not involve ionising radiation in image acquisition, it avoids the risks associated with radiation which include an increased risk of cancer, especially in younger populations (Lin, 2010). Many MRI sequences are non-invasive; it does not require the administration of exogenous contrast agents, the use of which can be contraindicated in certain populations such as patients with renal dysfunction, children, or contrast allergies (Haller et al., 2016). Due to its safety profile, MRI is more repeatable which can facilitate longitudinal imaging.

2.2.2 Diffusion-Weighted Imaging

Diffusion-weighted imaging (DWI) is an application of MRI that makes use of the molecular diffusion of water. This translational movement of particles in liquids or gasses, also known as Brownian motion, appears random owing to redirection of their trajectories by collisions with other particles (Brown, 1828; Einstein, 1905; D. Le Bihan et al., 2001). The intrinsic physical property of diffusion is not disrupted by MR effect or magnetic fields as water is not charged (Le Bihan et al., 2001). In DWI, researchers track this motion and consider how it is restricted in order to map structural boundaries such as cellular membranes, axons, and macromolecules in the brain (D. Le Bihan et al., 2001). From the voxel-wise displacement distribution of water, the geometry of the brain's microstructure can be characterised. A brain region can be said to have a high degree of anisotropy if the amount of diffusion of water molecules varies greatly between axes in three-dimensional space; in other words, if the local movement of water is highly directional (fig. 3 B). Coherently organised tracts of WM, having bundles of parallel axons, show anisotropic diffusion as the water preferentially diffuses along the length of the fibre bundles. The opposite of anisotropy, isotropy, occurs in areas of the brain with less directionality to its organisation (fig. 3 A). For example, the diffusion of cerebrospinal fluid (CSF) in the ventricles is less constrained, leading to highly isotropic, or non-directional, diffusion.

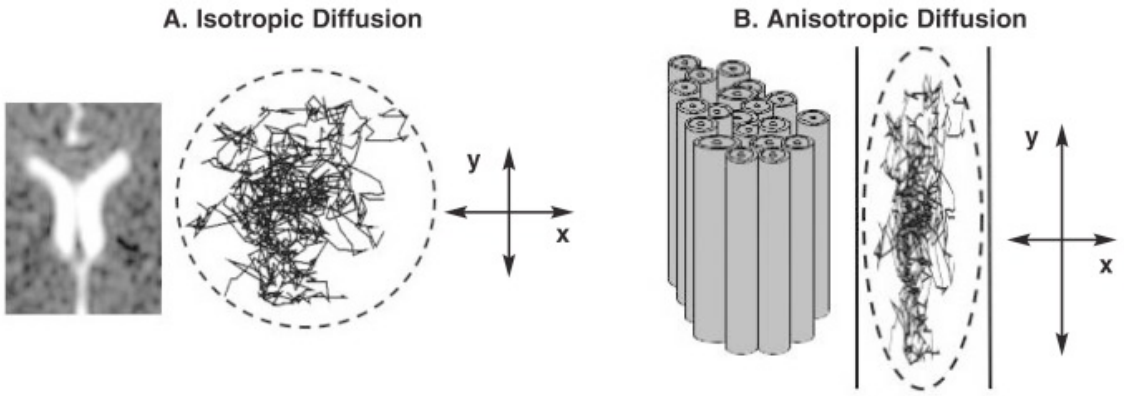

Figure 3: (A) Isotropic diffusion of CSF in ventricles, and (B) Anisotropic diffusion in a parallel bundle of axons. (Rosenbloom et al., 2003).

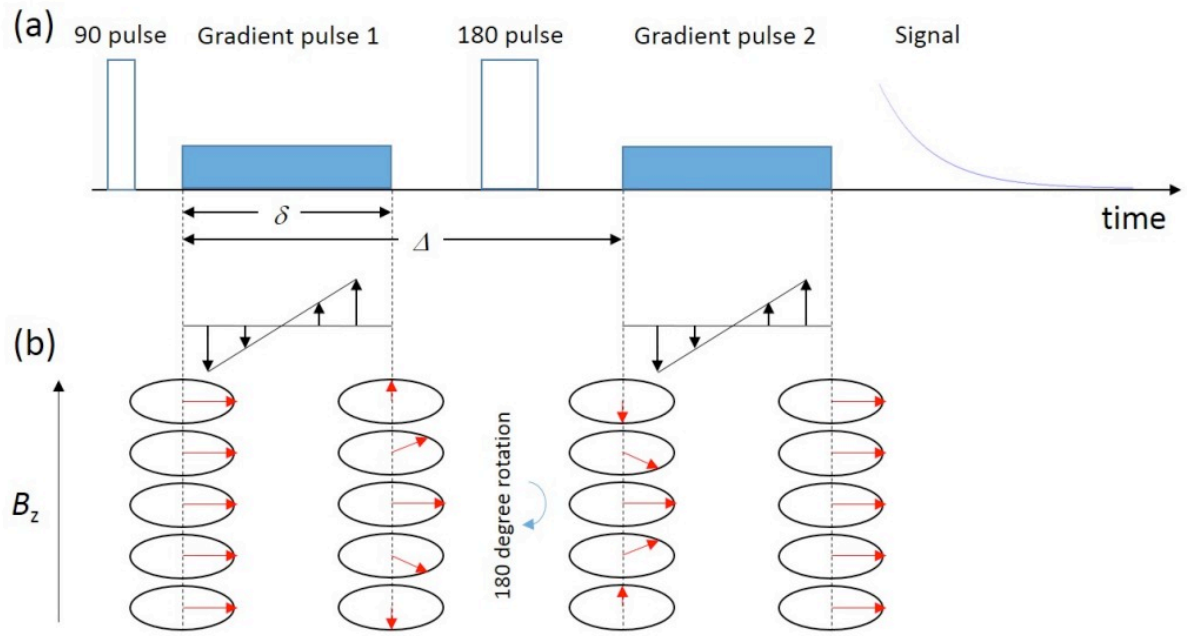

Figure 4: (a) Pulsed gradient spin echo (PGSE) sequence schematic. (b) The first gradient dephases spins and the second gradient rephases stationary spins, since there is a 180° RF pulse in between the gradients. (Magritek, 2016).

To acquire DWI images, a pulsed gradient spin echo sequence is used (fig. 4 a). A spin-echo sequence consists of two RF pulses of 90° and 180°. The 90° pulse reorients spins to a higher energy state. A gradient pulse causes the spins of diffusing water to dephase, then the 180° RF pulse is applied, followed by a second gradient pulse (Alexander et al., 1997; Reese et al., 2003; Stejskal & Tanner, 1965). Since the second gradient is after the 180° pulse, stationary spins are rephased (fig. 4 b) while moving spins (e.g. in diffusing water) are not. This spin dephasing causes a transverse relaxation signal attenuation, allowing for the measurement of diffusion in DWI. The DWI images can then be processed into specific metrics using different models. Two common models include the diffusion tensor imaging (DTI) model and the neurite orientation dispersion and density imaging (NODDI) model, each with their own assumptions, limitations and advantages.

DWI can be processed according to the conceptually simple DTI model to provide fractional anisotropy (FA), mean diffusivity (MD), axial diffusivity (AD), and radial diffusivities (RD). In situations where diffusion is not uniform in all directions (anisotropic, fig. 3 A), a single scalar value is insufficient to fully characterise it (Le Bihan et al., 2001). Within each voxel, DTI considers diffusion in terms of a scalar mean rate in each of the three axes of three-dimensional space (anterior-posterior, superior-inferior, and left-right).

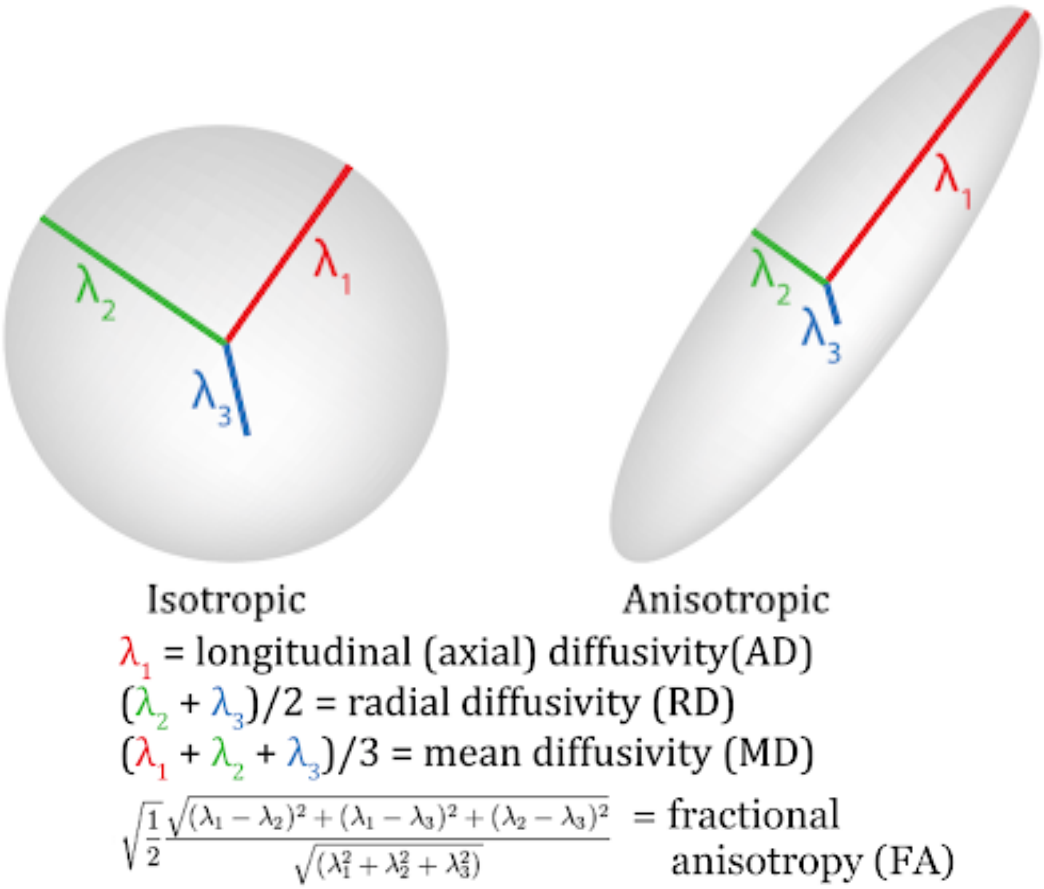

Figure 5: eigenvalues of diffusion, λ_1, λ_2, and λ_3. An isotropic property does not vary in magnitude between different directions of measurement; the measurement of an anisotropic property is direction dependent. AD, RD, MD, and FA are eigenvalue-dependent properties of diffusivity. (Tromp, 2015)

From these eigenvalues (fig. 5), several measures can be determined. Axial diffusivity (AD) is the diffusivity along the principal direction. Radial diffusivity (RD) is diffusivity in the plane perpendicular to the principal direction. Mean diffusivity (MD) describes the total average displacement across all directions and is rotationally invariant. Fractional anisotropy (FA) is a measure of spherical asymmetry of diffusion. It is the most used measure of anisotropy due to its high signal-to-noise and contrast to noise ratio (Alexander et al., 2000; Hasan et al., 2004; Papadakis et al., 1999; Sorensen et al., 1999). FA ranges from 0, describing isotropic diffusion, up to 1. Values above 0.7 indicate highly coherent fibres, such as in the corpus callosum (fig. 6). Values above 0.2 suggest WM, and values lower than 0.2 suggests GM and CSF.

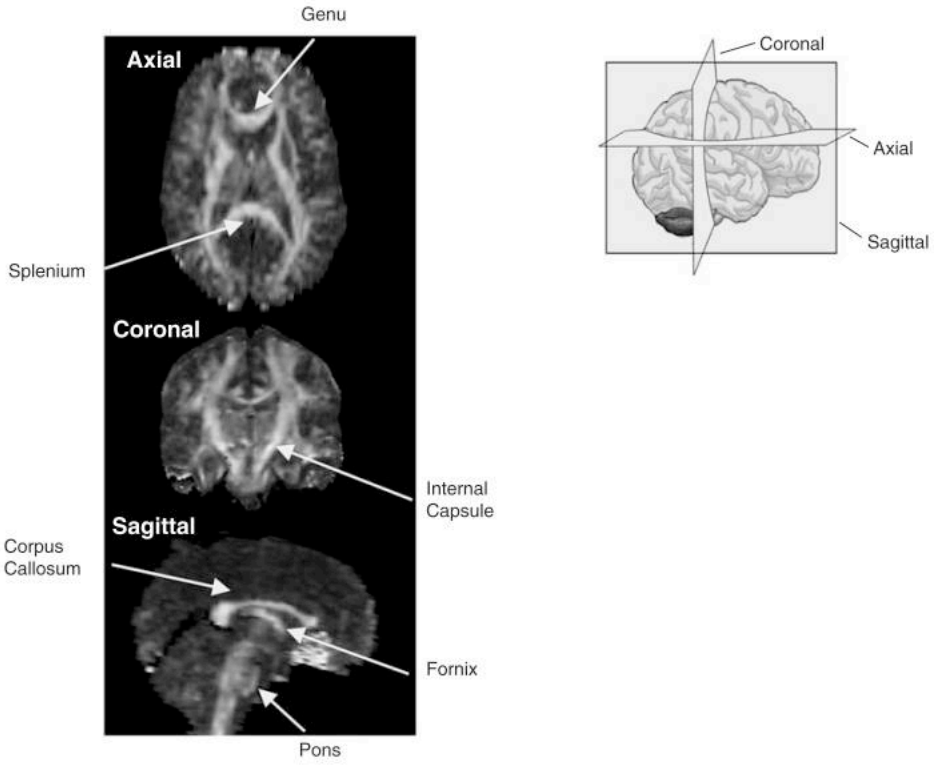

Figure 6: FA images of a 31-year-old healthy man's brain. Bright regions showing high FA correspond to WM tracts. Dark regions show isotropy. (Rosenbloom et al., 2003).

While DTI is an intuitive model, it is a gross measure which can lose some information due to the low resolution and simplicity of the model (Zhang et al, 2012). DTI essentially assumes a single compartment; that is, it assumes diffusion is occurring in an extracellular compartment only. The DTI model posits that each voxel has a single main orientation of diffusion. This is most appropriate in regions such as the corpus callosum, that have very large axon bundles with similar orientations. However, this assumption does not allow for identification of overlapping tracts, or non-parallel bundles of axons. In the case of overlapping tracts, diffusion would be highly directional in the intracellular compartment along the length of the tracts. As DTI only considers average diffusion rate per Cartesian axis, the magnitudes of the diffusion vectors of the overlapping tracts are lost as a simple, lower average diffusion rate (Descoteaux, 2015). In the case of crossing fibers, because there are competing diffusion directions that both enhance AD and RD, the DTI model results in an overly isotropic diffusion tensor that underestimates the contribution of fiber bundles. This can then be erroneously interpreted as a lower "WM integrity". DTI's simplicity affords an advantage when compared to other more complex models in requiring much shorter MRI acquisition durations. This advantage lends itself well to clinical applications in acute processes such as stroke, or in populations which may not tolerate longer scan times such as pediatric or dementia patients. However, it is limited in its interpretability.

The NODDI model is more complex and has more degrees of freedom than DTI. NODDI assumes a three-compartment schema, that is, that diffusion is constrained in the

intracellular space or compartment (intracellular volume fraction, ICVF), hindered in the extracellular space (isotropic volume fraction, ISOVF), or isotropic in CSF, which allows for different rates of diffusion between each compartment (Zhang et al., 2012). The orientation distribution of the intracellular space is modelled as restricted diffusion along axons of zero radius amounting to a collection of sticks, which make up a Watson distribution. The Watson distribution estimates the dispersion around a mean orientation at each voxel (Zhang et al., 2011, 2012). For the extracellular compartment, the diffusion is modelled as hindered diffusion in a Gaussian anisotropic process. It considers two diffusivities, both parallel (axial) and perpendicular (radial) to the orientation of the intrinsic intracellular diffusivity (Zhang et al., 2012). These values are determined from the measured neurite density (ND) while considering the effect of orientation dispersion (Jespersen et al., 2012). The third compartment, CSF, is modelled as isotropic Gaussian diffusion. The mathematical details behind this modeling is outside the scope of this project, however, the results of the NODDI model are metrics ND, and fibre orientation dispersion, also called orientation dispersion index (ODI). ND is an estimation of the volume of axons and dendrites in a voxel. ODI is a measure from 0 to 1, where 0 indicates perfect anisotropy and 1 indicates perfect isotropy.

NODDI's higher complexity allows for disentangling regions of crossing fibres from those having truly low neurite density. For example, in WM tracts where axons and therefore diffusion is largely aligned, such as in the corpus callosum, ND is high, and ODI is low and corresponds to high FA (fig. 7) (Zhang et al., 2012). In locations with lower alignment of axons (e.g. fanning and crossing fibres) however, ODI values are higher while ND values remain similarly high. This is expected, as WM has little variation in neurite density. In GM, ODI is higher and ND lower than in WM.

By considering these two metrics together, it is possible to differentiate between changes in WM fibre density and changes in fibre organisation (Zhang et al., 2012). The cost of the higher degrees of freedom of the NODDI model is that longer scan times are required to acquire more images, which may not be well tolerated by certain populations. Additionally, the more complex mathematical approach involved requires longer data processing times.

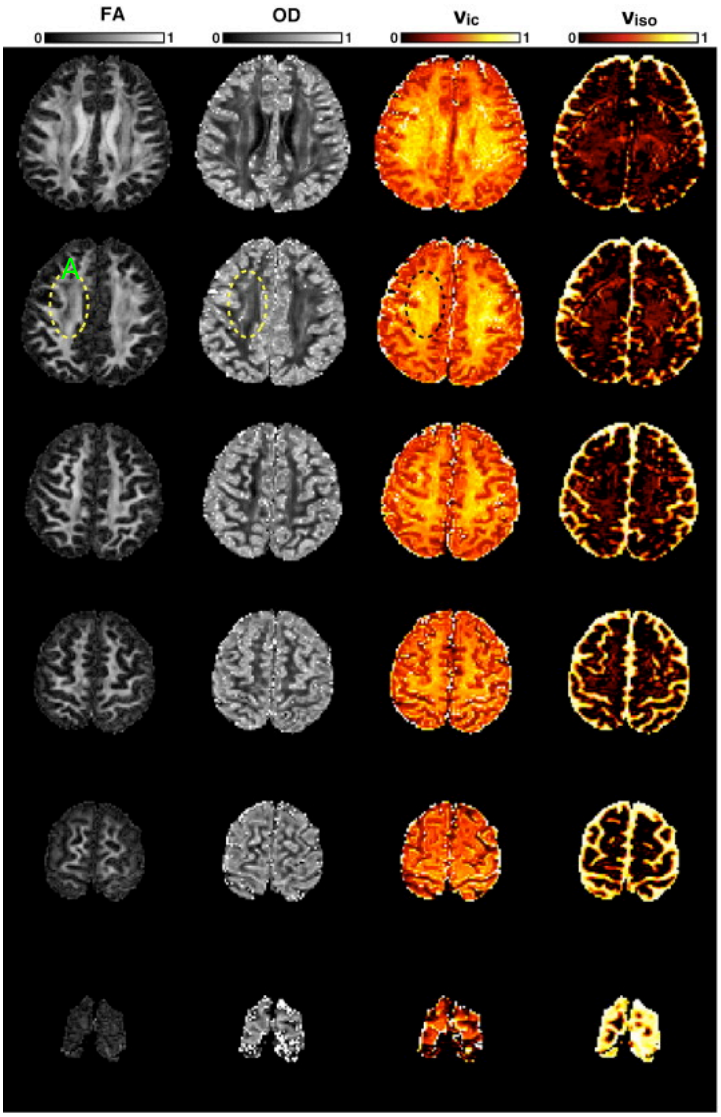

Figure 7: FA, orientation dispersion index OD, intra-cellular volume fraction v_{ic}, and isotropic (CSF) volume fraction v_{iso}. v_{ic} is mostly unchanging within white matter while ODI and FA have significant regional variations that are inversely correlated with each other. Row A: the dashed regions show the centrum semiovale where there is significant crossing and fanning of WM tracts. Here, lower FA values correspond to higher ODI values, with high v_{ic} (high neurite volume). Modified from Zhang et al., 2012.

2.2.3 Magnetisation Transfer Imaging (MTI)

Magnetisation transfer imaging (MTI) is a sequence that provides an estimate of the myelin content of the brain. Myelin is the characteristic lipid-heavy (70-85% lipids, 15-30% protein) insulation on axons which gives WM its colour (Morell and Quarles, 1999). The macromolecules that make up myelin are not directly visible through MRI as their bound

protons have highly restricted motion (e.g. rotation, vibration, and translation), so the signal decays too quickly for recording (Olsson, Andersen, et. al., 2020). To image myelin, MTI uses the binary spin-bath model, also known as the two pools model, as its theoretical basis (de Boer, 1995). The first "pool" is the group of bound protons, consisting of the protons of the macromolecules as well as those of the "hydration layer", the surrounding water molecules that are hydrogen-bonded to the macromolecules. This pool is referred to as the bound pool. The second "pool" referred to is the group of protons attached to free water, which allows those protons to be relatively unrestricted in their motion and have very long relaxation times compatible with imaging. This is called the free pool. The bound pool can be excited by a wide range of radiofrequencies while the free pool has a very narrow range of exciting frequencies.

In MTI, the approach is to first apply an MT pulse to excite the bound pool that then transfers magnetisation to the free pool. Shortly after the MT pulse, a second RF pulse targeting the free pool is applied. The resulting post-MT signal from the free pool (called MT-on) is measured. The MT-on signal is compared to an MT-off control obtained using the free pool pulse without MT pulse to determine the extent of transfer. To selectively saturate the bound pool, the frequency of the MT pulse is chosen to be within the bandwidth of the bound pool frequencies, which is separated from the free pool frequencies by several orders of magnitude (Henkelman, Stanisz and Graham, 2001). The resulting change in magnetisation in the bound pool then reorients the nearby free water protons through a process called cross relaxation. Following the transfer of magnetisation, the signal derived from the effect of a subsequent RF pulse on the free pool shows a signal attenuation of the free water pool proportional to the size and density of the bound pool (Edzes and Samulski, 1977). Since the net magnetisation of the bound pool is proportional to myelin content, the MT-related signal attenuation of the free pool is proportional to the amount of myelin present. In this way, MTI provides an indirect, relative measure of myelin content per voxel (Sled, 2018).

MTI can be processed into different metrics including magnetic transfer ratio (MTR) and magnetic transfer saturation (MTsat). MTR is a simple calculation of the difference between a control reading of the free pool signal (MT-off) and the free pool signal reduction (MT-on), as a ratio to MT-off (fig. 8). While this measure is intuitive, it is not specific. In measuring myelin content, it is sensitive to some confounders such as edema (Thiessen, Zhang et al, 2013). Inflammatory states which occur in some pathological conditions including multiple sclerosis (MS) can cause edema. Edema is an increase in cellular or interstitial fluid. This increase in the volume of the free water pool dilutes the signal attenuation and decreases the MTR. However, myelin loss is also a pathological feature of MS which can also cause a decrease in MTR. In MS, both myelin loss and edema can be signs of disease progression, but MTR can not specify which event is occurring (Levesque et al., 2005). Differentiating these conditions can be important in determining the appropriate course of treatment.

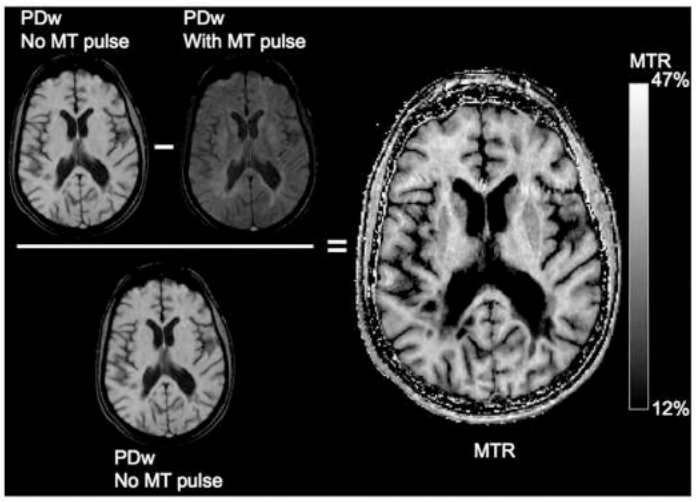

Figure 8: Stepwise demonstration of the calculation of an MTR image from proton density-weighted images (PDw). (Fox et al., 2011)

MTsat is a semi-quantitative measure of the fraction of free water that is saturated (magnetised) by a single MT pulse (Olsson et al, 2021). The mathematical model is more sophisticated than MTR and considers several additional factors which would otherwise distort the true saturation fraction. To make these adjustments, MTsat requires a T1-weighted map to account for relaxation time, a B1 map to account for local variations in flip angles, and both MT-on and MT-off (Olsson et al, 2021). The B_1 map is derived from the observed B_1 field and is a measure of the inhomogeneities of the B_1 field caused by, for example, attenuation of RF amplitude due to tissue conductivity (Helms et al., 2008), and is used for noise corrections in imaging. By adjusting for each of these factors, MTsat is much more specific than MTR. MTsat can measure myelin loss by removing the confound of edema. Additionally, compared to MTR, MTsat shows a greater image contrast and is more sensitive to demyelination and changes in tissue (Longoni et al, 2023).

2.2.4 Arterial Spin Labelling

Arterial spin labelling (ASL) is a family of MRI sequences that creates perfusion-weighted images, i.e. images that are weighted by how much blood is flowing through a particular area. It allows for the generated signal to distinguish between blood flow in vessels and the surrounding, mostly stationary, body tissues (Albert Einstein College of Medicine, 2014). It differs from typical MRI by employing selective application of the RF pulses to specific regions only. The RF is applied at the level of the carotids below the bifurcation for systematic labelling of all blood perfusing the brain. Application of this pulse means that the spins of the hydrogens at that location – including both tissue and blood – invert (become labelled). This labelled signal is then allowed to flow through the large vessels into the brain

until it reaches the capillaries. An image is then acquired and the duration between tagging and imaging is called the post-labelling delay (PLD) (Grade et al., 2015). At the time the image is acquired, the image is therefore a combination of a relaxed signal from the stationary tissue of that region and a perfused-weighted signal from the RF-labelled blood. This technique therefore uses labelled water as an endogenous, intravascular tracer. This perfusion-weighted image cannot be used in isolation since signal attenuation comes both from stationary spin relaxation and labelled water. To obtain a perfusion weighted signal, this first image is subtracted from an image without magnetic tagging. This subtraction image is not quantitative however. To make the perfusion-weighted image quantitative, an additional perfusion calibration image (M0) is needed with a long repetition time to estimate the baseline or fully-recovered magnetisation. This M0 map can be combined with the perfusion-weighted map into a kinetic model to obtain a quantitative measure of perfusion in units of ml blood/100g tissue/min. Kinetic modeling takes into consideration arterial transit time, incomplete spin inversion from the pulse, equilibrium magnetisation of blood, magnetisation decay via the exchange of magnetisation between static tissue and blood flow and magnetisation clearance via perfusion (Buxton et al., 1998). The final, flow-sensitive signal is proportional to the amount of inverted spin that arrives at the tissue in a given voxel (Haller et al., 2016).

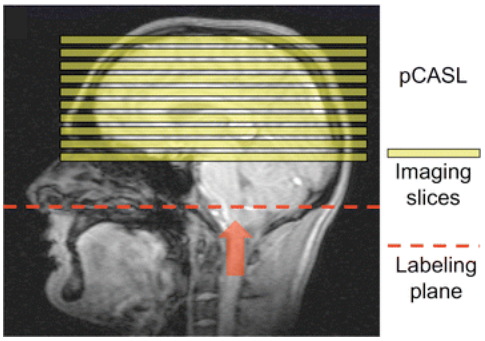

Figure 9: Blood passing through the arteries in the neck is continuously labelled and images are recorded distally. (Haller et al., 2016).

The main limitation of ASL include a low signal-to-noise ratio. As such, ASL typically involves many repetitions and averaging to better discern signal from noise. This repetition has two main downsides: it makes signal acquisition a very slow process, and it causes ASL to be highly sensitive to motion artefacts. The speed can be improved somewhat by coupling ASL with other techniques such as echo-planar imaging to make it faster. The signal-to-noise ratio can also be improved by use of stronger magnetic fields, for example, use of 3T instead of 1.5T MRIs where available.

The pseudocontinuous ASL (pCASL) sequence is a specific implementation of ASL which can be implemented on a standard clinical MRI scanner and has better signal-to-noise ratio than most other ASL sequences (fig. 9) (Grade et al., 2015). pCASL employs a series of very short (~1ms) RF pulses for approximately 1-2s, which is designed to label in a pseudo-steady state manner (Haller et al., 2016). The control images can be obtained from applying

the same frequency and duration of pulses with the pulse phase alternating by 180º for each subsequent pulse (Haller et al., 2016). pCASL also has its drawbacks, which include: the labelling efficiency of pCASL is generally lower than in some other ASL techniques (Haller et al., 2016); the pCASL labelling is sensitive to resonance offset at the labelling plane resulting in a shift; and some phase shift can occur between RF pulses (Grade et al., 2015). However, it is the most commonly used technique as it balances good signal-to-noise while limiting power deposition (W/kg; increases in core body temperature) into tissues.

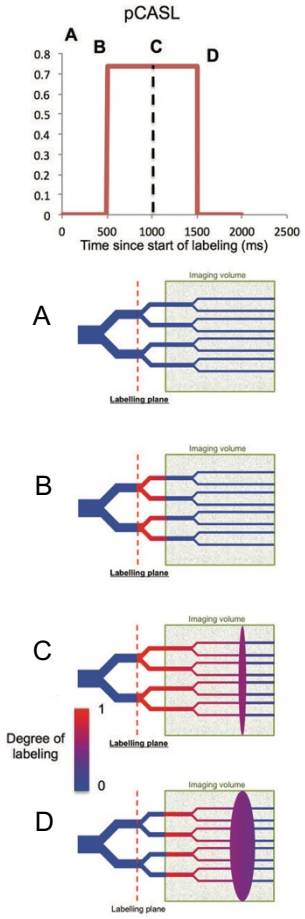

Figure 10: pseudo steady state application of RF pulse in pcASL.
A, t = 0;
B, t = ATT;
C, t = labelling duration;
D, t = ATT + labelling duration. Adapted from Haller et al., 2016.

The flow-sensitive signal can be processed to quantify perfusion and to estimate cerebral blood flow (CBF). CBF is a quantitative, absolute measure in units of millilitres of blood per 100 grams of tissue per minute. CBF is primarily affected by T1 relaxation time and labeling duration, which is the length of time the blood is able to retain its label. Labeling duration is an acquisition parameter that is determined by blood and tissue T1 (Detre et al., 2012). Other variables and imaging parameters can also influence CBF measures, including arterial transit time and post-labeling delay. Arterial blood transit time (ATT) is a function of flow velocity and describes the time required for blood to pass from arteries and into tissue. ATT has been well characterised in healthy subjects but can be affected by pathologies including cerebral vascular disease and stroke (fig. 11). The delay between the end of RF labelling and the beginning of imaging, during which time blood flows from the labelling plane into the brain, is referred to as the PLD. The PLD is of critical importance to imaging. It is chosen to best match with ATT and T1 relaxation time: the labelled blood must have arrived at the tissue while also still being well-labelled to produce a quality signal. If PLD is too short relative to the time taken for the first labelled blood of the bolus to arrive at the imaging volume (ATT, fig. 10 B), the signal will underestimate perfusion. If the PLD is longer than the longest time required for the labelled blood to pass from the labelling plane to the imaging volume (ATT plus the labelling duration, fig. 10 D), the ASL signal will lose its sensitivity when blood and tissue have a similar T1 (e.g. GM); this allows for CBF quantification but may decrease the signal-to-noise ratio (Haller et al., 2016). Both ATT and T1 relaxation time are on the order of seconds, and directly influence the signal strength. Since pathological states can increase ATT, the typical PLD used in healthy individuals may not attain the best results in individuals with vascular disease and it may be beneficial to use multiple PLD techniques to determine the most appropriate parameter. Underestimation of ATT can lead to an underestimation of CBF (Haller et al., 2016). A typical PLD value for healthy adults is 1500ms (Haller et al., 2016).

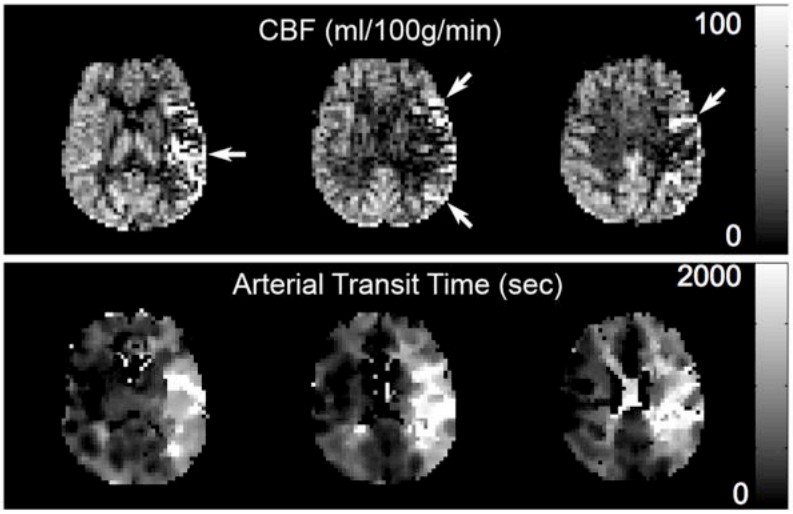

Figure 11: Imaging of left middle cerebral artery (MCA) stroke. Top: ASL CBF showing hyperperfusion (arrows) of left MCA territory. Due to the stroke, the labelled blood is delayed resulting in it being imaged in collateral blood flow vessels. Bottom: ATT shows delay to the left MCA region. (Detre et al., 2012)

A second metric which can be derived from ASL in combination with breathing manipulation is cerebrovascular reactivity (CVR). CVR provides a measure of cerebrovascular reserve, which quantifies the ability of the brain's vasculature to vasodilate in response to a vasodilatory stimulus. Cerebrovascular reserve can serve as an indicator of vascular health of the brain (Liu et al., 2019). This measurement can be obtained using different methods including breath-holding, hyperventilation, carbon dioxide (CO2) inhalation and administration of acetazolamide (Liu et al., 2019). Both breath-holding and hyperventilation require a high degree of cooperation from the participant which may be difficult for certain populations and variable adherence to the experimental protocol may decrease reliability of this method. CO2 inhalation requires gas delivery apparatus, but otherwise is a popular approach to CVR as it is a more controlled stimulus which yields large effect sizes (Liu et al., 2019).

CO2 is a potent vasodilator with both a rapid onset and resolution of effects (Liu et al., 2019). When serum concentrations of CO2 are increased from physiologically normal levels, blood pH decreases via dissociation of CO2 into bicarbonate and hydrogen ions. Both elevated CO2 and acidosis can trigger relaxation of the smooth muscle cells (SMCs) of the vascular endothelium (Liu et al., 2019). Relaxation of SMCs on the cellular level corresponds to a dilation of the blood vessels. In CVR, increasing inhaled CO2 leads to increased circulating CO2 that causes cerebrovascular vasodilation. This vasodilation results in elevated perfusion which can be captured using hemodynamic MRI signals such as ASL, allowing for the measurement of vascular reactivity to CO2 (Liu et al., 2019).

2.2.5 Quantitative Susceptibility Mapping

Susceptibility weighted imaging, the precursor to quantitative susceptibility mapping, uses MRI phase measurements to describe the spatial distribution of magnetic susceptibility in human tissues (Schweser et al., 2016). Magnetic susceptibility is an intrinsic physical property of matter that describes the extent to which a material becomes magnetised when in the presence of an external magnetic field (Duyn, 2013). The internal magnetic field can align in the same or in the opposite direction to the external field, referred to as paramagnetic or diamagnetic, respectively. A third classification, ferromagnetic, refers to very strongly paramagnetic materials that are not compatible with the use of MRI for safety reasons (Schweser et al., 2016).

Paramagnetism is a phenomenon of unpaired electrons causing a magnetic moment in an atom. When exposed to a magnetic field, the moment aligns in the direction of the external field (Duyn and Schenck, 2017). Diamagnetism arises from a particular molecular-level instance of Lenz's law: current induced by a changing external magnetic field is generated such that the magnetic field generated by the induced current opposes the change in the external magnetic field (Schenck, 1996). Essentially, the motion of the conductive electrons of a molecule change to produce a magnetic field which opposes the external field (Schenck, 1996).

With respect to neuroimaging, magnetic susceptibility is mostly affected by water, myelin, iron, and calcium (Schweser et al., 2016). Both myelin (−13 to −34 ppb relative to water, Duyn and Schenck, 2017) and calcium are more diamagnetic than water ($\chi = -9.05$ ppm, Duyn and Schenck, 2017) (Schweser et al., 2016). The susceptibility of iron depends on its form; the magnetic susceptibility of heme-iron depends on its oxygenation state (Bren et al., 2015). In 1936, Linus Pauling and Charles Coryell confirmed earlier findings that oxygenated hemoglobin is diamagnetic while deoxygenated hemoglobin, having four unpaired electrons per iron atom, is paramagnetic ($\chi = -7.9$ ppm, or about 1.2 ppm paramagnetic relative to water, Duyn and Schenck, 2017) (Bren et al., 2015).

To quantify susceptibility, we use a gradient recalled echo (GRE) pulse sequence is used, taking advantage of the FID signal (see section 2.2.1). First, an RF pulse generates an FID signal. Next, a dephasing gradient field alters the resonance frequencies of the tissue, which causes the FID signal to 'squelch', or causes an accelerated dephasing of the FID signal. Finally, a rephasing gradient pulse having the same magnitude but opposite polarity of the first gradient is applied. This reverses the phase shifts of the dephasing gradient, refocusing the spins and creating a gradient echo from the reemerging FID (fig. 12). GRE does not require any specialised hardware, making it an accessible sequence commonly used to susceptibility imaging.

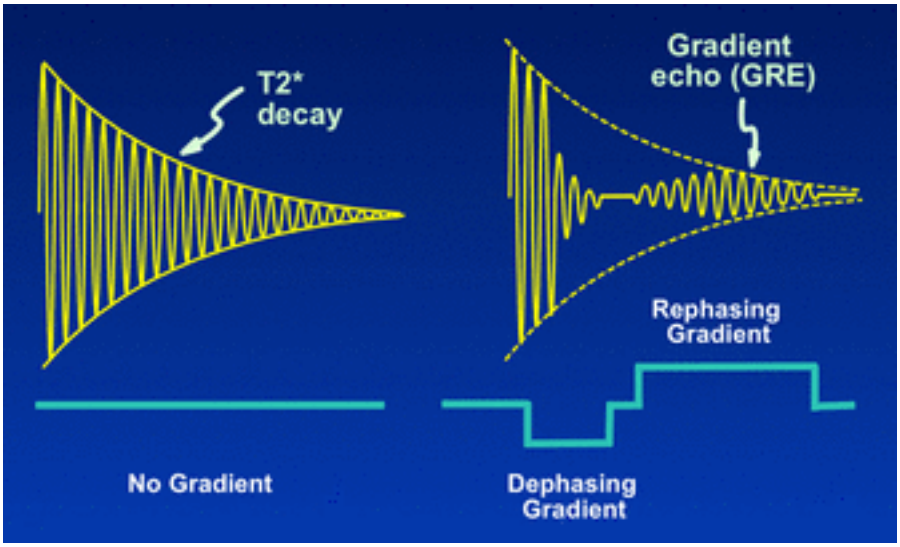

Figure 12: FID with T2* decay in the absence of gradient, and the dephasing of FID followed by the rephasing gradient resulting in a gradient echo. (Elster, n.d. e).

Quantitative Susceptibility Mapping (QSM) is a post-processing technique used on susceptibility weighted imaging that extracts the local magnetic susceptibilities from MRI phase measurements to describe the spatial distribution of magnetic susceptibility (Schweser et al., 2016). The process of QSM is complex, requiring several steps (Schweser et al., 2016 and Haacke et al., 2015). The variations in local magnetic susceptibility, such as from the presence of deoxyhemoglobin in veins, alter the local magnetic field. This local field perturbation can be measured and described by T_2*-weighted phase data, since the T_2* time constant is based on the rate of dephasing due to both spin-spin relaxation (T_2) and to field inhomogeneities (Liang and Lauterbur, 2000). A larger phase value correlates to a higher magnetic susceptibility value (Liang and Lauterbur, 2000). Since the susceptibility values within the venous system are related directly to the extent of deoxygenation of the blood (Fan et al., 2014), it is possible to derive the degree to which blood has been deoxygenated from magnetic susceptibilities (Bren et al., 2015).

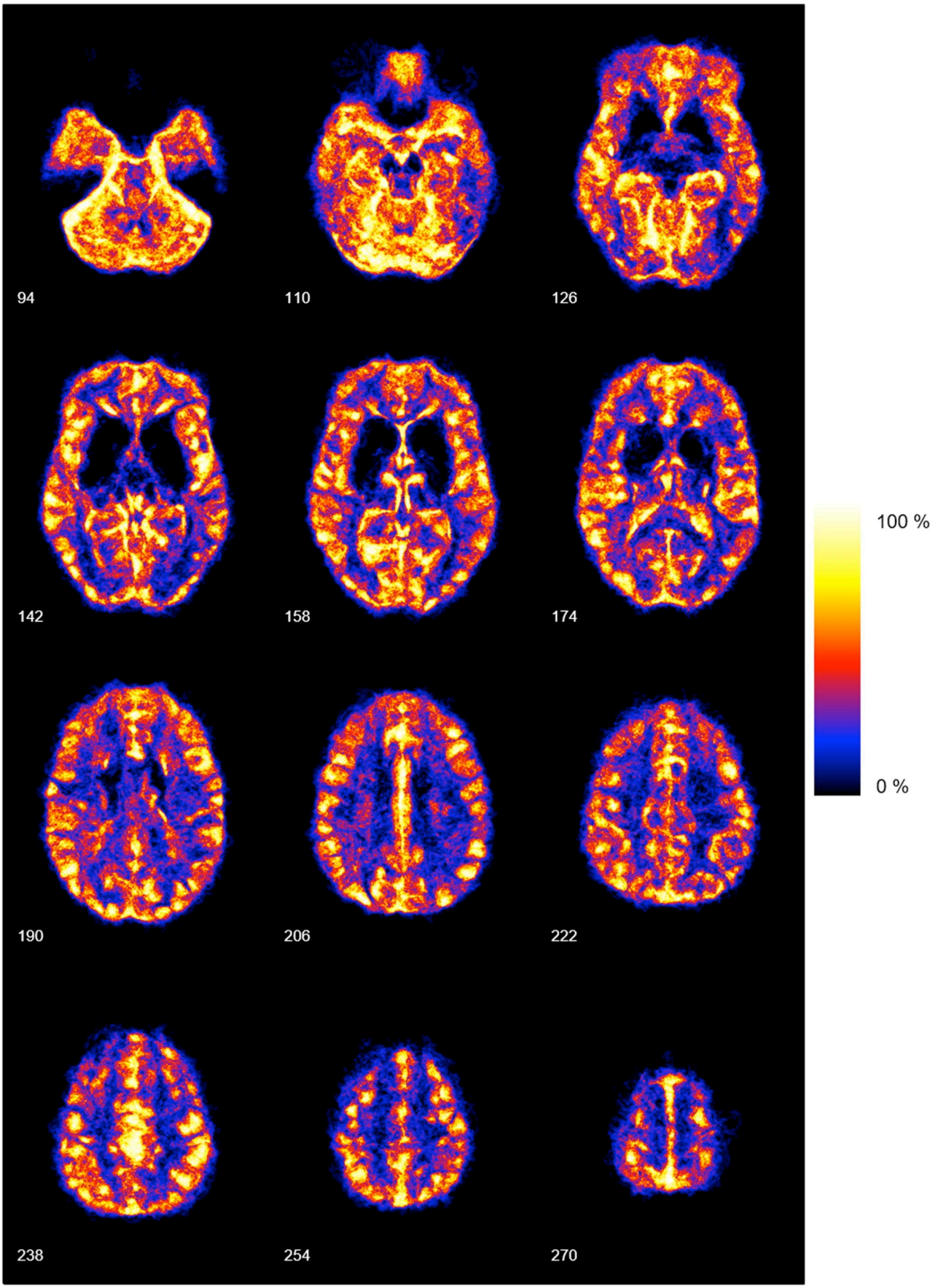

Figure 13: venous probability density map derived from 20 participants' QSM MRIs. Voxels with 100% probability are locations where all participants had a voxel classified as a venous vessel. The images show inferior to superior axial slices (Huck et al., 2019).

QSM can map the localisation of the venous system in the brain (fig. 13; Huck et al., 2019). The difference between QSM values of venous voxels and those of water can estimate the degree of deoxygenation in veins (Fan et al., 2014). This deoxygenation data is the basis for determining oxygen extraction fraction (OEF) across the brain.

OEF is a metric derived from QSM. OEF is the proportion of blood oxygen that has diffused out of the blood and into tissue in response to tissue metabolic usage of oxygen (Biondetti et al., 2023). The consumption rate is determined by the difference in oxygen saturation in arterial and venous blood as a proportion of arterial blood saturation (Biondetti et al., 2023). The arterial blood saturation is assumed to be 100%, and the venous blood saturation is determined from magnetic susceptibility of veins by QSM (Biondetti et al., 2023). OEF is an important biomarker and has been used to study both healthy (Huck et al., 2019) and disease states, including cerebrovascular disease (Fan et al., 2020) and Alzheimer's disease (Liu et al., 2020).

2.3 Computational Techniques

2.3.1 Machine Learning

Data driven techniques are central to the current scientific paradigm, enabling researchers to discover trends in the world through collecting and analysing larger volumes of data than was previously possible. These computational methods capable of handling large datasets are a powerful tool in identifying patterns in multivariate data. One specific method is machine learning (ML), a computational application of mathematics. ML is an algorithm that learns from data (Goodfellow et al., 2016). A general definition of learning is when a program or model has improved at a task according to a performance measure following experience or exposure to learning opportunities (Mitchell, 1997). Tasks may include fitting a line to data, identifying (or classifying) images, or dividing data into groups of similar instances (or clusters).

A frequently used distinction within ML is whether the learning is supervised or unsupervised. Supervised learning is when an ML algorithm is trained to associate input data with an expected output (Goodfellow et al., 2016) which serves as the ground truth. Often the inputs used for training are matched to outputs which are human generated; in other words, the inputs are 'labelled' by a 'supervisor'. An example of supervised learning is providing data for several plants (measurements, colour, shape, etc.) matched to their species name, then using the algorithm to identify or classify new examples of plants. Examples of supervised ML algorithms include linear regression, logistic regression, k nearest neighbours, decision trees,

support vector machine, and random forest. Unsupervised learning algorithms are given a set of inputs but no outputs and learn trends or features in the data. An example of unsupervised learning is the use of diffusion map nonlinear ML to untangle the free energy landscape of a single molecule from a multi-molecule simulation, and the impact thereon of environmental factors such as temperature and solvent conditions in order to examine its structure (Mansbach and Ferguson, 2015). Examples of unsupervised ML algorithms include gaussian mixture models, clustering (e.g. k-means, DBSCAN, OPTICS), covariance estimation, and manifold learning.

2.3.1 Clustering

Clustering, also called cluster analysis, is one application of ML that is very useful in scientific research. Cluster analysis is a family of algorithms that accomplish the unsupervised task of organising instances of data (data points, or samples) into sets or groups in datasets where group membership is not already labelled in the data. Clustering algorithms create clusters which should be optimised such that the instances within a cluster are as similar as possible to each other and as dissimilar as possible to instances of different clusters (Xu and Tian, 2015). The definition of similarity should be well-defined and consistent for a given model. In the case of quantitative features, similarity is usually defined as an inverse distance (with the distance between points being called the dissimilarity).

Clustering generally follows a standard set of ordered steps (Xu and Tian, 2015). First, the most representative features of the dataset must be extracted and selected. Secondly, an appropriate algorithm for the dataset in question must be chosen and hyperparameters of the algorithm identified; i.e. aspects of the algorithm must be defined such that they are adjusted to the specific characteristics of the dataset. For example, this might include selecting a proximity (inverse distance) measure and a clustering strategy appropriate to the clustering problem (fig. 14) (Xu and Wunsch, 2005). Next, the clusters must be evaluated for the validity of the resulting clustering solution to the problem of organising the specific dataset. Finally, the effective meaning behind the clustering results should be identified and explained.

Many different clustering methods exist. As described in the second step of clustering above, appropriate selection of clustering method for a given problem is essential to the generation of useful results (fig. 14).

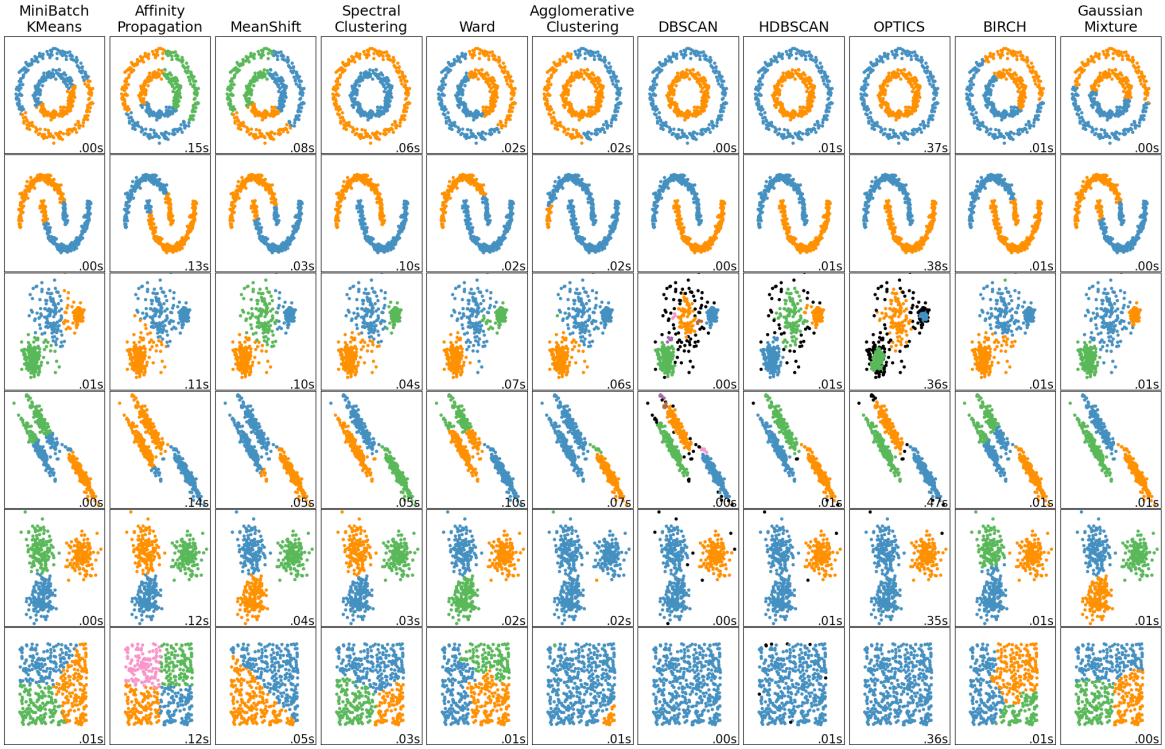

Figure 14: Examples of characteristics of a selection of cluster algorithms on two-dimensional toy datasets. Different types of data distributions are not well-categorised by every algorithm. For example, the first two rows of data are well characterised by spectral clustering, DBSCAN, HDBSCAN and OPTICS, but not the other algorithms. DBSCAN, HDBSCAN and OPTICS appear less appropriate for the fifth row of data. Spectral clustering categorises most datasets well; however, it separates the null dataset (last row) into meaningless separate clusters. Note the runtimes in the bottom right corner of each example, as a measure of the efficiency of each algorithm in processing each dataset. (scikit-learn developers, n.d. a).

Several common clustering algorithms are built on finding nearest neighbors to datapoints, which relies on finding the distance between points in the dataset (Cover and Hart, 1967). Starting with a query point selected from the dataset, an algorithm searches the space for the nearest points, which can be used to create clusters comprising k (an assigned positive integer value) members, plus the query point itself. A variant on finding a certain number of nearest neighbors is to identify neighbors within a fixed radius, which can be used to create a cluster by selecting all points within a certain distance or a specified radius (Bentley, 1975). Distance can be defined using a standard Euclidean distance or similar metric. When evaluating data with dimensions of varying scales, to avoid weighting the data based on scale it is necessary to normalise the data beforehand (Hastie, 2009).

This thesis project makes use of the nearest neighbours concept to define relevant subgraphs associated with voxel neighborhoods. This is slightly different from performing a clustering analysis. Clustering algorithms should maximise similarity between members of a group and minimise similarity between members of different groups. Additionally, measures of both similarity and dissimilarity should be well defined. The purpose of using clustering

algorithms in this project is to create subgraphs based on proximity. For this, clustering-related concepts are used, though the results are not true clusters as instances do not exclusively belong to only one group.

2.3.3 Strategies for Clustering High Dimensional and Large Datasets

Clustering datasets of high dimensional points or comprising many datapoints may be computationally impractical without first compressing or reorganising the data. Algorithms that can be useful for reducing dimensionality, essentially compressing datasets, include principal component analysis (PCA) and independent component analysis (ICA). An algorithm that can be helpful in managing datasets with a large sample size is K-dimensional (k-d) tree. Tree structures organise the data to make searches of the dataspace more efficient. In very large datasets, it may be advantageous to reduce dimensionality before creating a k-d tree. While this was not done in this thesis, exploring PCA and ICA for this application may be valuable in future work to reduce computational complexity. The greatest attention here, however, focuses on k-d tree as it was the sole algorithm selected for this research.

PCA is a statistical technique applied to research problems in many different disciplines including neuroscience (Kusztos et al., 2020), atmospheric science, and paleontology (Jolliffe and Cadima, 2016). To interpret high dimensional data, PCA linearly reduces the dimensionality of datasets while minimising information loss by preserving the variability of the data. The first principal component indicates the direction of greatest variation in the data (fig. 15 A). All subsequent components are orthogonal to all existing components and indicate the direction of next greatest variance subject to the orthogonality constraint. PCA essentially reduces to solving an eigenvector/eigenvalue problem, such that the eigenvectors are an orthonormal basis that describe the principal components, while the magnitude of the eigenvalue indicates the magnitude of variance along its principal component. It is most appropriate for use with data from a gaussian distribution. (Jolliffe and Cadima, 2016).

ICA is a dimensionality reduction method that separates multivariate data into source signals (Tharwat, 2021). It does so blindly – that is, without knowledge of the source signals or how they are combined. It assumes multiple independently sampled, non-gaussian signals exist within a given dataset. Each component carries the same weight, and the vectors are not necessarily orthogonal (fig. 15 B). ICA is often used as a pre-processing technique and has been applied to problems in neuroscience including MR image analysis (Gonçalves et al., 2009), fMRI, and EEG (Tharwat, 2021).

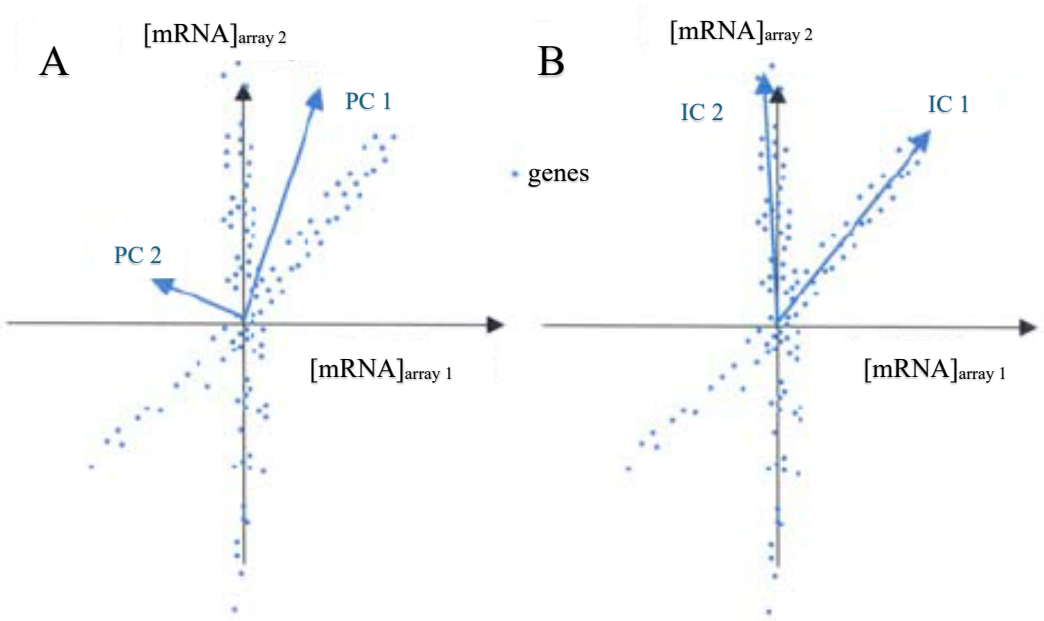

Figure 15: demonstration of dimensionality reduction techniques. In (a), the data is described by the two principal components. Note that the first component describes the axes of greatest variance, the second vector is orthogonal to the first, and that the two vectors have differing magnitudes (eigenvalues). In (b), the data is described by two independent components. Note that the vectors are not orthogonal. Both vectors describe the axes of greatest variance, and both are the same length. (Suri, 2003).

K-d tree is a type of data structure used for space partitioning with the aim of making searches of the dataspace more efficient. While multiple approaches to creating and using a k-d tree exist, the algorithm of interest for this thesis is described in Maneewongvatana and Mount (1999). A particular type of directed graph, trees organise datasets of points into a structure consisting of hierarchically connected nodes (fig. 16). The first node from which all other nodes branch is called the root node. In a binary tree, as in this k-d tree algorithm, all nodes but the edge nodes have one parent and two children. The first node, called the root node, has no parents. The final nodes, which have parents but no children, are called leaf nodes.

The binary tree structure describes the algorithm's execution in the data space. Each node is associated with a cell, that is, a rectangle of d-dimensions, where d is the dimensionality of the dataset. The cell of the root node contains all of the data points of the set. Each child node consists of a subset of data points contained within the root's cell, which is partitioned according to the "sliding midpoint" rule (fig. 16 iii) (Maneewongvatana and Mount, 1999). This trend continues recursively until the number of points contained within a node's cell is less than the 'bucket size' of the system, a threshold at which point the node becomes a leaf. Each internal node (non-root, non-leaf node) is associated with a splitting hyperplane. The algorithm uses a hyperplane to divide the space recursively by splitting along the middle of the longest axis only if there are points on both sides of the split (Narasimhulu et al., 2021).

Otherwise, it splits at the datapoint closest to the middle of the axis; that is, it 'slides' the splitting plane to the nearest data point (fig. 16 iii-v) (Narasimhulu et al., 2021).

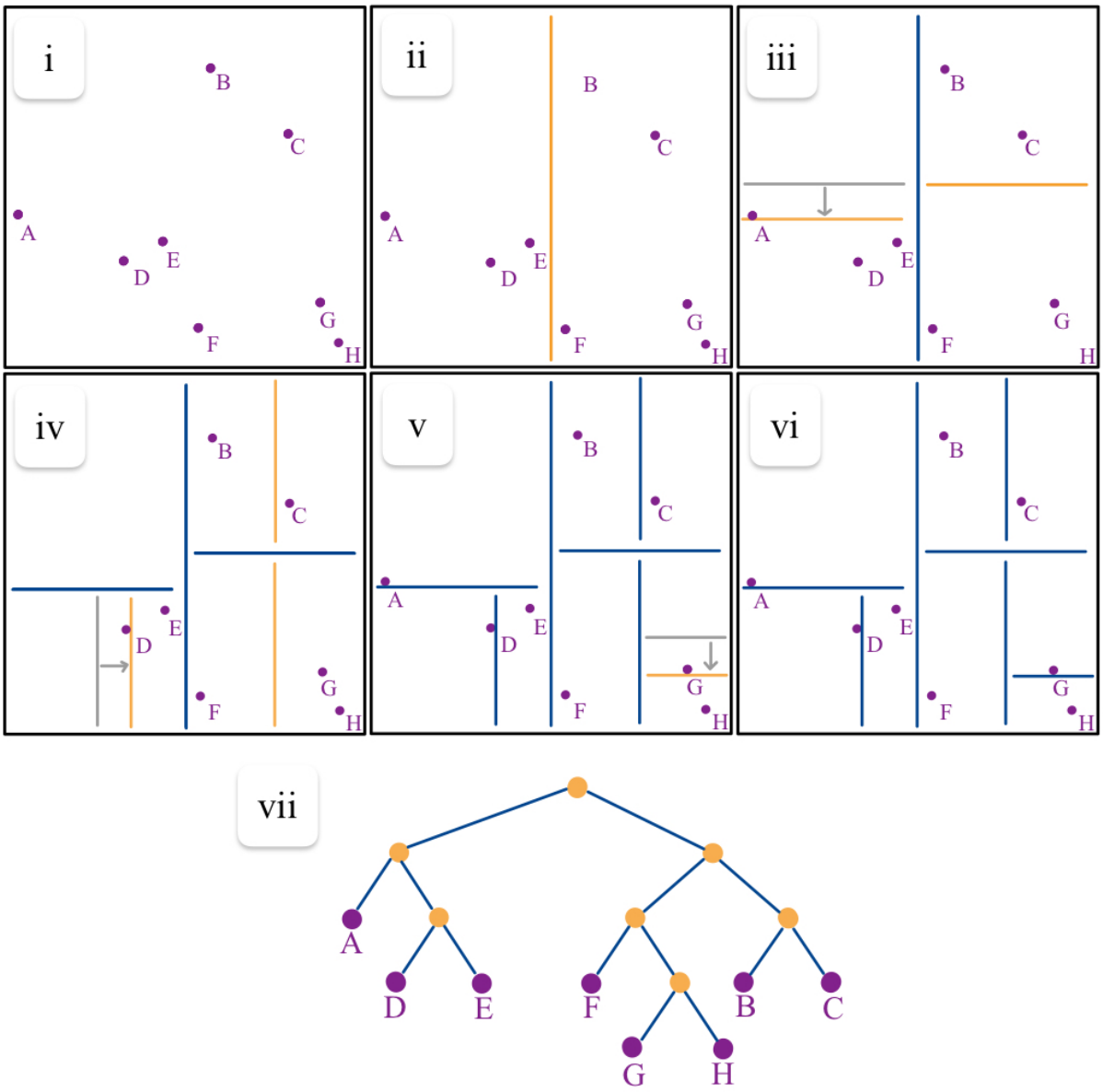

Figure 16: constructing a binary tree using sliding midpoint. Note that in iii, iv, and v, the division falls such that there are no data points in one of the cells, therefore the midpoint slides to the nearest datapoint. Adapted from Havran, 2016.

Once the k-d tree is constructed, it can be used to efficiently find the nearest neighbours to a query point using a simple recursive algorithm (Maneewongvatana and Mount, 1999). When the query point is part of the dataset, it is in a leaf on the tree. When looking for a set number k of nearest neighbours, from the query leaf, the distance to all other points associated with this node is calculated and the k shortest distances are recorded. The algorithm moves to the next parent node where it verifies where the query point is relative to that node's hyperplane. The query point is necessarily nearer to the child on the same side of the hyperplane; this child is assessed for nearness and compared to the recorded shortest distances.

If the hyperplane is at a shorter distance, this child is examined recursively. If it is not shorter it does not need to be explored, saving computational resources. When the search reaches and returns from the root, the shortest distances indicate the k nearest neighbours of the query point. The tree can also be searched for all points that are at a distance less than a specified radius. While searching the tree, any points that are at a distance less than the set radius are recorded as a neighbour. In following this algorithm, every leaf that is closer to the query point than any one of the neighbours on its initial proposed list of nearest neighbours will be verified, while not all points in the dataset will necessarily need to be evaluated.

In summary, the k-d tree structure allows the space to be more efficiently searched for the nearest neighbours of a query point. By pre-structuring the data using k-d tree, where the neighbours on the tree are also the neighbours in the dataset's space, the algorithm does not need to directly verify the distance to every point in the dataset (Maneewongvatana and Mount, 1999). While the algorithm can handle larger datasets, its efficiency declines with high dimensionality. For dimensions greater than 20, the running speed is about the same as direct comparison between each point and every other point in the dataset (SciPy community, n.d. h). Efficient searches for nearest neighbours in high-dimensional datasets remain an open problem in computer science.

2.4 Network Graphs

Graph theory is a mathematical tool that has been used in many scientific fields including chemistry, physics, and neurosciences (Zhou et al., 2020; Bullmore and Sporns, 2009; Ottet et al., 2013). A graph, also called a network or a network graph, is a framework for organising data consisting of nodes interconnected by edges (fig. 18) (Betzel & Bassett, 2017).

Nodes, also called vertices, can contain information about an object (Wilson 1996). For example, in a social network, nodes would represent users and might contain information such as age, gender, location, occupation, and hobbies. In chemistry, nodes might be atoms in a molecule. The information for each node could include the element type, weight, valence, and isotope (fig. 17 A, B).

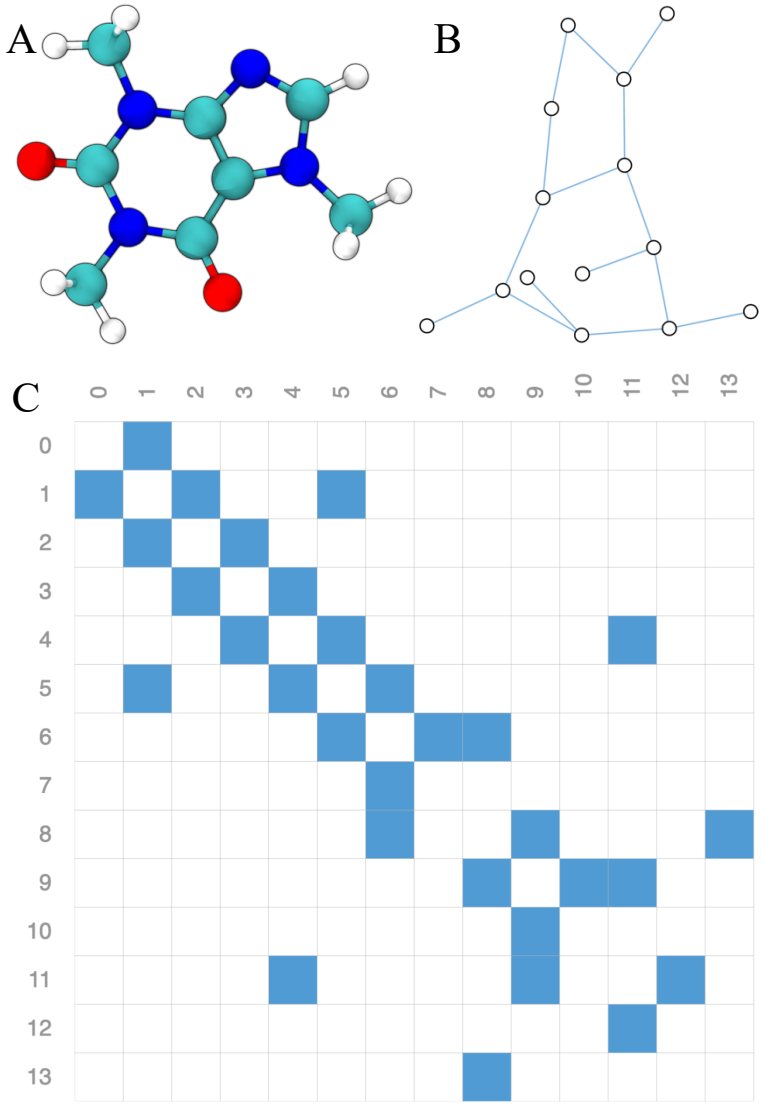

Figure 17: A. Caffeine molecular structure, B. related graph where black circles are nodes representing atoms and blue lines are edges representing covalent bonds, and C. the adjacency matrix of the graph with blue filling representing edges. Note that the adjacency matrix only shows the presence of bonds (unweighted edges) and not the strength or distance of the bonds (weighted edges). Adapted from Sanchez-Lengeling et al., 2021.

Edges provide information about the relationship between the objects of interest (nodes of the graph) (Wilson 1996). Edges may be associated with weights, with a higher weight indicating a stronger relationship between the connected nodes, typically depicted visually with a thicker line. Edges may also have a direction. For example, if one person owes money to another, the edge connecting them that describes the debt is directed from the debtor to the loaner, whereas the flow of money would be directed in the reverse direction. Edges may also be undirected, for example, describing the physical distance between two shops in a town or the covalent bonds between atoms in a molecule (fig. 17 A, B).

Graphs contain local or small-scale information about the data points, as well as large scale information about the system as a whole, which can be extracted by a number of metrics. Small scale information can include the nodes and edges themselves, node/edge centrality, and node/edge degree. Large scale information can include measures such as the degree distribution, assortativity, connection density, number of nodes, and longest path. More recently, deep learning tools known as graph neural networks have been developed to allow the use of such powerful algorithms with graphs as inputs (Sanchez-Lengeling et al., 2021).

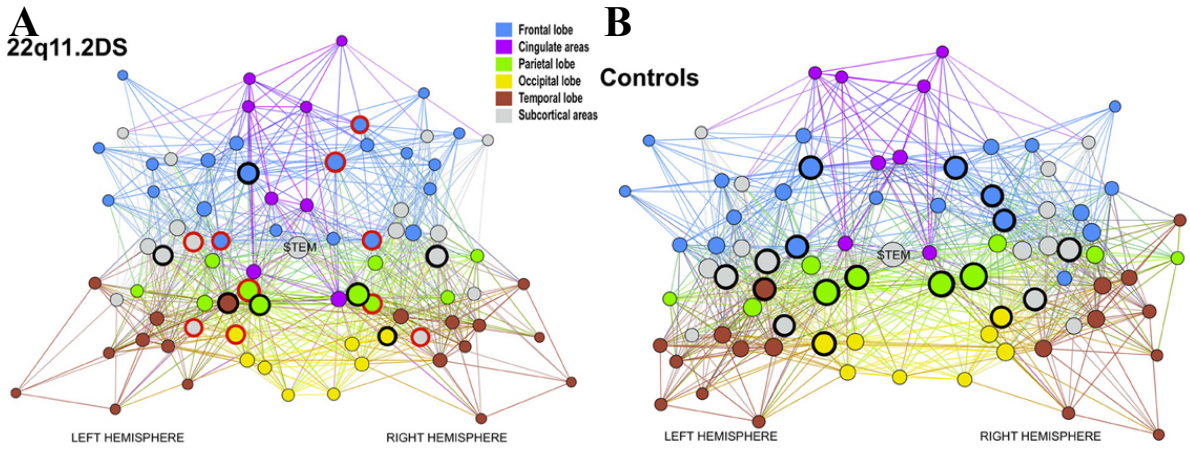

Figure 18: An example of a graph used in neuroscience. Here, structural and diffusion MRI data for two populations is displayed: A. patients with a genetic subtype of schizophrenia (22q11.2 deletion syndrome), and B. healthy controls. The circles are nodes which represent cortical regions. The lines are edges representing white matter bundles. The bold red circles are hubs that are altered in the patient population whereas the bold black circles are preserved hubs. The size of each node denotes degree. The colours represent brain regions. Adapted from Ottet et al., 2013.

The *degree* of a node refers to the number of connections to other nodes in the graph (Wilson 1996). A node that is not connected to any other node is called isolated and has degree 0. A node that has one edge to it is called an end-vertex and has degree 1. Since edges, by definition, connect two nodes, the sum of all node degrees must be even. The distribution of degrees in a graph is called the *degree distribution* and is a global measure of the system (Bullmore and Sporns, 2009). More complex networks tend to have more positively skewed distributions, with most nodes having low degrees and a few nodes having higher degrees. *Assortativity* is the connectedness of like-degree nodes. A graph where high-degree nodes are connected to each other is said to have positive assortativity. *Connection density* is the proportion of edges in a graph compared to the number of possible edges in a graph.

A *path* is defined as a traversal of the graph, from one node, along an edge, to a subsequent node, to another subsequent edge and so on, such that no node appears more than once in the sequence (Wilson 1996). The shortest *path length* is the fewest number of edges connecting two nodes (Bullmore and Sporns, 2009). *Centrality* is a node attribute

corresponding to the number of shortest paths passing through the node. A high centrality suggests that the node has high importance to the communication and connectivity within the network; such a node is called a *hub*.

One way to identify local patterns within a graph is to analyze patterns in its subgraphs. A subgraph is any subset of nodes and edges of a graph, such that all nodes and edges of the subgraph are a part of the larger graph (Wilson 1996). Subgraphs can be created by removing nodes and edges of a whole graph. A *network motif* (fig. 19) is a subgraph consisting of a pattern of nodes and edges that appear frequently in a single graph or across sets of graphs (Bullmore and Sporns, 2009). Their distribution throughout the graphs suggests similarity of functions in different regions. It can be computationally demanding to determine the motifs in a graph. A *module* or community is a type of subgraph whose nodes are highly interconnected within the module, and less connected to nodes outside of the module (fig. 20). The identification of communities in a graph is the province of community detection algorithms. Hub nodes (cf. previous paragraph) can be further categorized according to their relationship with the communities of a graph. Provincial hubs are hub nodes that are highly connected within a single community. Connector hubs are high centrality nodes of one community that are connect to nodes in another community.

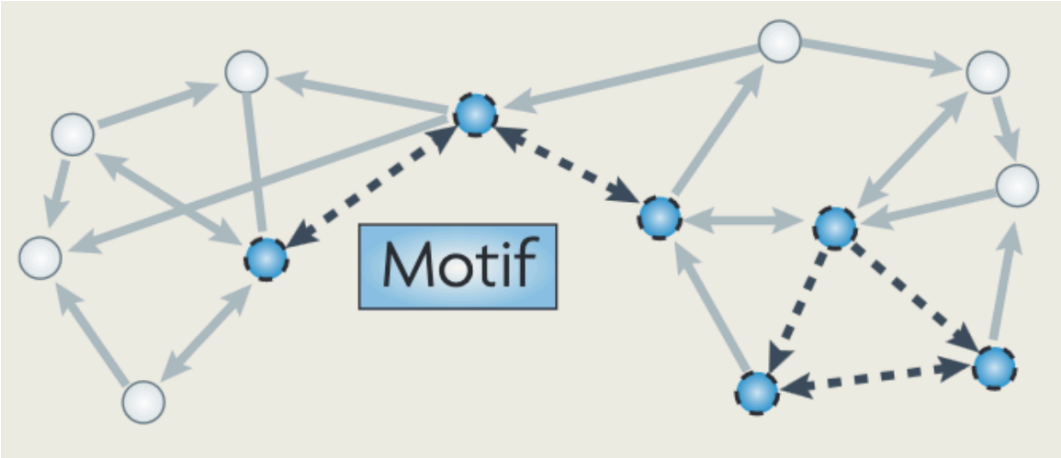

Figure 19: Repeating patterns of nodes and edges are called motifs. Adapted from Bullmore and Sporns, 2009.

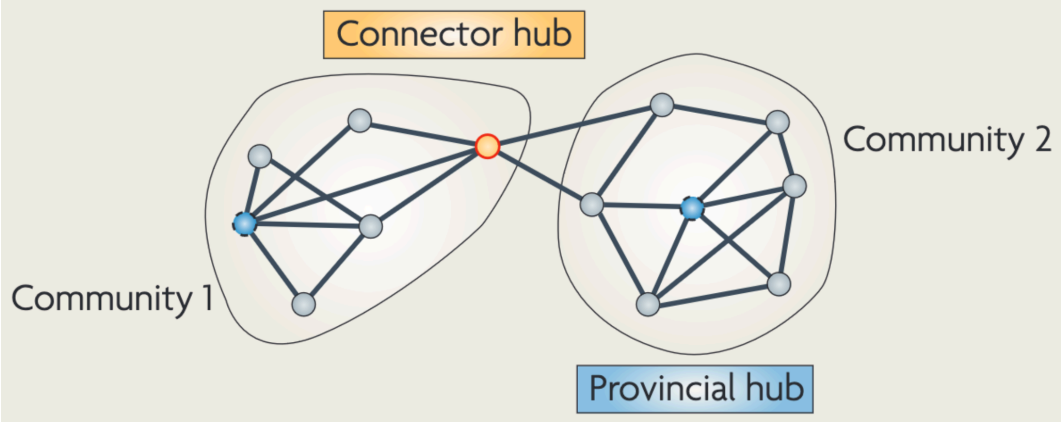

Figure 20: Sets of highly connected nodes with relatively low external connectivity are called modules. A specific module is called a community. Provincial hubs are highly central to their community, whereas connector hubs are central nodes connected to nodes both within and outside of their community. Adapted from Bullmore and Sporns, 2009.

There are many different ways to represent graphs (fig. 18; fig. 17 B, C). When representing a graph computationally, a matrix representation is a memory-efficient format for storage, especially of larger graphs. One such matrix is an adjacency matrix (fig. 17 C), which represents the relationships between nodes. For a graph of n vertices, an n x n adjacency matrix is constructed such that the entry at ij describes the edge between the i-th and j-th nodes. When quantified, adjacency is inversely related to distance; nodes which are highly adjacent will have a minimal distance between them and vice versa. A particularly memory-efficient tool for computational storage of adjacency matrices is the sparse matrix. A sparse matrix is an object designed for use when a matrix contains a large number of zero elements. It stores all non-zero entries as a tuple of row-column indices along with the value of the entry (Educative, n.d.; SciPy community, n.d. g). If an entry is queried and the tuple not found, the entry is assigned a zero. Because of this approach, sparse matrices do not save to memory the large number of zero entries, which can greatly reduce the memory requirements for the object.

3. Related work

Here, we include a brief overview of literature related to various aspects of this research. We note that, to our knowledge, the derivation of the brain physiological connectome via graph-based methodology for use in generative deep learning-based data augmentation is a novel application.

3.1 Networks and Neuroscience

Conceptualizing the brain as a series of overlapping networks with anatomical or functional connectivity is a well-established framework in neuroscience. Typically, connectivity is either based on structural connections that are inferred from diffusion-based imaging techniques, or functional networks that are defined for regions having similar temporal signal characteristics. In its traditional form, network analysis of the brain almost always refers to spatial networks. For example, Fornito et al. (2013) examined the use of network graphs and topological analysis in neurosciences and examining the "human connectome". They also discuss several studies using generative models in analysing topology in brain networks. For example, a study by Raj et al. examined neurodegenerative disease by modeling disease progression using a diffusion model on a diffusion-weighted image-based structural connectome (Raj et al., 2012). This resulted in a hypothesised pattern resembling the atrophy of both Alzheimer's disease and fronto-temporal dementia. Betzel and Basset's (2017) paper on brain networks reviewed the use of generative network models which focused on structural neuronal networks across several species including humans. A recent study by Waikhom and Patgiri (2023) reviewed many applications of graph neural networks, including in neuroscience research. In their article, they discuss how whole-brain network graphs have been used to show differences in structural and functional connectivity in brain imaging data between several patient populations and control, including the ways in which these studies make use of functional connectivity centrality to determine the important regions in the brain involved in diseases such as bipolar disorder (Deng et al. 2019; Zhou et al. 2017), retinitis pigmentosa (Lin et al. 2021), and diabetic optic neuropathy (Xu et al. 2020). Finally, a seminal paper by Sporns (2018) discusses graph theory as a mathematical tool for interpreting brain network graphs, which are used in a large array of studies of network properties of the brain.

3.2 Data Augmentation in Medical Imaging

Previous work in data augmentation for deep learning applications in medical imaging has primarily been focused on the domain of image classification and segmentation (Chlap et al., 2021), but there have been a few other approaches. One study proposed a novel approach to data augmentation by adding noise to fMRI time series data which performed well on a sex classification task (Kusztos et al., 2020). Nguyen et al. (2020) proposed a new, coregistration-based method for generating fMRI data for augmentation, which demonstrated an improvement in treatment response predictions of patients to antidepressants.

3.3 Generative Deep Learning and Neuroscience

Generative deep learning has been applied to problems across scientific fields and beyond. We mention here several of the most pertinent studies to this thesis. In neurosciences, one study successfully applied generative adversarial networks (GAN) and latent diffusion models in neuroimaging to generate structural synthetic brain imaging sets (Pinaya et al., 2022). An encoder-decoder network was also used to map cerebral hemodynamic function (Hou et al., 2023). Wang et al. (2021) reviewed the use of generative deep learning in synthesising brain imaging data between modalities, for example, using a generative adversarial network (GAN) to predict a patient's brain computed tomography image from their brain MRI (Kazemifar et al., 2019). GANs have also been used to model disease progression and natural aging of the brain (Wang et al., 2023). Specifically, a GAN was used in predicting the growth of glioblastoma brain tumors (Kamli et al. 2020), and a conditional attention GAN was used with positron emission tomography to predict myelin content in multiple sclerosis patients (Wei et al., 2020).

4. Methods

4.1 Participants

MR imaging (the COV-IRM dataset) was collected from 15 women who experienced long Covid, aged 25-86 (average ± standard deviation of 47.87 ± 14.19 years). Data was also collected from 4 female controls with acquisition ongoing at the time of writing, ages 54-81 (61.75 ± 18.73) with the intent to match recruited controls in age, sex, and hypertension status to the long Covid cohort. Controls may have a history of mild Covid infection if they were vaccinated but were excluded if they suffered more serious Covid-related symptoms such a loss of taste and/or of smell, confusion, dizziness, concentration difficulties, severe headache, respiratory difficulties, or elevated fever.

All participants were screened for contraindications to testing apparatus including 3T MRI, the RespirAct gas delivery system, and the cardiac stress test exercise bicycle; most contraindications are common to more than one apparatus. Contraindications for 3T MRI include claustrophobia and ferromagnetic implants. Conditions affecting the cardiac stress test include severe exercise intolerance or exercise limitation due to non-cardiopulmonary conditions such as arthritis or claudication. The complete list of contraindications to the RespirAct system can be found in the RA-MR Operator's manual (Thornhill Medical, 2017).

General exclusion criteria consist of conditions and illnesses which may be associated with altered brain function. Potential participants were screened and excluded for any history of neurological illness, psychiatric illness, cognitive impairment or dementia with a mini-mental status exam score less than 25 (Folstein et al., 1975), thyroid disease, diabetes, systemic hormone therapy, non-Covid respiratory illnesses, or other chronic, progressive illnesses unrelated to Covid. Other exclusions were current and regular smokers; excessive alcohol consumption (defined as > 2 drinks per day); ICU admission or intubation due to Covid; surgery under general anesthesia within the preceding six months; cardiac insufficiency; recent acute coronary events; restrictive lung disease; supplemental oxygen requirements; severe discomfort to hypercapnia; or severe heart failure or pulmonary disease.

All participants were fluent in English or French to facilitate informed consent and for the validity of neuropsychological testing. The study was approved by the Comité d'éthique de la recherche et du développement des nouvelles technologies (CÉRDNT) de l'Institut de Cardiologie de Montréal.

4.2 Breathing Manipulation

For CVR measurement, a breathing manipulation took place during the first sub-acquisition (pCASL sequence) in the MRI scanner. CO2 delivery was performed using a computer-controlled gas manipulation device (RespirAct device by Thornhill Research Inc., Toronto, Canada) (Slessarev et al., 2007). The device administers gases to the participant in the MRI through a closed system and face mask. It controls the flow rate and concentrations for prospective targeting of end-tidal values following a pre-determined protocol. It also measures end-tidal oxygen and CO2 partial pressures. Initially, participants breathe room air (normocapnia) through the mask for two minutes. This is followed by two minutes of breathing at CO2 levels at 5 mmHg above the participant's resting state CO2 (hypercapnia). Next, two more minutes of normocapnia are then followed by two minutes breathing elevated oxygen levels of 150 mmHg above resting O2. Finally, two more minutes of normocapnia are followed by breathing a mixture of 5 mmHg CO2 and 150 mmHg oxygen above resting state for two minutes. The test concludes with two final minutes of normocapnia for a 14-minute total protocol.

4.3 MRI Acquisition and Processing

Acquisitions were performed at the Institut de Cardiologie de Montréal. The scanner is a 3T Siemens Magnetom Skyra with a 32-channel array coil. A T1-weighted magnetization prepared rapid acquisition gradient echo (MPRAGE) was collected with a TR= 2300ms, TE= 2.32ms and flip angle of 8° and a 0.9 x 0.9 x 0.9 mm resolution. This sequence provides structural information about the brain and an image to better visualize the spatial information from the remaining metrics. This data is not considered by the clustering algorithm but will be used for data visualisation.

Multi-shell diffusion-weighted imaging (DWI) data were acquired with a 2x2x2 resolution using a pulsed gradient spin-echo sequence (TR = 6000 ms, TE = 106 ms). Diffusion-sensitizing gradients of different strengths (b-values of 300, 700, 2500 s/mm^2) were applied in several different directions (10 at 300 s/mm^2; 30 at 700 s/mm^2; 64 at 2500 s/mm^2). Non-diffusion-weighted images (b-value of 0 s/mm^2) were also acquired to provide a baseline. The DWI acquisitions lasted 14 minutes total. As mentioned in section 2.2.2, DWI was processed using both the DTI and NODDI models to provide the following measures: ISOVF, ICVF, FA, MD, AD, RD, ND, and ODI.

Myelin content is estimated with magnetisation transfer saturation (MTsat) from magnetization transfer imaging (MTI). This is measured using two spoiled gradient-echo

(SPGE) acquisitions, one with a preparatory MT pulse and one without (MT-on and MT-off, respectively) along with a T1-weighted image and a B1 map. MT-on and MT-off images were acquired with these parameters: TR = 33 ms, TE = 3.81 ms, flip angle = 10° and a resolution of 2x2x2 mm. These two sequences were identical except for the off-resonance preparatory pulse (MT pulse) applied during the MT-on acquisition. The T1w image was acquired with a 2x2x2 resolution, TR = 15 ms, TE = 3.81 ms, and flip angle = 25°. Additionally, a B1 map was collected to correct for transmit field inhomogeneities. These acquisitions lasted about 9 minutes total. qMRLab (Karakuzu et al., 2020), a Matlab-based toolbox, was used to compute MTsat from the MTI data, using the T1w image and B1 field map to correct for T1 relaxation dependence and field inhomogeneities.

Investigations into measures of cerebrovascular health are done with a pseudocontinuous arterial spin labelling (pCASL) sequence. The data was acquired over 14 minutes using a 1300 ms post labeling delay and labeling duration of 1.5 s, with TR = 4000 ms and TEs = 10 ms and 30 ms. These parameters were adjusted to this project's specific population in pilot data. The 10 ms echo of this sequence provides CBF and the 30 ms echo to assess CVR.

Both echos of the pCASL data were motion corrected and brain extracted using FSL BET. For CBF quantification, a General Linear Model (GLM) was used to regress out motion parameters in MATLAB. The first pair of tag and control images were removed for signal stabilization and reliable motion estimation. Perfusion weighted images were generated by subtracting the tag and control images using surround subtraction of the first 3min of acquisition (Wu et al., 2011). The average of the control images was used to create the M0 image. CSF masks were derived from the M0 images to serve as the M0 mask for final CBF quantification. 10 voxels were selected from the same axial slice containing the lateral ventricles for each participant, following the method described in Intzandt et al. (2020). Quantitative CBF maps were then calculated using average perfusion maps and FSL's BASIL toolkit. To correct for partial volume effects, structural images were segmented into GM, WM, and CSF using CAT12. The segmented tissues were then transformed to the ASL native space using the transformation matrix obtained from registering the mean ASL image to the T1 space using ANTS. Partial volume correction was performed using a method introduced in a referenced study (Asllani et al., 2008).

The TE = 30 ms time series was used for CVR quantification. A blood oxygen level dependent (BOLD) time series was obtained through surround subtraction of the TE = 30 ms time series. A GLM was used to estimate the fractional change in BOLD signal during the hypercapnia block. This percent change in BOLD signal was divided by change in end-tidal CO2 in mmHg to obtain voxel-wise CVR maps.

Quantitative susceptibility mapping (QSM) was derived from the phase and magnitude data from all uncombined coil channels of a 6-minute 3D gradient-recalled echo sequence. QSM data acquisitions were performed using a 32-channel coil, using a 3D multi-echo

gradient-echo sequence (TR/TE1/TE2/TE3/TE4/flip angle = 20 ms/6.92 ms/13.45 ms/19.28 ms/26.51 ms/9°, 0.7 x 0.7 x 1.4 mm^3 voxel size). Raw unwrapped phase data were combined and wrapped using the ROMEO toolbox (Dymerska et al., 2021). The QSM maps were reconstructed using the TGV toolbox (Langkammer et al., 2015). QSM maps were referenced to ventricular CSF susceptibility values. Then, a Recursive Ridge Filtering method (Bazin et al., 2016) was used to extract draining veins and calculate venous OEF (Fan et al., 2014).

Non-brain voxels were removed from all computed maps using FSL's brain extraction tool (Jenkinson et al., 2005). In addition, masking was applied to retain only voxels of interest for each modality (e.g., keeping only white matter voxels for metrics derived from DWI). Masks of white matter, grey matter, and CSF were generated by segmenting a structural image (MPRAGE T1w) using the FAST algorithm from the FSL toolbox (Smith et al., 2012; Smith, 2002; Zhang et al., 2001; Patenaude et al., 2011; Smith et al., 2004). Masking was performed by multiplying the appropriate binary mask by the map (e.g., FA maps were multiplied by white matter mask).

Current protocols ensure all research MRIs are verified by radiologists for incidental findings. All findings are reported to the participant and their family doctor who will be responsible for follow-up.

4.4 Subgraphs

The primary purpose of this project was to develop the first step of a data augmentation and analysis methodology in the context of the COV-IRM dataset. As this dataset was still being acquired at the time of writing, toy data was used for the purposes of developing the method. The toy data was selected from a dataset employing the same MRI protocol and included 7 white matter features (AD, FA, MD, RD, ICVF, OD, ISOVF); the other features described in section 2.2 were not available at the time of development. The code was developed with this toy data to mimic the data management challenges of the COV-IRM data to allow for transferability. As the toy dataset consists of 359 938 white matter voxels from a single participant's brain, graphical visualisation is not possible. For the purposes of data visualisation for this thesis, a downsampled version (1073 white matter voxels) of this toy data was generated and processed using the same technique as the original toy data (see section 5.1). The graphs were plotted using Gephi (Bastian et al., 2009).

The code is organised according to the publicly available Shablona template for small scientific python projects (Rokem, 2015) with the commands for this segmented into two scripts (fig. 21). The first script is used for reformatting the original MRI data, and the second script is used for subgraph generation. Specifically, the first script (*git_generate_featurematrix.py*), generates a matrix of features of size N x M, where N is the

number of voxels plus one row for column labels, and M is the number of features, or MRI metrics. This matrix (referred to in the code as *labeled_matrix*) is composed of a flattened array of white matter voxels (though the same process will later be applied to grey matter as well) and their features derived from MRI data .nii files, where the MRI data is in the form of a cubic array. The second script (*git_ckdtree_clustering2.py*) uses the output of the first script (i.e. *labeled_matrix*) to create subgraphs of nearest neighbours in matrix form; it saves as csv files both an adjacency matrix (*adj*) and a matrix of features (*cluster*) per subgraph of neighbour voxels. In the following sections, I describe the composition of the scripts and how they were used to generate results in this project, with the names of arguments in parentheses along with the specific values that were used in this project to validate the code with the toy data.

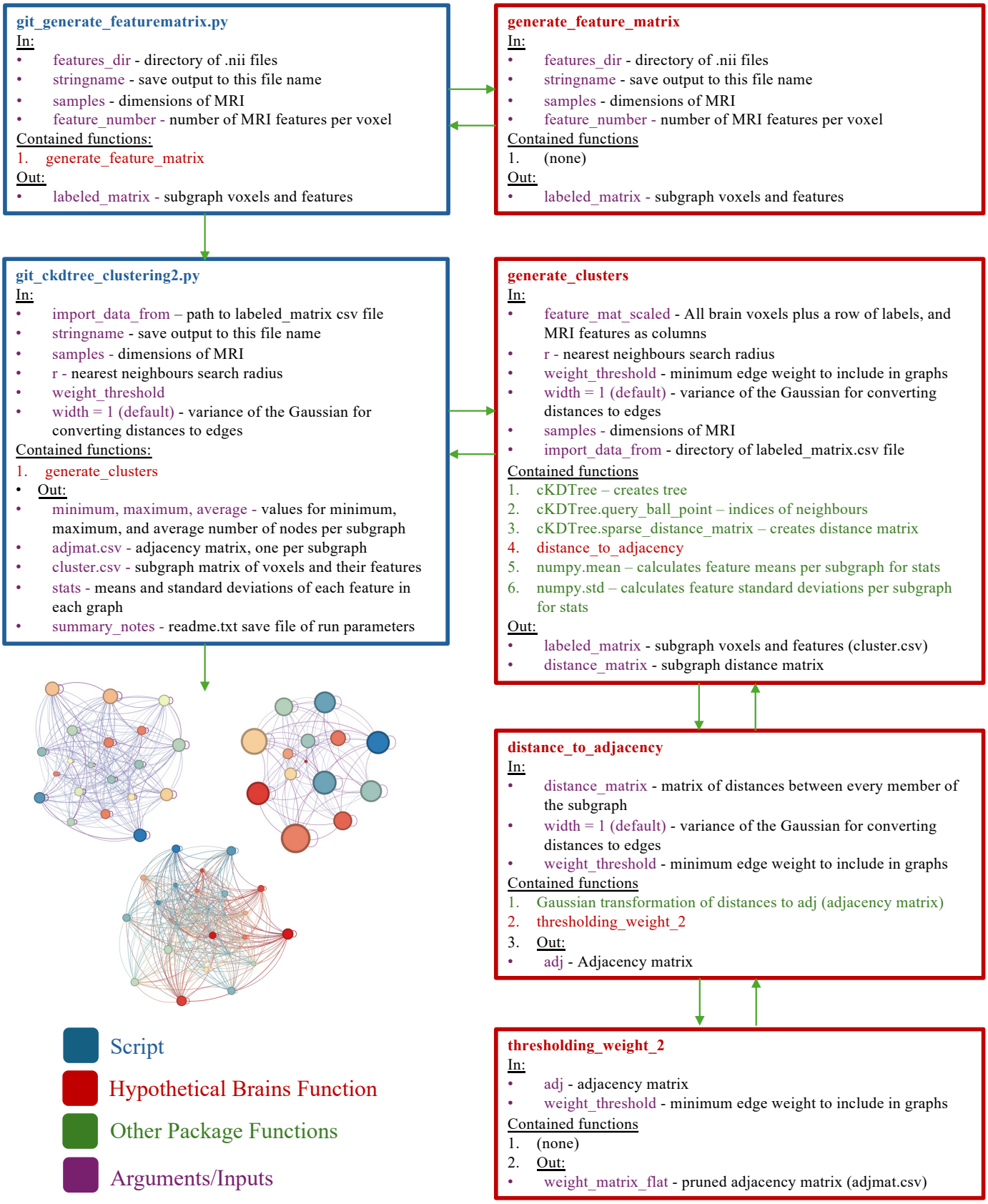

Figure 21: workflow of method indicating direction of information flow and argument passing.

4.4.1 Generating a Matrix of Features

The first script (*git_generate_featurematrix.py*) runs a single function, generate_feature_matrix. It requires user-supplied arguments for input, as follows: the dimensions of the cubic MRI data (*samples = 256*), the number of features per voxel (*feature_number = 7*), the directory where the .nii files are saved (*features_dir = '/home/lwright/Desktop/TrialData/'*), and the name for the .csv matrix save file (*stringname = 'brain_aug17_features'*).

This command produces the flattened array of MRI data with one voxel per row, and MRI metrics (features) as columns. It uses the nibabel package (Brett et al., 2024) in python to load the MRI data from .nii files, one for each of the seven features. These data are then masked to exclude voxels imaging only air or containing only noisy data (for example, for DTI, voxels containing mostly grey matter are excluded). These data are collected in a 2D array of features per voxel (*voxel_number* rows and *feature_number* columns). Each column is scaled using StandardScaler from sklearn.preprocessing package. This shifts the mean of each column of feature data to zero and scaling to unit variance (Scikit-learn developers, n.d. b), allowing features with naturally very different scales to be more straightforwardly comparable. A row of numerical labels is added (0. to 6.) to identify the columns, and the feature matrix saved as csv and a print message confirming the completion of the script is displayed ("Attributes saved").

4.4.2 Creating Subgraphs of Nearest Neighbours

The second script runs one function, generate_clusters, and requires user specified arguments. These arguments include, as previously described, the same voxels dimension variable (*samples=256*), the same csv file name (*stringname = 'brain_aug17_features'*), and the whole path to the csv file (*import_data_from*). It imports the data saved in the csv file by the previous script (*feature_mat_scaled*) and removes the labels which are unnecessary for the computerised analysis.

This script runs one function, *generate_clusters*, to create subgraphs by performing a nearest neighbours search. It requires as input the feature matrix, and two new variables that must also be specified: the radius for the nearest neighbour search (*r = 0.3*); and the minimum edge weight to store in the adjacency matrix (*weight_threshold = 0.1*), where any weight below this is set to 0. The radius was selected to obtain subgraphs on the order of one to two thousand voxels, and the edge threshold was selected conservatively to allow for pruning as appropriate. One optional input is the *width*, a parameter that describes the variance of the Gaussian distribution (*width = 1*, default) used to smooth the distribution of edges as an optional pruning

method when converting from distance to weight (see further discussion below). This command accepts, but does not require, the *samples* and *import_data_from* variables as input at present. They are used currently only for documentation purposes and are included in an output summary file describing the parameters used for a given run.

The function *generate_clusters* begins by starting a timer (*now = datetime.now()*); creating a string (*dateonly*) with the full date and time of the initiation of the run for file, directory, and documentation purposes; and creating a run directory with subdirectories for csv files. The files and run directory both have name configurations as follows: HypoBrains_Y2023_M12_D04_H12_M51_S19_v359938_r0.3 [project name, year, month, date, time, *voxel_number = 359938*, *r = 0.3*]. Next, a tree object is created from the feature matrix using cKDTree in the python scipy package (SciPy community, n.d. a). This method was selected for use in finding nearest neighbours owing to its computational power for large datasets and ease of implementation. In this case, nearest neighbors are determined by distances in feature space, not spatial proximity of voxels, in order to create a "physiological connectome." A set of variables were instantiated for later use in collecting statistics on the subgraphs: *minimum = 0, maximum = voxel_number, average = 0*, and stats (a 3D array of zeros, voxel_number x feature_number x 2) for collecting both the mean value and standard deviation in the two layers of the matrix of each feature for each subgraph.

4.4.2.1 Looping Through Voxels

We construct a number of subgraphs equal to the number of voxels in the dataset, by using the cKDTree to find the feature-space neighbours (within a radius *r*) of each voxel. This is done as follows. A loop iterates over each voxel in the dataset (with iteration variable *i*). The indices of all of this voxel's neighbours within *r* radius distance are determined by cKDTree.query_ball_point (SciPy community, n.d. b). Inputs used are: the already instantiated tree; the *i*-th entry of the feature matrix (*feature_mat_scaled[i]*) as the starting point for the neighbours search, which corresponds to the voxel under consideration; the radius *r = 0.3*; the distance calculations were by Euclidean distance by selecting the Minkowski p-norm to be *p = 2*. (National Institute of Standards and Technology, 2017); the default eps = 0 was used for an exact determination of neighbours instead of approximation; parallel processing using all processors was selected (*workers = -1*); the returned array of indices was sorted (*return_sorted = True*) for ease of use in subsequent steps; and *return_length* = False so that the index values are returned, as is required.

For each voxel *i*, a 2D array is created (*cluster*, of size M x N, where M is the number of neighbours and N the number of features, plus one more row for feature label headings) in which we collect the indices and features of its neighbours. A second tree is created from the

sub-cluster array (*tree2*), which is then used to determine the distances between all member voxels (*distance_matrix)* with cKDTree.sparse_distance_matrix (SciPy community, n.d. c). This command takes the cluster tree as input twice, as well as a maximum distance metric above which the distance is encoded as zero. This distance (max_distance = *100000*) was selected to be an arbitrarily high value to retain all distance measures; this is to keep edge pruning for weighted edges which is more easily implemented. The distance matrix is in the default "dictionary of keys" (dok) format (SciPy community, n.d. f).

The next step in *generate_clusters* is converting the distance matrix to the adjacency matrix of the corresponding *i*-th subgraph. This is accomplished using a defined function, *distance_to_adjacency*. It takes the distance matrix, and the aforementioned Gaussian width and weight threshold variables. The distance matrix is first converted to dense format as neither dok nor sparse matrices can be mathematically manipulated as is required here. Next, the entries of the distance matrix (*x*, here) are converted to weights. There are multiple ways to do so. For visualisation, an exponential function was used according to the formula:

$$weight = e^{\frac{-x}{width}}$$

Since distances are always positive values, this process still ensures that all returned weights are in the range (0, 1]. However, for future applications of this method, a Gaussian is recommended for its rigor:

$$weight = e^{\frac{-x^2}{width}}$$

Since this formula is Gaussian, this process ensures that all returned weights are in the range (0, 1]. Zero is not included in the set, which means that this method will always produce an edge given any distance no matter how large. As such, pruning methods are extremely important for practical use.

Here, two options exist for edge pruning. The first is the width (variance) of the Gaussian. As the width parameter was 1, this method was not used at this time. These weights are instead pruned according to the weight threshold parameter via another defined function, *thresholding_weight_2*. This function takes only the weight matrix and the weight threshold and compares the entries in the matrix to the threshold; any entry below threshold is assigned a weight of 0. The matrix is then converted to a sparse matrix format (csr_matrix) (SciPy community, n.d. e), the diagonal elements of the matrix set at 1.0, to minimise machine error in the adjacency value for any voxel to itself. This adjacency matrix is saved as csv.

Finally, at the end of the loop for the *i*-th voxel, the number of voxels in the subgraph is compared to the maximum and minimum variables. If the number of voxels is larger than the maximum variable, the number is assigned to the variable, and vice versa for minimum. The number is also added to the average variable for later calculation. The mean and standard deviation for each feature for the *i*-th subgraph is calculated (np.mean and np.std, respectively,

both with *axis = 0* to collect data columnwise, Numpy Developers, n.d.) and assigned to a row of the stats 2D array. The feature matrix of the subgraph is also saved as csv. A print statement confirms the completion of creating the *i*-th subgraph and gives the time progression of the run.

4.4.2.2 Collecting Subgraph Statistics

Outside of the loop, once all subgraphs are created and saved to file, the set of feature means and standard deviations are each saved as separate csv files. Finally, the average variable, which has summed the total number of voxels in all subgraphs, is divided by the total number of voxels in the original feature matrix to calculate the average number of voxels per graph. A final runtime is printed out, then a summary of parameters and stats is created and saved to the run directory as a readme.txt file. These include: the run name; the source file name; the number of samples (dimensions of the MRI data); the total number of voxels and features; the radius; the maximum, minimum and average number of voxels per subgraph; and the run time.

4.4.3 Verification and Analysis

The package includes two alternative methods for creating both adjacency and distance matrices (by loop or using pdist and squareform from scipy.spatial.distance (SciPy community, n.d. d). They were both found to be less efficient or unable to process the full-sized toy data set. They were included in the code package for unit testing the less intuitive blackbox of cKDTree.sparse_distance_matrix.

There is also a defined function (*meanogram*) in the code package for graphing a histogram of the means or standard deviations distribution of features, which was not used in the script as it is computationally intensive and can be done separately from this subgraphing process when needed. This function requires the subgraph stats matrix, a variable indicating which layer of the matrix to use (*metric* = 0 or 1; means or standard deviations, respectively), the bin count for the histogram, and the directory for saving the picture file of the plot. This function was used in troubleshooting during development and retained for future use, if needed.

A final function is present for calculating the free energy surfaces (FES) of all combinations of features (see 5.1, fig. 23), called *free_energy_surface_allfeatures*. This analysis approach is related to concepts from statistical mechanics and thermodynamics but

can be generalised to probability theory (Crooks, 2007). The probability of a given state may be related to a generalised system-specific "free energy" F_{s_i} of that state:

$$p_{s_i} = \frac{e^{-\beta F_{s_i}}}{Z} \rightarrow F_{s_i} \sim -\frac{1}{\beta} \ln p_{s_i}$$

Where Z (the partition function) is a normalising constant and β is inversely related to the system "temperature." If the sampling of microstates is high enough to approach the underlying distribution of microstates, then the count of any given microstate is inversely related to the Helmholtz free energy of that state. The partition function essentially becomes counts of a given microstate, where a microstate here is a specific combination of MRI features. The higher the count, the lower the free energy and the more likely the combination of features.

free_energy_surface_allfeatures uses pyemmaplots.plot_free_energy (CMB-group, 2020) to show the relationship between pairs of features and the favourability of their combination. This function takes as inputs the feature matrix, the preferred directory location for saving the graph, the maximum and minimum free energies to be included in the plot (*vmax = 10 and vmin = 0*, respectively, as extreme default values for later visual adjustment), the number of histogram bins used in each dimension, and the width of the border around the subplots (*border = 1*). The output is a set of FES plots comparing each pair of features from the feature matrix input. The FES results used *vmin = 0, vmax = 6*, and *nbins = 1000*.

5. Results and Discussion

In this thesis, we developed a method to create physiological connectomes that could be used in the future for data augmentation. Results with toy data and are presented here to demonstrate proof of functionality of the method development. First is an overview of the downsampled version of the toy data and properties of the downsampled whole brain graphs. Next is an evaluation of the co-occurrence and possible codependence of pairs of MRI features as examined on the original, full resolution data, which is then compared to theoretical predictions. This is followed by an examination of the relationship between these trends of pairwise feature relationships and graph connectivity. Finally, a demonstration of these same pairwise features is presented in a sample subgraph, to verify the similarities and differences observed between the properties of the subgraph and those of the whole brain.

Theory predicts several trends in pairwise relationships between features. Mean diffusivity (MD), and axial diffusivity (AD) and radial diffusivity (RD) are all components of the same model, have shared variable dependencies, and are not independent (see section 2.2.2, fig. 5), (Tromp, 2015). As such, it would be expected for these features to correlate. Though they arise from separate models, ODI and FA should be inversely related (Zhang et al., 2012) as they describe opposite phenomena: isotropy and anisotropy, respectively. FA and MD relay similar information about diffusion (Alexander et al., 2011). ISOVF and ICVF represent extracellular and intracellular volume fraction, respectively. It would be expected that these demonstrate an inverse relationship. Since ODI is related to isotropic diffusion, it would likely positively correlate to ISOVF, as diffusion is less constrained in the extracellular space (Zhang et al., 2012). Similarly, FA should correlate to ICVF due to microstructure constraints increasing anisotropy (Schilling et al., 2023). Likewise, FA and AD should also correlate since AD represents diffusion along the long axis of the tensor (Tromp, 2023).

5.1 Whole Brain Graph

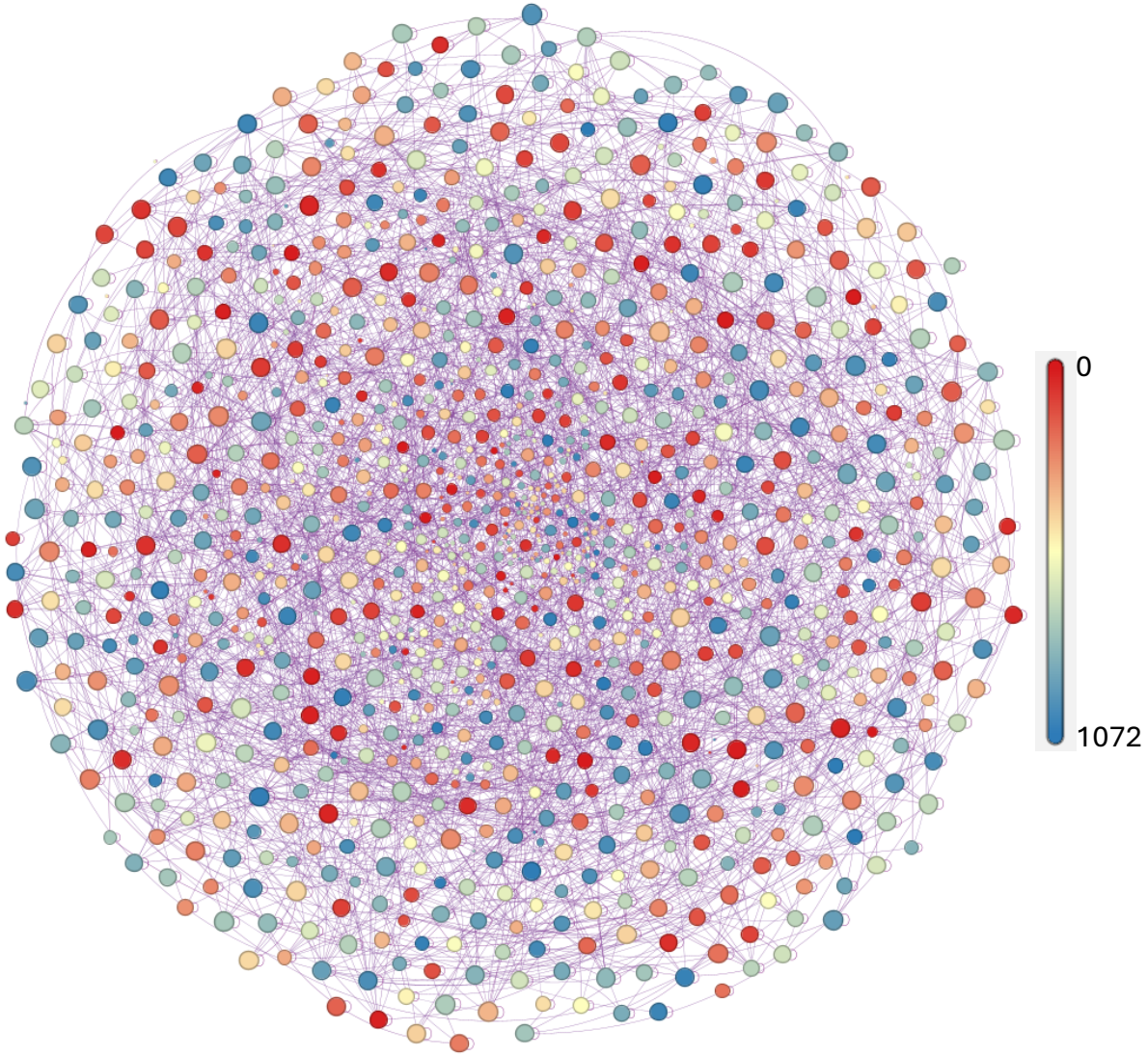

Figure 22: whole brain network graph of down-sampled white matter data (20 x 20 x 20 voxels; 1073 white matter voxels, 7 features) for demonstration and visualisation purposes. Node size correlates to connectivity (weighted degree). Edges were filtered during graph rendering to retain relationships of greatest similarity between nodes (top 2511 edges by edge weight retained). Node colour relates to the voxel ID number, based on voxel location in flattened data array. Note that a typical white matter cluster of this project's full resolution brain MRI would also be, on average, approximately this size.

This whole brain graph (fig. 22) shows that the voxel ID values are qualitatively dispersed randomly among all the nodes large enough to visualise. Since large node size relates to high degree, this suggests that connectivity and location are not strongly correlated, in terms of the flattened array. It is not possible to be sure of the location of the flattened array voxels

in 3D space, however, without converting the flattened array back to a 3D spatially organised array. Future work should investigate these relationships in image space.

This graph also illustrates the complexity of the network, even in downsampled data. Similarly, it is difficult to analyse full resolution subgraphs due to their complexity. As such, further work is needed to identify trends in this type of voluminous and highly complex dataset.

5.2 Co-occurrence of Pairwise Features

To obtain a more robust measure of relationships between features in the data, a set of free energy surface graphs comparing each pair of features in the non-downsampled data was created (fig. 23). These free energy surfaces were created to explore the co-incidence of various combinations of the different MRI metrics. Low free energy (the blue/purple colours) shows a higher likelihood of values for different metrics. White and red show less common metric combinations. The metrics values have been scaled and therefore the x and y labels cannot directly be interpreted in terms of their physical meaning, though the general relationship patterns are preserved. Future work should assess these relationships using unscaled data to preserve physical meaning.

The free energy surface is mostly consistent with the expected theoretical trends. FA and ODI show a negative correlation. This is expected as a high FA is indicative of a highly coherent diffusion direction, and therefore low dispersion of tracts. However, ICVF and ISOVF were expected to have a negative correlation but do not, demonstrating a larger variance in co-occuring values. This may be due to the complexity of the three-compartment model used to fit these parameters, and could be related to the myelin component not captured in these two compartments (Le Bihan et al., 2001). Since MD reflects the overall diffusivity including both axial and radial components, AD, MD and RD are expected be positively correlated. AD and MD do show a degree of positive correlation with a less homogenous set of values as shown by the dispersion of their FES. RD and MD show a minor correlation, whereas RD and AD have a consistent core of values with a wider FES spread for other co-occurances. While the co-dependence of RD and AD was lower than expected, this likely stems from the fact that in voxels with isotropic diffusion or crossing fibers, AD and RD would vary in the same direction, whereas in voxels with a more defined direction of diffusion, AD and RD would be inversely related. These two types of voxels are all represented here, showing a high degree of variance in the relationship between AD and RD depending on the composition of the voxel (Tromp, 2023). RD and FA, and AD and ODI, demonstrate an inverse relationship, for similar reasons to RD and AD demonstrating a negative correlation. Namely, high FA is indicative of voxels with a predominant diffusion direction, and therefore a lower radial diffusivity, while a high AD is also often associated with a dominant diffusion direction and therefore low dispersion.

RD and ODI are positively correlated, which complements RD and FA's negative correlation. Finally, ICVF and MD, and ICVF and RD both show a negative correlation. This is consistent with the interpretation of constrained diffusion within axons, so that a greater intacellular volume fraction is associated with an overall lower diffusivity (Zhang et al., 2012).

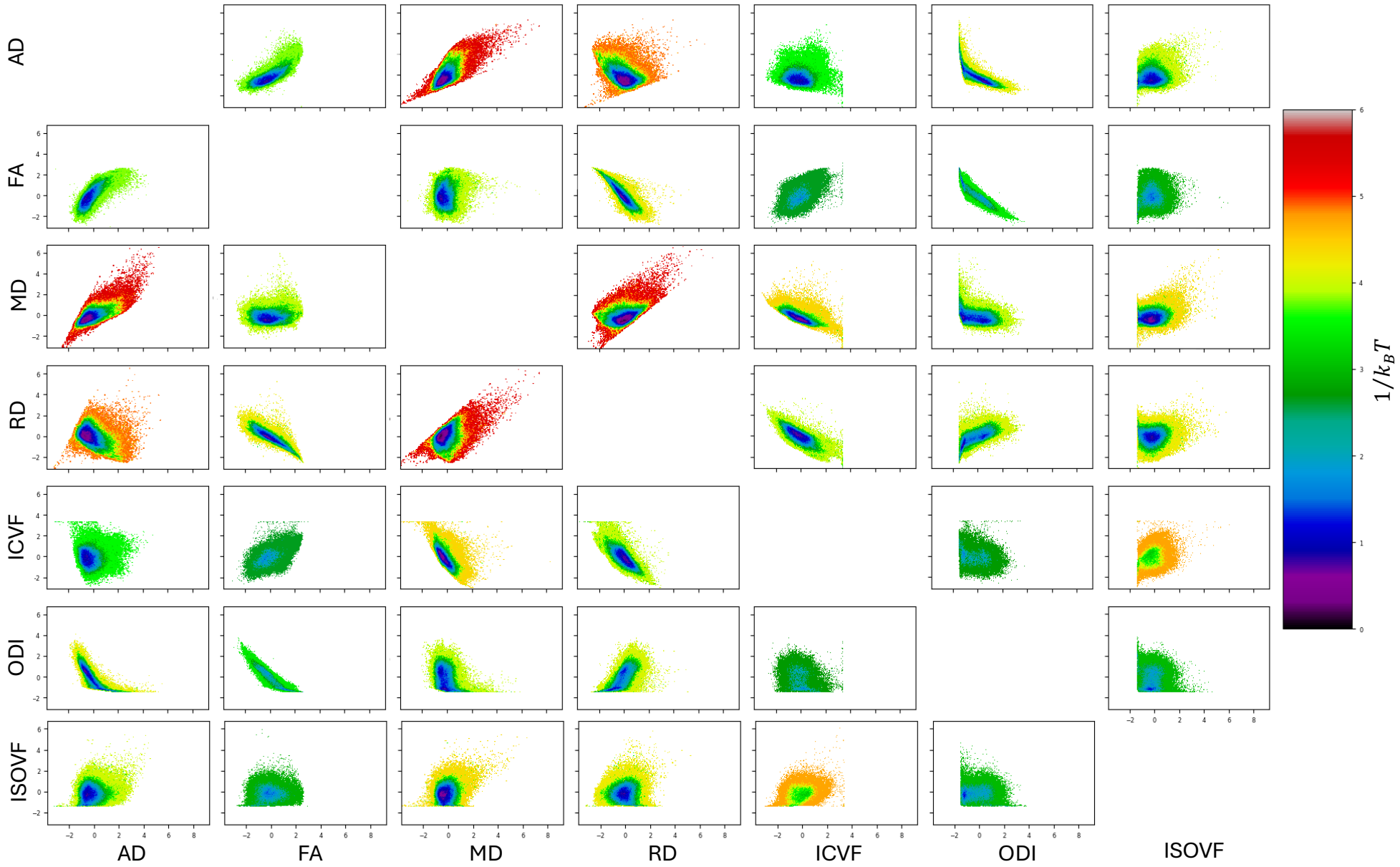

Figure 23: Free energy surfaces for full resolution brain MRI 359 938 white matter voxels, 100 voxel bins. Higher probability combinations of features have lower energy and vice versa. Note an inverse relationship between FA and ODI, between RD and FA, between AD and ODI, between ICVF and RD, and between ICVF and MD; a positive correlation between AD and MD, between ODI and RD, and between RD and MD. RD and AD, and ISOVF and ICVF do not demonstrate a clear trend.

5.3 The Relationship Between Node Connectivity and MRI metrics

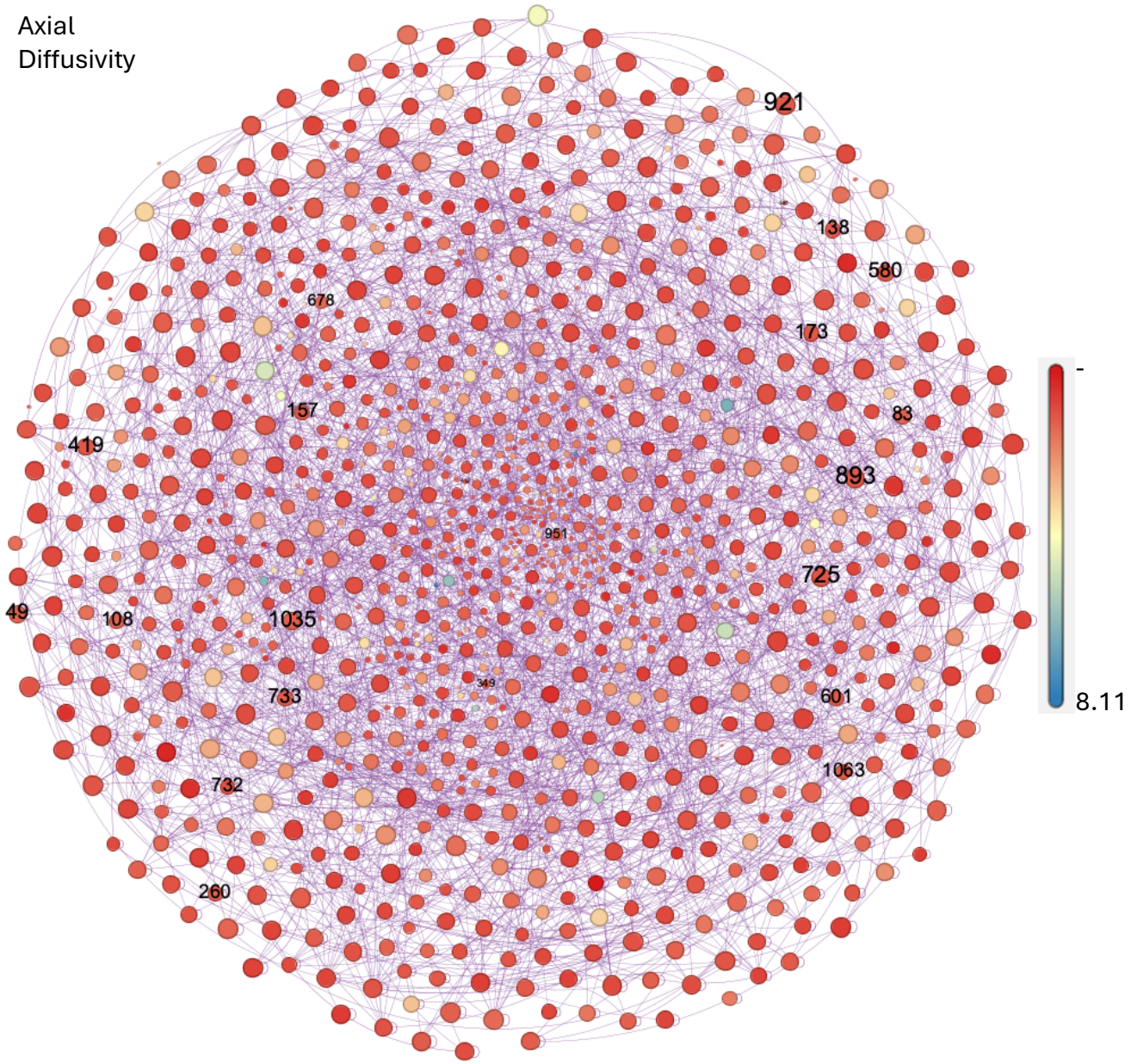

Figure 24: whole brain network graph of down-sampled white matter data (20 x 20 x 20 voxels; 1073 white matter voxels, 7 features) with node colour representing scaled values of axonal diffusion. Both mean diffusivity and radial diffusivity demonstrate similar characteristics. Node size correlates to connectivity (weighted degree). Edges filtered during graph rendering to retain relationships of greatest similarity between nodes (top 2511 edges by edge weight retained). Subgraph of voxel 138 nodes labeled.

In order to select a subgraph of interest, an understanding of the dataset as a whole is required. In this vein, we investigate the relationship between physiological connectivity—that

is, node degree in the whole-brain graph extracted from all features—and the individual node attributes, which are the MRI white-matter metrics. We then demonstrate our method by producing a local subgraph neighborhood demonstrating several of these relationships. In particular, the selected node had below average values for AD, MD, RD, and an inverse relationship between FA and ODI.

In fig. 24, we investigate the relationship between connectivity and axial diffusion. The values appear predominantly negative. Both mean diffusivity and radial diffusivity also appear predominantly negative when similarly analysed. The data, however, has been scaled to 0 mean and standard deviation of 1 as expected from StandardScaler and verified manually from the feature matrix. The apparent uniformity of this graph is likely because the positive values when visualised according to weighted degree are too small to be observed. As the larger nodes are more visible, the low values appear to be associated with higher degree and visually dominate the graph. Highly weighted edges are the strongest connections between voxels across all features and are of particular interest in this method. This dominance of negative AD in the more highly connected nodes would be consistent with greater metric similarities across voxels which are not characterized by a single diffusion direction. Accordingly, a voxel reflecting this below average value characteristic for RD, AD, and MD was selected and its associated subgraph visualised for examination.

The fractional anisotropy values (fig. 25) for high degree nodes appear highly non-homogenous with an average value likely near the FA mean. The orientation dispersion index values for high degree nodes appears non-homogenous with an average value moderately negative, as there is much more red-orange representation than blue-green. Neither FA nor ODI appear highly correlated with connectivity. As mentioned in the previous subsection, it is expected that FA is inversely related to ODI which grossly appears to be the case when viewing the data from this perspective.

The isotropic volume fraction (fig. 26) is the portion of the voxel which is estimated to be extracellular space (Zhang et al., 2012). Most of the highly connected voxels are below average values of ISOVF; that is, they are predominantly red, orange, or dark yellow. The values are less homogenous than the MD, AD, and RD graphs, suggesting a less strong correlation between ISOVF and connectivity. Considering that this data is taken from a healthy individual without a large amount of atrophy, it is expected that most voxels have a large partial volume of WM fibers. Since ISOVF is a measure of extracellular space, it is reasonable that ISOVF would be less highly connected than MD, RD, and AD.

Intracellular volume fraction (fig. 26) is the portion of the voxel which is estimated to be intracellular space (Zhang et al., 2012). Most of the highly connected voxels have ICVF values near the mean; that is, they are predominantly light yellow, with a notable presence of both light orange and light green.

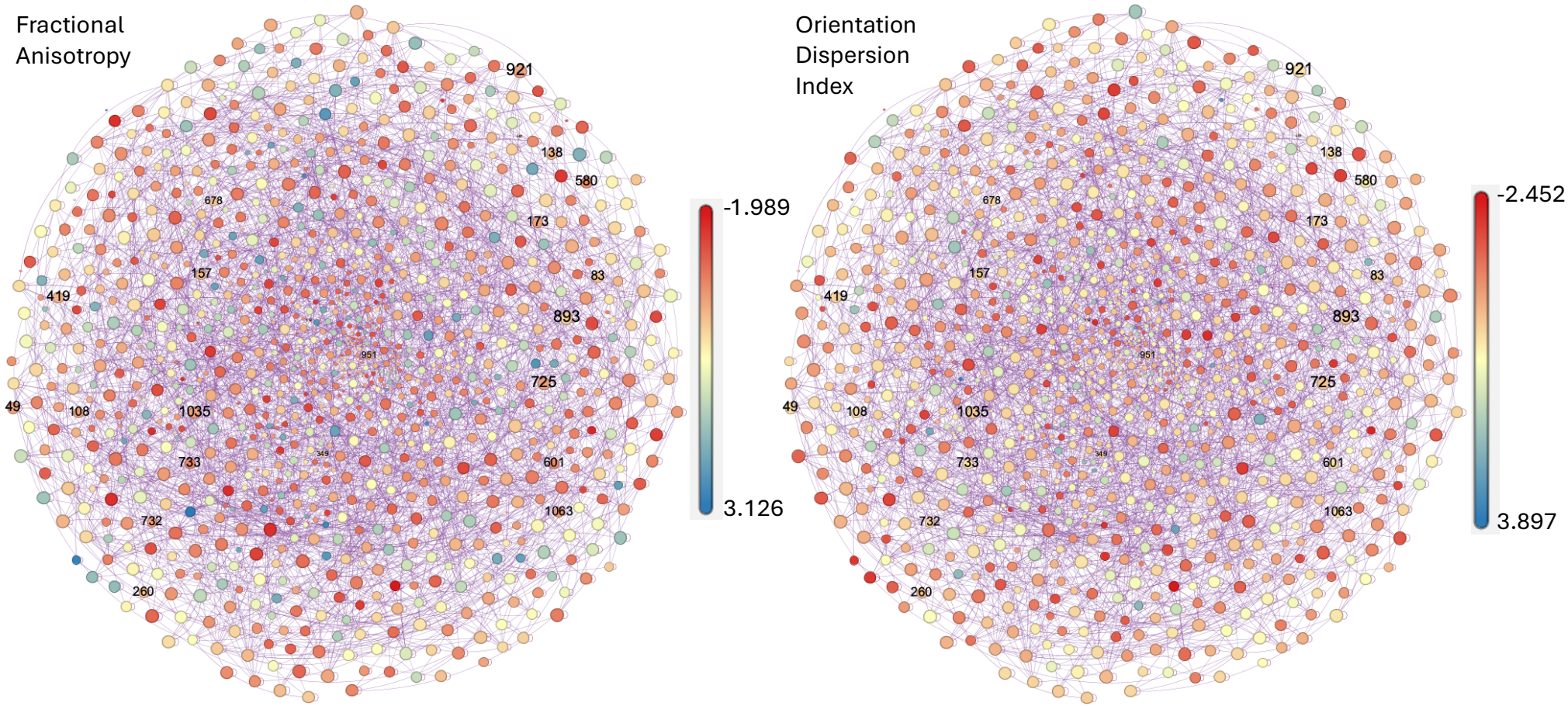

Figure 25: whole brain network graphs of down-sampled white matter data (20 x 20 x 20 voxels; 1073 white matter voxels, 7 features) with node colour representing scaled values of fractional anisotropy and orientation dispersion index as indicated. Node size correlates to connectivity (weighted degree). Edges filtered during graph rendering to retain relationships of greatest similarity between nodes (top 2511 edges by edge weight retained). Subgraph of voxel 138 nodes labeled.

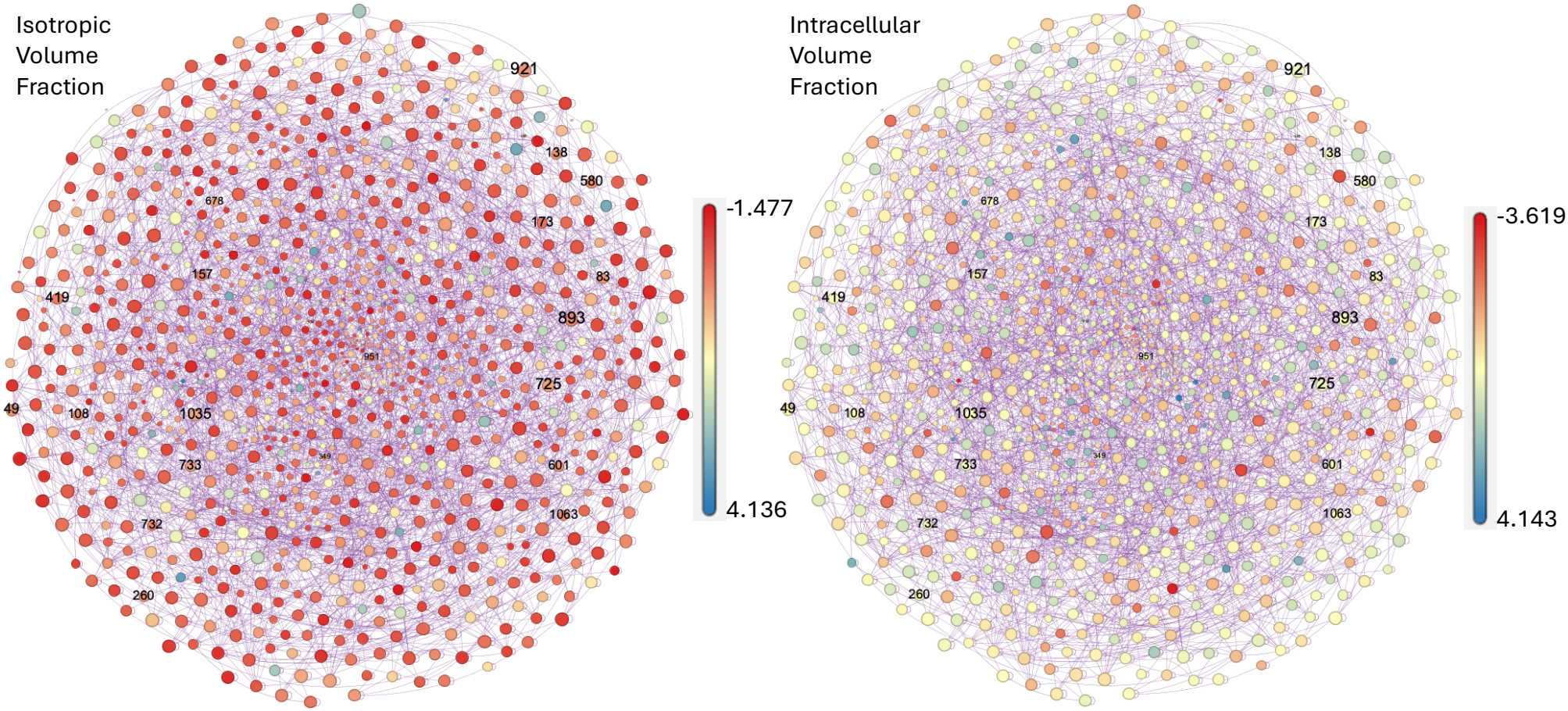

Figure 26: whole brain network graphs of down-sampled white matter data (20 x 20 x 20 voxels; 1073 white matter voxels, 7 features) with node colour representing scaled values of isotropic volume fraction and intracellular volume fraction as indicated. Node size correlates to connectivity (weighted degree). Edges filtered during graph rendering to retain relationships of greatest similarity between nodes (top 2511 edges by edge weight retained). Labelled nodes are those associated with the subgraph of voxel #138.

5.4 Demonstration of Local Neighbourhood Subgraph Production and Analysis

Index	AD	FA	MD	RD	ICVF	ODI	ISOVF
49	-0.102	-0.415	-0.009	0.056	0.055	0.329	0.046
83	-0.107	-0.083	-0.071	-0.044	0.093	0.150	0.084
108	-0.110	-0.456	-0.062	-0.027	0.224	0.215	-0.005
138	**-0.308**	**-0.229**	**-0.189**	**-0.101**	**0.454**	**0.354**	**0.041**
157	-0.264	-0.154	-0.201	-0.150	0.173	0.165	-0.171
159	-0.226	-0.304	-0.151	-0.095	0.018	0.188	-0.009
173	-0.341	-0.436	-0.354	-0.354	0.371	0.163	-0.345
260	-0.135	-0.106	-0.178	-0.204	0.451	-0.120	0.051
349	-0.312	-0.026	-0.246	-0.193	0.325	-0.085	0.000
412	-0.286	0.261	-0.211	-0.152	0.500	0.047	-0.093
419	-0.357	-0.279	-0.285	-0.226	0.186	0.092	-0.222
580	-0.285	-0.668	-0.094	0.043	0.158	0.509	0.031
581	-0.238	-0.403	-0.091	0.015	0.373	0.429	0.146
601	-0.271	-0.719	-0.326	-0.356	0.524	0.306	0.052
678	-0.024	-0.080	0.080	0.151	0.208	0.245	0.241
725	-0.410	-0.458	-0.334	-0.271	0.538	-0.059	0.231
732	-0.427	-0.481	-0.358	-0.299	0.559	0.122	-0.159
733	-0.211	-0.735	-0.083	0.009	0.577	0.325	0.206
893	-0.344	0.033	-0.246	-0.170	0.299	-0.059	-0.125
921	-0.327	-0.404	-0.284	-0.245	0.397	0.340	0.081
951	-0.141	-0.301	-0.081	-0.037	0.150	0.096	-0.156
1035	-0.156	-0.471	-0.091	-0.042	0.218	-0.015	0.284
1063	-0.291	-0.550	-0.187	-0.109	0.145	0.248	-0.064
MIN	-0.427	-0.735	-0.358	-0.356	0.018	-0.120	-0.345
MAX	-0.024	0.261	0.080	0.151	0.577	0.509	0.284
MEAN	-0.247	-0.324	-0.176	-0.122	0.304	0.173	0.006

Table 1: Voxel #138 nearest neighbours subgraph (r = 0.6, weight_threshold = 0.0), consisting of 23 nodes (average number of nodes per subgraph: 23.99) and 276 edges all of which are visualised.

As a demonstration of the subgraphs of the 1073 voxel dataset, voxel 138 subgraph was selected for discussion. It was selected in part because it had negative values for MD, AD and RD. Within the downsampled dataset, the highest degree nodes have a clear trend with respect to values of MD, RD, and AD: they tend to be small negative values, which is reflected here in this subgraph. Additionally, ODI is positive, and FA is a negative number, demonstrating the expected inverse relationship of these two features. Finally, ISOVF is a very small positive value, and ICVF is a larger positive value. This would be consistent with the most highly connected nodes (representing voxels with a higher similarity across metrics with other voxels,

irrespective of spatial location) being voxels with axons within them, but not completely made up of large axons, and for these axons to not have all the same orientation. While the volume fractions (ISOVF and ICVF) relationship does not follow the typically expected trend (the two values should be inverse), the voxel 138 subgraph offers an excellent opportunity to demonstrate the concepts involved in this method. To visualise the MRI features, the same data and graph configuration are presented with node colour reflecting the scaled values of a specified feature. It is also worthwhile to note that this subgraph also confirms that highly similar voxels are not necessarily located close together in space, indicating that this method successfully reflects physiological similarity rather than anatomic connectivity.

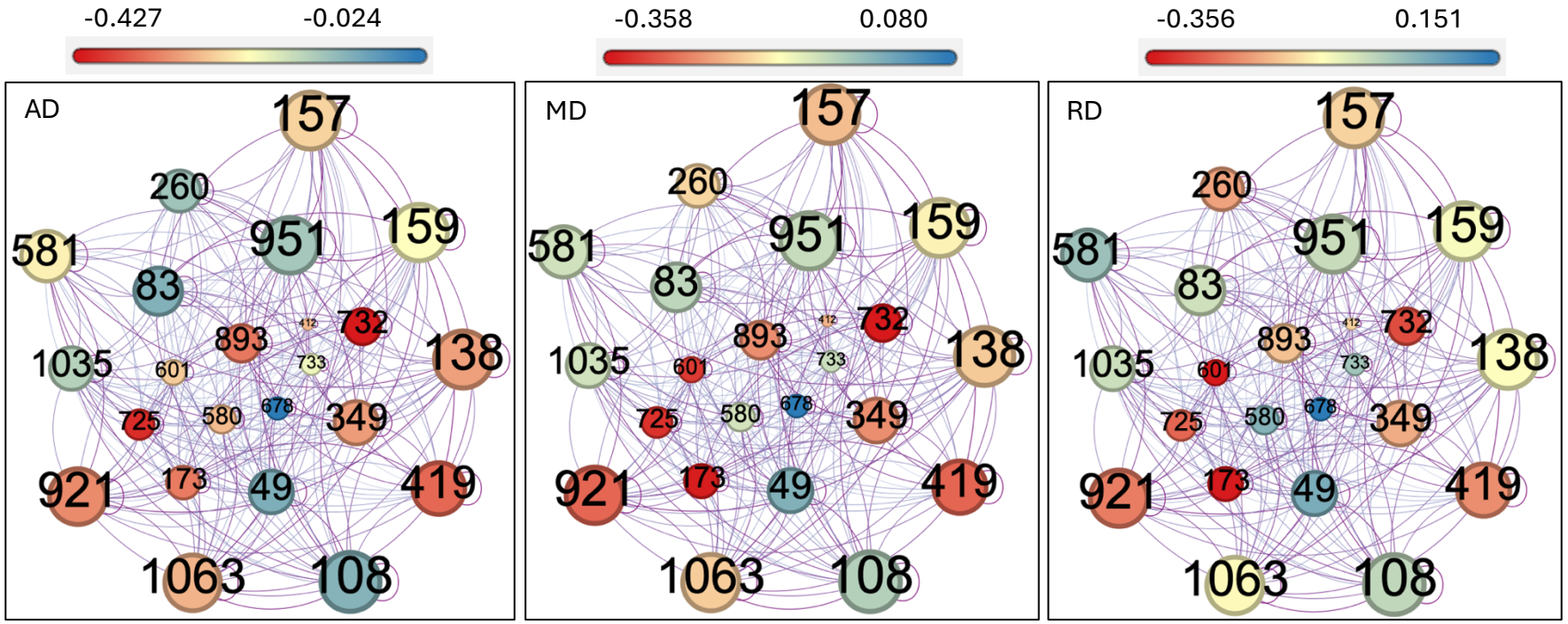

Figure 27: Voxel 138 subgraphs. The node size represents the weighted degree in terms of the subgraph connectivity. The labels are the indices by whole brain flattened voxel array locations. The colours describe the relative value of the features AD, MD, and RD, as labelled. The darkness of the edge colour indicates the weight of the edge.

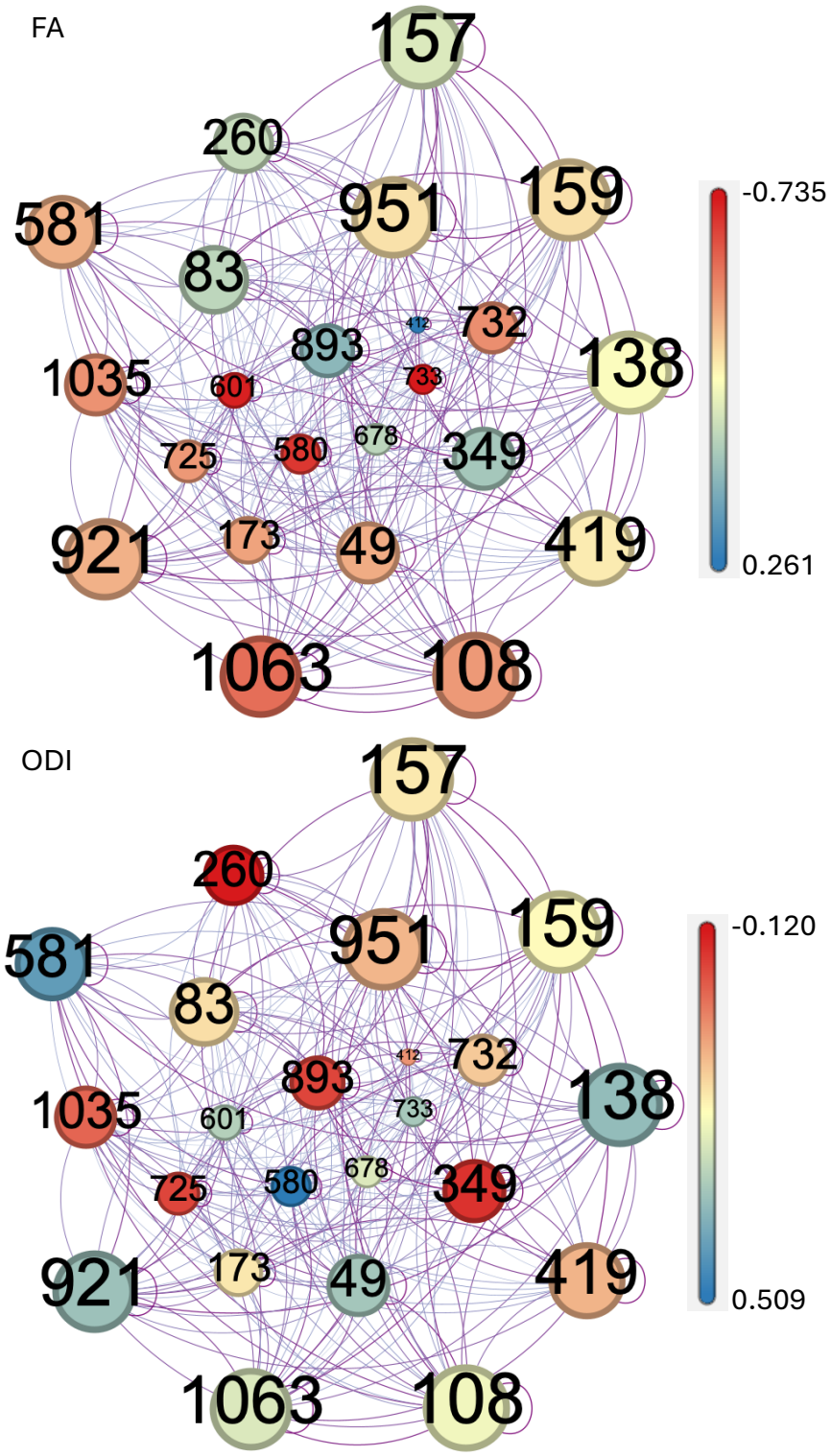

Figure 28: Voxel 138 subgraphs. The node size represents the weighted degree in terms of the subgraph connectivity. The labels are the indices by whole brain flattened voxel array locations. The colours describe the value of the features FA and ODI, as labelled. The darkness of the edge colour indicates the weight of the edge.

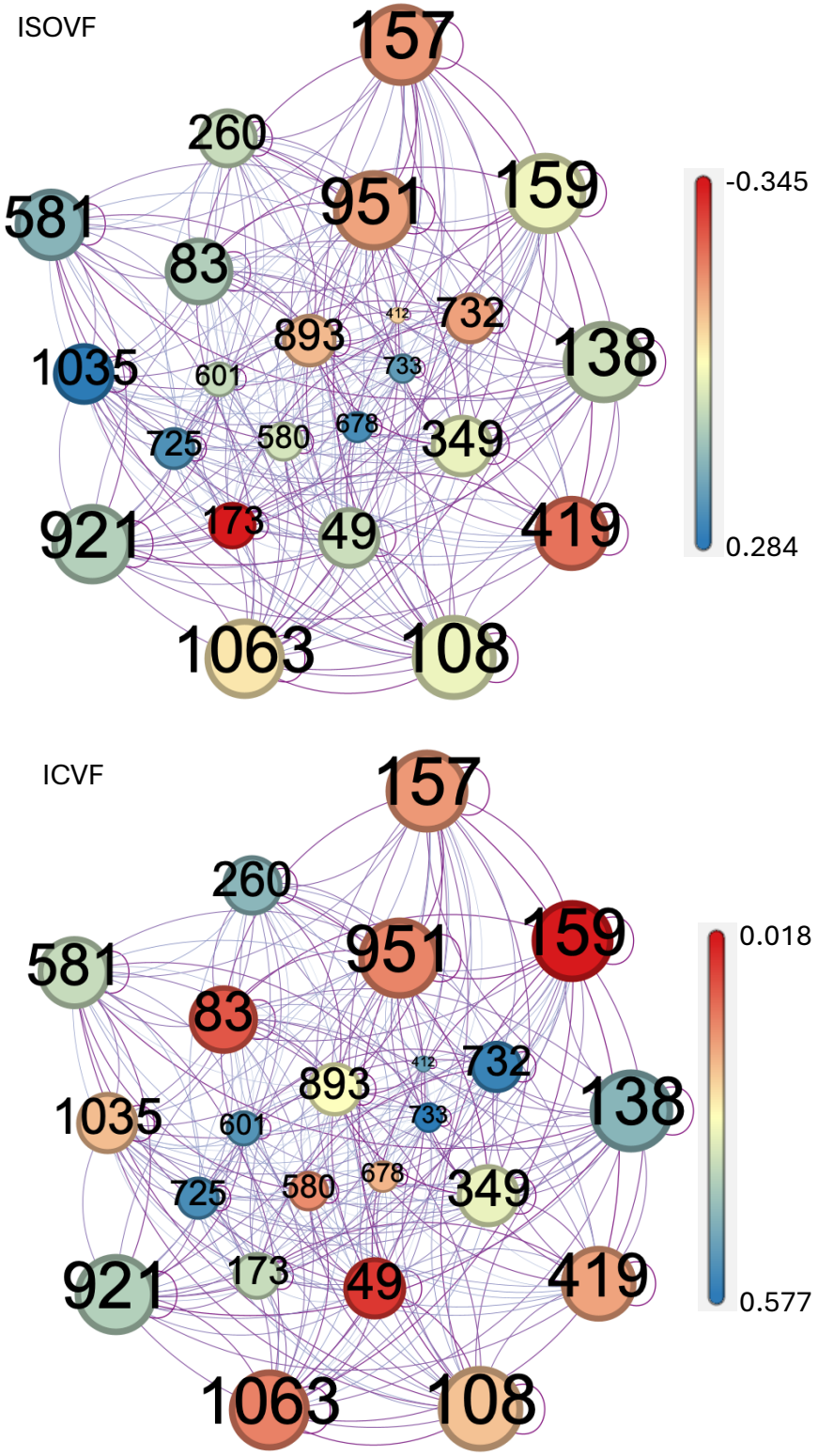

Figure 29: Voxel 138 subgraphs. The node size represents the weighted degree in terms of the subgraph connectivity. The labels are the indices by whole brain flattened voxel array locations. The colours describe the value of the features ISOVF and ICVF, as labelled. The darkness of the edge colour indicates the weight of the edge.

Voxel 138 subgraph consists mostly of relatively high degree voxel nodes (fig. 27-29) that demonstrate several of the expected characteristics associated with the highly weighted degree nodes of the whole graph (fig. 22). As noted in table 1, the mean values of AD, MD, and RD are negative, which aligns with the trend across large degree voxels in the whole brain graphs. In subgraph 138, AD, MD and RD generally correlate well, with few examples of variation between subgraph voxels (Fig 27), such as voxel 580 in MD and voxel 260 in AD. This is mostly to be expected from the FES plot (fig. 21), where MD correlates roughly to both AD and RD, though AD and RD don't show a strong relationship. With respect to FA and ODI, many voxels demonstrate the expected inverse relationship (260, 580, 581, 601, 893, 921, and 1063, to name several). This relationship is strongly supported by the FES plot, which shows a clear negative correlation between FA and ODI. Finally, both ISOVF and ICVF demonstrate the heterogenous makeup of values expected of high degree voxels according to whole graph observations (fig. 29). While some voxels do demonstrate inverse values between these two features (49, 159, 173, 580, 732, and 1035), many voxels show a similarity between values (157, 260, 419, 581, and 733). While this latter trend is inconsistent with theoretical expectations and likely reflects the complexity of the model used to estimate them and its three-compartment nature, it aligns with the FES plot, which demonstrates a lack of clear relationship between the two features. More work is needed to assess whether voxels without an inverse relationship have a high myelin content. Overall, we have demonstrated the production of a local neighborhood subgraph reflecting similar trends to the whole-brain data. This demonstrates the possibility of generalizing from small subgraphs to whole-brain data, which is crucial for performing data augmentation from local neighborhood data.

5.5 Challenges

One of the greatest challenges in developing this methodology has been data management. Nearly 360 000 voxels leading to an equal number of clusters is computationally demanding for both memory in processing and storage between steps in the method. To meet this challenge, verifying what data is important and what information may be discarded has been imperative. While further investigations into alternative visualisation options may aid in the understanding of the data (e.g. chord diagram or circle plots), it does not address the larger issue of data management.

Coupling pruning with sparse matrices has been the primary strategy for managing memory requirements. To this end, this method offers multiple opportunities and methods for edge pruning. As the formula for converting from a distance matrix to an adjacency matrix is Gaussian, it does not produce edges of weight 0 leading to a complete graph, where every node is connected to every other node by an edge. Without pruning, the adjacency matrices would

be fully populated, producing an object of a memory-prohibitive size. By pruning low-weighted edges, the size of the matrices can be managed well with sparse matrices.

Using sparse matrices for organizing data decreases memory requirements as only non-zero values occupy storage. Since below a user-defined threshold, pruning assigns edges a value of 0, a large portion of the adjacency graphs are occupied by zeros. In using sparse matrices, these zeros do not add to the strain of this methodology on memory requirements.

Another strategy employed for decreasing memory requirements is to save matrices to storage. This frees up memory, as once they are saved in storage, they are overwritten in the memory. This introduces a new challenge: storage requirements. Even when using only simple text-based formats for file storage, the large number of subgraphs rapidly occupies significant portions of memory. The number of files can even be an issue, as many supercomputers have strict limits for the number of files allowed in storage. One strategy being developed for this difficulty was to stack adjacency matrices such that each file has multiple adjacency matrices saved to storage in a 3D array. This method would allow for saving batches of subgraphs to storage, however, until they are stored, the subgraphs occupy memory. Thus, it is an optimization problem. Maximise the number of adjacency matrices in memory before storing the set of matrices and removing them from memory.

All of these steps, however, still leave a precarious draw on computational resources when considering that there are nearly 360 000 voxels in the white matter alone in a single brain. This process will also need to be repeated in grey matter voxels to have a complete overview of brain changes in patient populations. As such, a final strategy may require decreasing the resolution of the MRI data to decrease the total number of voxels. The noise introduced by this approach could be at least partially mitigated through employing established downsampling methods developed to decrease resolution without confounding data by avoiding the combination of spatially adjacent voxels with very different functions or partial volume compositions (Huo et al., 2018). This diminishes the amount of data available for subsequent steps in the methodology but may be a necessary step without major cost when considering the magnitude of data involved.

A final option for data management is reconsidering the approach to the subgraphing process itself. While making use of a dimension reduction technique such as PCA or ICA would almost certainly improve the computation time of the distance matrix, it would not decrease the size of the distance and adjacency matrices. Exploring an alternative nearest neighbours method which creates true clusters – where each node is a member of only one cluster – could decrease redundancy in the subgraphs and also in the volume of data. However, the redundancy may have value in reinforcing particularly important connections in the connectome by increasing its representation in the dataset, which would be lost in using true clusters.

6. Conclusions and Future Work

In this thesis, we set out to lay the groundwork for defining physiological connectomes which can form the basis for data augmentation approaches. This novel methodological approach to brain MRI analysis has been successfully applied to toy data to create neighborhood subgraphs of the physiological connectome. While these early results are encouraging regarding the functionality of the method, much more work will be needed before the method reaches the full potential of its applications. In the short term, as discussed in the previous section, further work is required to organise and manage the volume of data involved in this project. A formal calculation on the error rate and error propagation in the feature values and edges would be helpful to verify which edges are valid measures and which are within the error range, reflecting noise in the data. This information could help guide edge pruning in future work. A rigorous comparison of neighborhood size with error properties would also be useful. As the project unfolds, it should become more clear what size and composition of subgraphs is best suited to the analysis methods suggested by this project, possibly allowing for more aggressive pruning and a more efficient methodology. Additional statistical analysis to verify the subgraphs and compare to the whole could be insightful. It may also be useful to restore the feature data to its unscaled values to assess any correlation with 3D spatial coordinates.

In the future, after working through the data management concerns, the next phase of the project is to proceed towards data augmentation and topological analysis (fig. 30).

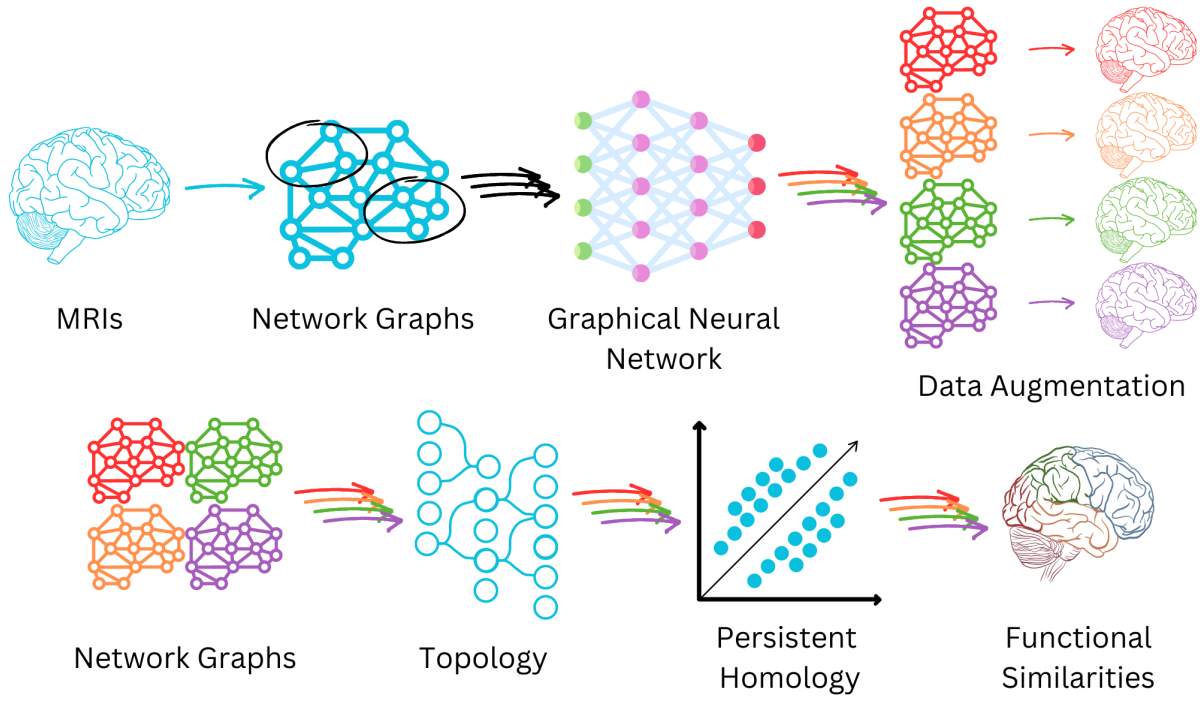

Figure 30: graphical abstract of project overview.

This thesis has addressed the first step in the visualized methodological pipeline, creating subgraphs of brain MRI voxels. Subsequent steps in this project will use these subgraphs to train a graphical neural network for the purposes of using generative deep learning for data augmentation purposes. The trained model will be used to produce 'hypothetical brains' – synthetic brain MRI data consistent with the (local) trends of acquired data without directly reproducing the original dataset. The synthetic data will then be topologically analysed using a technique called persistent homology (Naitzat et al., 2020), the results of which should give insight into the most important connections in the brains of people suffering from long covid. The process will be repeated using control brain MRIs to determine which of these connections are unique to the pathological changes associated with long covid.

Generative deep learning (GDL) uses deep neural networks to create a model for the underlying probability distribution of the properties of the training data (Pinaya et al., 2022). This distribution can then be sampled from to generate synthetic data that resembles the original data to perform data augmentation (Pinaya et al., 2022). To our knowledge, the use of GDL to create synthetic data mimicking the brain physiological connectome is a novel application. Using the PyTorch package for deep learning (Paszke et al., 2019), deep learning models will be trained on brain network subgraphs produced using the methodology of this thesis. The large quantity of local subgraphs will provide sufficient data to train notoriously data-hungry deep learning models. The accuracy of the deep learning model in learning these properties will be tested by using the models to synthesise local subgraph properties of a "hypothetical brain" which can be compared to a hold-out test set of brain graphs.

The next step in the project will be to analyse the properties demonstrated by the generated hypothetical brain network graphs. The network graphs will be modelled as flow networks (Rocks et al., 2020). Persistent homology will then be applied to the biological flow network model. Persistent homology identifies functional similarities between heterogenous, topologically structured systems, such as flow networks modelling the cerebral vasculature of multiple brains (Rocks et al., 2020). The use of persistent homology should allow for identification of common functional similarities between sets of diverse brains. The network will be further analysed by adapting existing functional flow network methods (Rocks et al., 2020). Finally, by comparing the findings of these analyses to control, the functional changes associated with long covid may be identified, offering both a basis for future investigations into specific brain regions, and informing research into targeted treatment options.

In conclusion, this thesis forms the basis for a novel approach to modeling and analysing brain imaging data with the intent that the approach be generalizable to other complex biological systems.

Bibliography

Ajčević, M., Iscra, K., Furlanis, G., Michelutti, M., Miladinović, A., Buoite Stella, A., Ukmar, M., Assunta Cova, M., Accardo, A., & Manganotti, P. (2023). Cerebral hypoperfusion in post-COVID-19 cognitively impaired subjects revealed by arterial spin labeling MRI. *Scientific Reports*, *13*(1), 5808. https://doi.org/10.1038/s41598-023-32275-3

Albert Einstein College of Medicine. (2014, Sep 23). Introducing MRI: Arterial spin labeling (54 of 56) [Video]. YouTube. https://www.youtube.com/watch?v=o6gqqeDfKGM

Alexander, A. L., Hasan, K., Kindlmann, G., Parker, D. L., & Tsuruda, J. S. (2000). A geometric analysis of diffusion tensor measurements of the human brain. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *44*(2), 283-291. https://doi.org/10.1002/1522-2594(200008)44:2%3C283::aid-mrm16%3E3.0.co;2-v

Alexander, A. L., Hurley, S. A., Samsonov, A. A., Adluru, N., Hosseinbor, A. P., Mossahebi, P., Tromp, D. P. M., Zakszewski, E., & Field, A. S. (2011). Characterization of cerebral white matter properties using quantitative magnetic resonance imaging stains. *Brain Connectivity*, *1*(6), 423-446. https://doi.org/10.1089/brain.2011.0071

Alexander, A. L., Tsuruda, J. S., & Parker, D. L. (1997). Elimination of eddy current artifacts in diffusion-weighted echo-planar images: The use of bipolar gradients. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 38(6), 1016-1021. https://doi.org/10.1002/mrm.1910380623

Asllani, I., Borogovac, A., & Brown, T. R. (2008). Regression algorithm correcting for partial volume effects in arterial spin labeling MRI. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *60*(6), 1362-1371. https://doi.org/10.1002/mrm.21670

Bastian, M., Heymann, S., & Jacomy, M. (2009, March). Gephi: An open source software for exploring and manipulating networks. In *Proceedings of the international AAAI conference on web and social media* (Vol. 3, No. 1, pp. 361-362). https://doi.org/10.1609/icwsm.v3i1.13937

Bazin, P. L., Plessis, V., Fan, A. P., Villringer, A., & Gauthier, C. J. (2016, April). Vessel segmentation from quantitative susceptibility maps for local oxygenation venography. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)* (pp. 1135-1138). IEEE. https://doi.org/10.1109/ISBI.2016.7493466

Bentley, J. L. (1975). *A survey of techniques for fixed radius near neighbor searching*. Stanford University. Retrieved June 10, 2024, from https://exhibits.stanford.edu/stanford-pubs/catalog/bh705ys8528

Betzel, R. F., & Bassett, D. S. (2017). Generative models for network neuroscience: Prospects and promise. *Journal of The Royal Society Interface*, *14*(136), 20170623. https://doi.org/10.1098/rsif.2017.0623

Biondetti, E., Cho, J., & Lee, H. (2023). Cerebral oxygen metabolism from MRI susceptibility. *Neuroimage*, *276*, 120189. https://doi.org/10.1016/j.neuroimage.2023.120189

Bren, K. L., Eisenberg, R., & Gray, H. B. (2015). Discovery of the magnetic behavior of hemoglobin: A beginning of bioinorganic chemistry. *Proceedings of the National Academy of Sciences*, *112*(43), 13123-13127. https://doi.org/10.1073/pnas.1515704112

Brett, M., Markiewicz, C. J., Hanke, M., Côté, M.-A., Cipollini, B., McCarthy, P., Jarecka, D., Cheng, C. P., Larson, E., Halchenko, Y. O., Cottaar, M., Ghosh, S., Wassermann, D., Gerhard, S., Lee, G. R., Baratz, Z., Wang, H.-T., Papadopoulos Orfanos, D., Kastman, E., … freec84. (2024). nipy/nibabel: 5.2.1 (5.2.1). Zenodo. https://doi.org/10.5281/zenodo.10714563

Brown, R. (1828). XXVII. A brief account of microscopical observations made in the months of June, July and August 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. *The Philosophical Magazine*, *4*(21), 161-173.

Bullmore, E., & Sporns, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3), 186-198. https://doi.org/10.1038/nrn2575

Buxton, R. B., Frank, L. R., Wong, E. C., Siewert, B., Warach, S., & Edelman, R. R. (1998). A general kinetic model for quantitative perfusion imaging with arterial spin labeling. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *40*(3), 383-396. https://doi.org/10.1002/mrm.1910400308

Carfì, A., Bernabei, R., & Landi, F. (2020). Persistent symptoms in patients after acute COVID-19. *Jama*, *324*(6), 603-605.

Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., & Haworth, A. (2021). A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 65(5), 545-563. https://doi.org/10.1111/1754-9485.13261

CMB-group. (2020). pyemma.plots.plot_free_energy. *PyEMMA 2.5.7 documentation.* Retrieved June 19, 2024, from http://www.emma-project.org/latest/api/generated/pyemma.plots.plot_free_energy.html

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21-27. https://doi.org/10.1109/TIT.1967.1053964

Crooks, G. E. (2007). Measuring thermodynamic length. *Physical Review Letters*, 99(10), 100602. https://doi.org/10.1103/PhysRevLett.99.100602

De Boer, R. W. (1995). Magnetization transfer contrast. *Part*, *1*, 64-73.

Deng, W., Zhang, B., Zou, W., Zhang, X., Cheng, X., Guan, L., Lin, Y., Lao, G., Ye, B., Li, X., Yang, C., Ning, Y., & Cao, L. (2019). Abnormal degree centrality associated with cognitive dysfunctions in early bipolar disorder. *Frontiers in Psychiatry*, *10*, 140. https://doi.org/10.3389/fpsyt.2019.00140

Descoteaux, M. (1999). High angular resolution diffusion imaging (HARDI). *Wiley Encyclopedia of Electrical and Electronics Engineering*, 1-25. https://doi.org/10.1002/047134608X.W8258

Detre, J. A., Rao, H., Wang, D. J., Chen, Y. F., & Wang, Z. (2012). Applications of arterial spin labeled MRI in the brain. *Journal of Magnetic Resonance Imaging*, 35(5), 1026-1037. https://doi.org/10.1002/jmri.23581

Douaud, G., Lee, S., Alfaro-Almagro, F., Arthofer, C., Wang, C., McCarthy, P., Lange, F., Andersson, J. L. R., Griffanti, L., Duff, E., Jbabdi, S., Taschler, B., Keating, P., Winkler, A. M., Collins, R., Matthews, P. M., Allen, N., Miller, K. L., Nichols, T. E., & Smith, S. M. (2022). SARS-CoV-2 is associated with changes in brain structure in UK Biobank. *Nature*, *604*(7907), 697-707.

Duyn, J. (2013). MR susceptibility imaging. *Journal of Magnetic Resonance*, *229*, 198-207. https://doi.org/10.1016/j.jmr.2012.11.01

Duyn, J. H., & Schenck, J. (2017). Contributions to magnetic susceptibility of brain tissue. *NMR in Biomedicine*, *30*(4), e3546. https://doi.org/10.1002/nbm.3546

Dymerska, B., Eckstein, K., Bachrata, B., Siow, B., Trattnig, S., Shmueli, K., & Robinson, S. D. (2021). Phase unwrapping with a rapid opensource minimum spanning tree algorithm (ROMEO). *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *85*(4), 2294-2308. https://doi.org/10.1002/mrm.28563

Educative. (n.d.). *Sparse matrices in Python*. Retrieved June 17, 2024, from https://www.educative.io/answers/sparse-matrices-in-python

Edzes, H. T., & Samulski, E. T. (1977). Cross relaxation and spin diffusion in the proton NMR of hydrated collagen. *Nature*, *265*(5594), 521-523. https://doi.org/10.1038/265521a0

Einstein, A. (1905). über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen der Physik*, *322*(8), 549-560.

Elster, A. (n.d. a). *T1 Relaxation: Definition*. Questions and Answers in MRI. Retrieved May 29, 2024, from https://mriquestions.com/what-is-t1.html

Elster, A. (n.d. b). *T2 Relaxation: Definition*. Questions and Answers in MRI. Retrieved May 29, 2024, from https://mriquestions.com/what-is-t2.html

Elster, A. (n.d. c). *T2 vs T2*.* Questions and Answers in MRI. Retrieved May 29, 2024, from https://mriquestions.com/t2-vs-t2.html

Elster, A. (n.d. d). *Free Induction Decay.* Questions and Answers in MRI. Retrieved May 29, 2024, from https://mriquestions.com/free-induction-decay.html

Elster, A. (n.d. e). *Gradient Echo (GRE).* Questions and Answers in MRI. Retrieved May 20, 2024, from https://mriquestions.com/gradient-echo.html

Fan, A. P., Bilgic, B., Gagnon, L., Witzel, T., Bhat, H., Rosen, B. R., & Adalsteinsson, E. (2014). Quantitative oxygenation venography from MRI phase. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *72*(1), 149-159. https://doi.org/10.1002/mrm.24918

Fan, A. P., Khalil, A. A., Fiebach, J. B., Zaharchuk, G., Villringer, A., Villringer, K., & Gauthier, C. J. (2020). Elevated brain oxygen extraction fraction measured by MRI susceptibility relates to perfusion status in acute ischemic stroke. *Journal of Cerebral Blood Flow & Metabolism*, *40*(3), 539-551. https://doi.org/10.1177/0271678X198279

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). “Mini-mental state”: a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189-198. https://doi.org/10.1016/0022-3956(75)90026-6

Fornito, A., Zalesky, A., & Breakspear, M. (2013). Graph analysis of the human connectome: Promise, progress, and pitfalls. *Neuroimage*, *80*, 426-444. https://doi.org/10.1016/j.neuroimage.2013.04.087

Fox, R. J., Beall, E., Bhattacharyya, P., Chen, J. T., & Sakaie, K. (2011). Advanced MRI in multiple sclerosis: Current status and future challenges. *Neurologic Clinics*, *29*(2), 357-380. https://doi.org/10.1016/j.ncl.2010.12.011

Giunta, S., Giordani, C., De Luca, M., & Olivieri, F. (2024). Long-COVID-19 autonomic dysfunction: an integrated view in the framework of inflammaging. *Mechanisms of Ageing and Development*, 111915.

Gonçalves, N., Nikkilä, J., & Vigário, R. (2009). Partial clustering for tissue segmentation in MRI. In *Advances in Neuro-Information Processing: 15th International Conference, ICONIP 2008, Auckland, New Zealand, November 25-28, 2008, Revised Selected Papers, Part II 15* (pp. 559-566). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-03040-6_68

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Grade, M., Hernandez Tamames, J. A., Pizzini, F. B., Achten, E., Golay, X., & Smits, M. (2015). A neuroradiologist's guide to arterial spin labeling MRI in clinical practice. *Neuroradiology*, *57*, 1181-1202. https://doi.org/10.1007/s00234-015-1571-z

Guillén, N., Pérez-Millan, A., Falgàs, N., Lledó-Ibáñez, G. M., Rami, L., Sarto, J., Botí, M., Arnaldos-Pérez, C., Ruiz-García, R., Naranjo, L., Segura, B., Balasa, M., Sala-Llonch, R., Lladó, A., Gray, S. M., Johannesen, J. K., Pantoni, M. M., Rutledge, G. A., Sawant, R., ... Sanchez-Valle, R. (2024). Cognitive profile, neuroimaging and fluid biomarkers in post-acute COVID-19 syndrome. *Scientific Reports*, *14*(1), 12927.

Haacke, E. M., Liu, S., Buch, S., Zheng, W., Wu, D., & Ye, Y. (2015). Quantitative susceptibility mapping: Current status and future directions. *Magnetic Resonance Imaging*, *33*(1), 1-25. https://doi.org/10.1016/j.mri.2014.09.004

Haller, S., Zaharchuk, G., Thomas, D. L., Lovblad, K. O., Barkhof, F., & Golay, X. (2016). Arterial spin labeling perfusion of the brain: emerging clinical applications. *Radiology*, *281*(2), 337-356.

Hasan, K. M., Alexander, A. L., & Narayana, P. A. (2004). Does fractional anisotropy have better noise immunity characteristics than relative anisotropy in diffusion tensor MRI? An analytical approach. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *51*(2), 413-417. https://doi.org/10.1002/mrm.10682

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

Havran, Vlastimil. (2016). [Lecture slides]. Data Structures for Computer Graphics: point based representations and data structures. Retrieved June 10, 2024 from https://slideplayer.com/slide/6318865/

Helms, G., Dathe, H., Kallenberg, K., & Dechent, P. (2008). High-resolution maps of magnetization transfer with inherent correction for RF inhomogeneity and T1 relaxation obtained from 3D FLASH MRI. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *60*(6), 1396-1407. https://doi.org/10.1002/mrm.21732

Henkelman, R. M., Stanisz, G. J., & Graham, S. J. (2001). Magnetization transfer in MRI: a review. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo*, *14*(2), 57-64. https://doi.org/10.1002/nbm.683

Hidalgo-Tobon, S. S. (2010). Theory of gradient coil design methods for magnetic resonance imaging. *Concepts in Magnetic Resonance Part A*, *36*(4), 223-242. https://doi.org/10.1002/cmr.a.20163

Horowitz, T., Dudouet, P., Campion, J. Y., Kaphan, E., Radulesco, T., Gonzalez, S., Cammilleri, S., Ménard, A., & Guedj, E. (2024). Persistent brain metabolic impairment in long COVID patients with persistent clinical symptoms: a nine-month follow-up [18F] FDG-PET study. *European Journal of Nuclear Medicine and Molecular Imaging*, 1-8. https://doi.org/10.1007/s00259-024-06775-x

Hosp, J. A., Reisert, M., Dressing, A., Götz, V., Kellner, E., Mast, H., Arndt, S., Waller, C. F., Wagner, D., Rieg, S., Urbach, H., Weiller, C., Schröter, N., & Rau, A. (2024). Cerebral microstructural alterations in Post-COVID-condition are related to cognitive impairment, olfactory dysfunction and fatigue. *Nature Communications*, *15*(1), 4256. https://doi.org/10.1038/s41467-024-48651-0

Hou, X., Guo, P., Wang, P., Liu, P., Lin, D. D. M., Fan, H., Li, Y., Wei, Z., Lin, Z., Jiang, D., Jin, J., Kelly, C., Pillai, J. J., Huang, J., Pinho, M. C., Thomas, B. P., Welch, B. G., Park, D. C., Patel, V. M., ... Lu, H. (2023). Deep-learning-enabled brain hemodynamic mapping using resting-state fMRI. *npj Digital Medicine*, *6*(1), 116. https://doi.org/10.1038/s41746-023-00859-y

Huck, J., Wanner, Y., Fan, A. P., Jäger, A. T., Grahl, S., Schneider, U., Villringer, A., Steele, C. J., Tardif, C. L., Bazin, P. L., & Gauthier, C. J. (2019). High resolution atlas of the venous brain vasculature from 7 T quantitative susceptibility maps. *Brain Structure and Function*, 224, 2467-2485. https://doi.org/10.1016/j.cag.2011.02.001

Huisman, T. A. G. M. (2010). Diffusion-weighted and diffusion tensor imaging of the brain, made easy. *Cancer Imaging*, *10*(1A), S163. https://doi.org/10.1102/1470-7330.2010.9023

Huo, J., Wu, J., Cao, J., & Wang, G. (2018). Supervoxel based method for multi-atlas segmentation of brain MR images. *NeuroImage*, *175*, 201-214. https://doi.org/10.1016/j.neuroimage.2018.04.001

Intzandt, B., Sabra, D., Foster, C., Desjardins-Crépeau, L., Hoge, R. D., Steele, C. J., Bherer, L., & Gauthier, C. J. (2020). Higher cardiovascular fitness level is associated with lower cerebrovascular reactivity and perfusion in healthy older adults. *Journal of Cerebral Blood Flow & Metabolism*, *40*(7), 1468-1481. https://doi.org/10.1177/0271678X198628

Jenkinson, M., Pechaud, M., & Smith, S. (2005). BET2: MR-based estimation of brain, skull and scalp surfaces. In *Eleventh Annual Meeting of the Organization for Human Brain Mapping, 2005*.

Jespersen, S. N., Leigland, L. A., Cornea, A., & Kroenke, C. D. (2011). Determination of axonal and dendritic orientation distributions within the developing cerebral cortex by diffusion tensor imaging. *IEEE Transactions on Medical Imaging*, *31*(1), 16-32. https://doi.org/10.1109/TMI.2011.2162099

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. http://doi.org/10.1098/rsta.2015.0202

Jones, J., Murphy, A., Bell, D., Gaillard, F., & Di Muzio, B. (2021, Sep 19). *Net magnetisation vector*. Radiopedia.org. https://doi.org/10.53347/rID-6316

Kamli, A., Saouli, R., Batatia, H., Naceur, M. B., & Youkana, I. (2020). Synthetic medical image generator for data augmentation and anonymisation based on generative adversarial network for glioblastoma tumors growth prediction. *IET Image Processing*, *14*(16), 4248-4257. https://doi.org/10.1049/iet-ipr.2020.1141

Karakuzu, A., Boudreau, M., Duval, T., Boshkovski, T., Leppert, I., Cabana, J. F., Gagnon, I., Beliveau, P., Pike, G. B., Cohen-Adad, J., & Stikov, N. (2020). qMRLab: Quantitative MRI analysis, under one umbrella. *Journal of Open Source Software*, *5*(53). https://doi.org/10.21105/joss.02343

Kazemifar, S., McGuire, S., Timmerman, R., Wardak, Z., Nguyen, D., Park, Y., Jiang, S., & Owrangi, A. (2019). MRI-only brain radiotherapy: Assessing the dosimetric accuracy of synthetic CT images generated using a deep learning approach. *Radiotherapy and Oncology*, *136*, 56-63. https://doi.org/10.1016/j.radonc.2019.03.026

Kusztos, R., Dimitri, G. M., & Lió, P. (2020). Neural Models for Brain Networks Connectivity Analysis. In *Computational Intelligence Methods for Bioinformatics and Biostatistics:*

15th International Meeting, CIBB 2018, Caparica, Portugal, September 6–8, 2018, Revised Selected Papers 15 (pp. 212-226). Springer International Publishing. https://doi.org/10.1007/978-3-030-34585-3_19

Langkammer, C., Bredies, K., Poser, B. A., Barth, M., Reishofer, G., Fan, A. P., Bilgic, B., Fazekas, F., Mainero, C., & Ropele, S. (2015). Fast quantitative susceptibility mapping using 3D EPI and total generalized variation. *Neuroimage*, *111*, 622-630. https://doi.org/10.1016/j.neuroimage.2015.02.041

Le Bihan, D., Mangin, J. F., Poupon, C., Clark, C. A., Pappata, S., Molko, N., & Chabriat, H. (2001). Diffusion tensor imaging: concepts and applications. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *13*(4), 534-546. https://doi.org/10.1002/jmri.1076

Levesque, I., Sled, J. G., Narayanan, S., Santos, A. C., Brass, S. D., Francis, S. J., Arnold, D. L., & Pike, G. B. (2005). The role of edema and demyelination in chronic T1 black holes: a quantitative magnetization transfer study. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *21*(2), 103-110. https://doi.org/10.1002/jmri.20231

Li S., Murphy A, Knipe H., & Gaillard, F. (2020, Apr 2). *Energy difference between spin up and spin down states*. Radiopaedia.org. https://doi.org/10.53347/rID-54148

Liang, Z. P., & Lauterbur, P. C. (2000). *Principles of magnetic resonance imaging* (pp. 1-7). Bellingham: SPIE Optical Engineering Press.

Lin, E. C. (2010, December). Radiation risk from medical imaging. *Mayo Clinic Proceedings* (Vol. 85, No. 12, pp. 1142-1146). Elsevier. https://doi.org/10.4065/mcp.2010.0260

Lin, Q., Zhu, F. Y., Shu, Y. Q., Zhu, P. W., Ye, L., Shi, W. Q., Min, Y. L., Li, B., Yuan, Q., & Shao, Y. (2021). Altered brain network centrality in middle-aged patients with retinitis pigmentosa: A resting-state functional magnetic resonance imaging study. *Brain and Behavior*, 11(2), e01983. https://doi.org/10.1002/brb3.1983

Liu, P., De Vis, J., & Lu, H. (2019). Cerebrovascular reactivity (CVR) MRI with CO2 challenge: A technical review. *NeuroImage*, *187*, 104. https://doi.org/10.1016/j.neuroimage.2018.03.047

Liu, Y., Dong, J., Song, Q., Zhang, N., Wang, W., Gao, B., Tian, S., Dong, C., Liang, Z., Xie, L., & Miao, Y. (2021). Correlation between cerebral venous oxygen level and cognitive status in patients with Alzheimer's disease using quantitative susceptibility mapping. *Frontiers in Neuroscience*, *14*, 570848. https://doi.org/10.3389/fnins.2020.570848

Longoni, G., Martinez Chavez, E., Young, K., Brown, R. A., Bells, S., Fetco, D., Kim, L., Grover, S. A., Costello, F., Reginald, A., Bar-Or, A., Marrie, R. A., Arnold, D. L.,

Narayanan, S., Branson, H. M., Banwell, B. L., Sled, J. G., Mabbott, D. J., & Yeh, E. A. (2023). Magnetization transfer saturation reveals subclinical optic nerve injury in pediatric-onset multiple sclerosis. *Multiple Sclerosis Journal*, *29*(2), 212-220. https://doi.org/10.1177/13524585221137500

Magritek. (2016, July 18). *Gradients in NMR spectroscopy – Part 5: the pulsed gradient spin echo (PGSE) experiment.* Retrieved May 20, 2024, from https://magritek.com/2016/07/18/gradients-in-nmr-spectroscopy-part-5-the-pulsed-gradient-spin-echo-pgse-experiment/

Mansbach, R. A., & Ferguson, A. L. (2015). Machine learning of single molecule free energy surfaces and the impact of chemistry and environment upon structure and dynamics. *The Journal of Chemical Physics*, *142*(10). https://doi.org/10.1063/1.4914144

Mercadante, A. A., & Tadi, P. (2023). Neuroanatomy, gray matter. *StatPearls [Internet]*. StatPearls Publishing. Retrieved May 29, 2024, from https://www.ncbi.nlm.nih.gov/books/NBK553239/

Mitchell, T. M., & Mitchell, T. M. (1997). *Machine learning* (Vol. 1, No. 9). New York: McGraw-hill.

Maneewongvatana, S., & Mount, D. M. (1999). Analysis of approximate nearest neighbor searching with clustered point sets. *arXiv preprint cs/9901013*.

Morell, P., & Quarles, R. H. (1999). The myelin sheath. *Basic Neurochemistry: Molecular, Cellular and Medical Aspects*, *6*. Retrieved May 29, 2024, from https://www.ncbi.nlm.nih.gov/books/NBK20402/

Naitzat, G., Zhitnikov, A., & Lim, L. H. (2020). Topology of deep neural networks. *Journal of Machine Learning Research*, 21(184), 1-40. https://doi.org/10.1089/brain.2011.0071

Narasimhulu, Y., Suthar, A., Pasunuri, R., & Vadlamudi, C. V. (2021). CKD-Tree: An Improved KD-Tree Construction Algorithm. *ISIC,* (pp. 211-218).

National Institute of Biomedical Imaging and Bioengineering. (n.d.). *Magnetic resonance imaging (MRI).* Retrieved December 4, 2022, from https://www.nibib.nih.gov/science-education/science-topics/magnetic-resonance-imaging-mri

National Institute of Standards and Technology. (2017). Minkowski distance. Retrieved Jun 19, 2024, from https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/minkdist.htm

Nguyen, K. P., Fatt, C. C., Treacher, A., Mellema, C., Trivedi, M. H., & Montillo, A. (2020, March). Anatomically informed data augmentation for functional MRI with

applications to deep learning. *Medical Imaging 2020: Image Processing* (Vol. 11313, pp. 172-177). SPIE. https://doi.org/10.48550/arXiv.1910.08112

Numpy Developers. (n.d.). Statistics. *NumPy v2.0 Manual.* Retrieved June 21, 2024, from https://numpy.org/doc/stable/reference/routines.statistics.html

Olsson, H., Andersen, M., Wirestam, R., & Helms, G. (2021). Mapping magnetization transfer saturation (MTsat) in human brain at 7T: Protocol optimization under specific absorption rate constraints. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *86*(5), 2562-2576. https://doi.org/10.1002/mrm.28899

Ottet, M. C., Schaer, M., Debbané, M., Cammoun, L., Thiran, J. P., & Eliez, S. (2013). Graph theory reveals dysconnected hubs in 22q11DS and altered nodal efficiency in patients with hallucinations. *Frontiers in Human Neuroscience*, *7*, 402. https://doi.org/10.3389/fnhum.2013.00402

Paterson, R. W., Brown, R. L., Benjamin, L., Nortley, R., Wiethoff, S., Bharucha, T., Jayaseelan, D. L., Kumar, G., Raftopoulos, R. E., Zambreanu, L., Vivekanandam, V., Khoo, A., Geraldes, R., Chinthapalli, K., Boyd, E., Tuzlali, H., Price, G., Christofi, G., Morrow, J., McNamara, P., McLoughlin, B., ... Zandi, M. S. (2020). The emerging spectrum of COVID-19 neurology: clinical, radiological and laboratory findings. *Brain*, *143*(10), 3104-3120. https://doi.org/10.1093/brain/awaa240

Papadakis, N. G., Xing, D., Houston, G. C., Smith, J. M., Smith, M. I., James, M. F., Parsons, A. A., Huang, C. L., Hall, L. D., & Carpenter, T. A. (1999). A study of rotationally invariant and symmetric indices of diffusion anisotropy. *Magnetic Resonance Imaging*, *17*(6), 881-892. https://doi.org/10.1016/S0730-725X(99)00029-6

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32. https://doi.org/10.48550/arXiv.1912.01703

Patenaude, B., Smith, S. M., Kennedy, D. N., & Jenkinson, M. (2011). A Bayesian model of shape and appearance for subcortical brain segmentation. *Neuroimage*, *56*(3), 907-922. https://doi.org/10.1016/j.neuroimage.2011.02.046s

Pinaya, W. H., Tudosiu, P. D., Dafflon, J., Da Costa, P. F., Fernandez, V., Nachev, P., Ourselin, S., & Cardoso, M. J. (2022, September). Brain imaging generation with latent diffusion models. In *MICCAI Workshop on Deep Generative Models* (pp. 117-126). Cham: Springer Nature Switzerland. https://doi.org/10.48550/arXiv.2209.07162

Raj, A., Kuceyeski, A., & Weiner, M. (2012). A network diffusion model of disease progression in dementia. *Neuron*, *73*(6), 1204-1215. https://doi.org/10.1016/j.neuron.2011.12.040

Raveendran, A. V., Jayadevan, R., & Sashidharan, S. (2021). Long COVID: an overview. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *15*(3), 869-875.

Reese, T. G., Heid, O., Weisskoff, R. M., & Wedeen, V. J. (2003). Reduction of eddy-current-induced distortion in diffusion MRI using a twice-refocused spin echo. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *49*(1), 177-182.

Rocks, J. W., Liu, A. J., & Katifori, E. (2020). Revealing structure-function relationships in functional flow networks via persistent homology. *Physical Review Research*, 2(3), 033234. https://doi.org/10.1103/PhysRevResearch.2.033234

Rokem, A. (2015). GitHub - uwescience/shablona: A template for small scientific python projects. Retrieved Jun 29, 2023, from https://github.com/uwescience/shablona

Rosenbloom, M., Sullivan, E. V., & Pfefferbaum, A. (2003). Using magnetic resonance imaging and diffusion tensor imaging to assess brain damage in alcoholics. *Alcohol Research & Health*, *27*(2), 146.

Sanchez-Lengeling, B., Reif, E., Pearce, A., & Wiltschko, A. B. (2021). A gentle introduction to graph neural networks. *Distill*, 6(9), e33. https://doi.org/10.23915/distill.00033

Schenck, J. F. (1996). The role of magnetic susceptibility in magnetic resonance imaging: MRI magnetic compatibility of the first and second kinds. *Medical Physics*, *23*(6), 815-850. https://doi.org/10.1118/1.597854

Schilling, K. G., Chad, J. A., Chamberland, M., Nozais, V., Rheault, F., Archer, D., Li, M., Gao, Y., Cai, L., Del'Acqua, F., Newton, A., Moyer, D., Gore, J., Lebel, C., & Landman, B. A. (2023). White matter tract microstructure, macrostructure, and associated cortical gray matter morphology across the lifespan. *Imaging Neuroscience*, *1*, 1-24. https://doi.org/10.1162/imag_a_00050

Schweser, F., Deistung, A., & Reichenbach, J. R. (2016). Foundations of MRI phase imaging and processing for Quantitative Susceptibility Mapping (QSM). *Zeitschrift für medizinische Physik*, *26*(1), 6-34. https://doi.org/10.1016/j.zemedi.2015.10.002

Scikit-learn developers. (n.d. a). Comparing different clustering algorithms on toy datasets. *scikit-learn 1.5.0 documentation*. Retrieved June 11, 2024, from https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html

Scikit-learn developers. (n.d. b). StandardScaler. *scikit-learn 1.5.0 documentation*. Retrieved June 19, 2024, from https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

SciPy community. (n.d. a). scipy.spatial.cKDTree. *SciPy v1.13.1 Manual*. Retrieved June 19, 2024, from https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.cKDTree.html

SciPy community. (n.d. b). scipy.spatial.cKDTree.query_ball_point. *SciPy v1.13.1 Manual*. Retrieved June 19, 2024, from https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.cKDTree.query_ball_point.html

SciPy community. (n.d. c). scipy.spatial.cKDTree.sparse_distance_matrix. *SciPy v1.13.1 Manual*. Retrieved June 19, 2024, from https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.cKDTree.sparse_distance_matrix.html

SciPy community. (n.d. d). scipy.spatial.distance. *SciPy v1.13.1 Manual*. Retrieved June 19, 2024, from https://docs.scipy.org/doc/scipy/reference/spatial.distance.html

SciPy community. (n.d. e). scipy.sparse.csr_matrix. *SciPy v1.13.1 Manual*. Retrieved June 21, 2024, from https://docs.scipy.org/doc/scipy/reference/generated/scipy.sparse.csr_matrix.html

SciPy community. (n.d. f). scipy.sparse.dok_matrix. *SciPy v1.13.1 Manual*. Retrieved June 21, 2024, from https://docs.scipy.org/doc/scipy/reference/generated/scipy.sparse.dok_matrix.html

SciPy community. (n.d. g). Sparse matrices (scipy.sparse). *SciPy v1.13.1 Manual*. Retrieved June 17, 2024, from https://docs.scipy.org/doc/scipy/reference/sparse.html

SciPy community. (n.d. h). Spatial kdtree class. *SciPy v1.14.0 Manual*. Retrieved June 5, 2024, from https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.KDTree.html

Serai, S. D. (2022). Basics of magnetic resonance imaging and quantitative parameters T1, T2, T2*, T1rho and diffusion-weighted imaging. *Pediatric Radiology*, *52*(2), 217-227.

Sled, J. G. (2018). Modelling and interpretation of magnetization transfer imaging in the brain. *Neuroimage*, *182*, 128-135. https://doi.org/10.1016/j.neuroimage.2017.11.065

Slessarev, M., Han, J., Mardimae, A., Prisman, E., Preiss, D., Volgyesi, G., Ansel, C., Duffin, J. and Fisher, J.A. (2007), Prospective targeting and control of end-tidal CO2 and O2 concentrations. *The Journal of Physiology*, 581: 1207-1219. https://doi.org/10.1113/jphysiol.2007.129395

Smith, R. E., Tournier, J. D., Calamante, F., & Connelly, A. (2012). Anatomically-constrained tractography: improved diffusion MRI streamlines tractography through effective use of anatomical information. *Neuroimage*, *62*(3), 1924-1938. https://doi.org/10.1016/j.neuroimage.2012.06.005

Smith, S. M. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, *17*(3), 143-155. https://doi.org/10.1002/hbm.10062

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M., & Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, *23 Suppl 1*, S208–S219. https://doi.org/10.1016/j.neuroimage.2004.07.051

Sorensen, A. G., Wu, O., Copen, W. A., Davis, T. L., Gonzalez, R. G., Koroshetz, W. J., Reese, T. G., Rosen, B. R., Wedeen, V. J., & Weisskoff, R. M. (1999). Human acute cerebral ischemia: Detection of changes in water diffusion anisotropy by using MR imaging. *Radiology*, *212*(3), 785-792. https://doi.org/10.1148/radiology.212.3.r99se24785

Sporns, O. (2018). Graph theory methods: applications in brain networks. *Dialogues in Clinical Neuroscience*, 20(2), 111-121. https://doi.org/10.31887/DCNS.2018.20.2/osporns

Stejskal, E. O., & Tanner, J. E. (1965). Spin diffusion measurements: spin echoes in the presence of a time-dependent field gradient. *The Journal of Chemical Physics*, *42*(1), 288-292. https://doi.org/10.1063/1.1695690

Suri, R. E. (2003, September). Application of independent component analysis to microarray data. In *IEMC'03 Proceedings. Managing Technologically Driven Organizations: The Human Side of Innovation and Change (IEEE Cat. No. 03CH37502)* (pp. 375-378). IEEE. https://doi.org/10.1109/KIMAS.2003.1245073.

Tharwat, A. (2021). Independent component analysis: An introduction. *Applied Computing and Informatics*, *17*(2), 222-249. https://doi.org/10.1016/j.aci.2018.08.006

Thornhill Medical. (2017). *RespirAct RA-MR Operator's Manual*. Thornhill Research Inc. Retrieved on Jan 29, 2024 from https://thornhillmedical.ca/resources/respiract-ra-mr/manual

Thiessen, J. D., Zhang, Y., Zhang, H., Wang, L., Buist, R., Del Bigio, M. R., Kong, J., Li, X. M., & Martin, M. (2013). Quantitative MRI and ultrastructural examination of the cuprizone mouse model of demyelination. *NMR in Biomedicine*, *26*(11), 1562-1581. https://doi.org/10.1002/nbm.2992

Tromp, D. (2015, October 8). *The diffusion tensor, and its relation to FA, MD, AD and RD*. Diffusion Imaging. https://www.diffusion-imaging.com/2015/10/what-is-diffusion-tensor.html

Tromp, D. (2023). DTI Scalars (FA, MD, AD, RD)-How do they relate to brain structure?. *Authorea Preprints*. https://doi.org/10.15200/winn.146119.94778

Waikhom, L., & Patgiri, R. (2023). A survey of graph neural networks in various learning paradigms: methods, applications, and challenges. *Artificial Intelligence Review*, 56(7), 6295-6364. https://doi.org/10.1007/s10462-022-10321-2

Wang, R., Bashyam, V., Yang, Z., Yu, F., Tassopoulou, V., Chintapalli, S. S., Skampardoni. I., Sreepada, L. P., Sahoo, D., Nikita, K., Abdulkadir, Wen, J., & Davatzikos, C. (2023). Applications of generative adversarial networks in neuroimaging and clinical neuroscience. *Neuroimage*, *269*, 119898. https://doi.org/10.1016/j.neuroimage.2023.119898

Wang, T., Lei, Y., Fu, Y., Wynne, J. F., Curran, W. J., Liu, T., & Yang, X. (2021). A review on medical imaging synthesis using deep learning and its clinical applications. *Journal of Applied Clinical Medical Physics*, *22*(1), 11-36. https://doi.org/10.1002/acm2.13121

Wei, W., Poirion, E., Bodini, B., Tonietto, M., Durrleman, S., Colliot, O., Stankoff, B., & ... Ayache, N. (2020). Predicting PET-derived myelin content from multisequence MRI for individual longitudinal analysis in multiple sclerosis. *NeuroImage*, *223*, 117308. https://doi.org/10.1016/j.neuroimage.2020.117308

Wilson, R.J. (1996). *Introduction to graph theory* (4th ed.). Pearson.

Wu, W. C., Jiang, S. F., Yang, S. C., & Lien, S. H. (2011). Pseudocontinuous arterial spin labeling perfusion magnetic resonance imaging—A normative study of reproducibility in the human brain. *Neuroimage*, *56*(3), 1244-1250. https://doi.org/10.1016/j.neuroimage.2011.02.080

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, *2*, 165-193. https://doi.org/10.1007/s40745-015-0040-1

Xu, L., Li, Q., Myers, M., Chen, Q., & Li, X. (2019). Application of nuclear magnetic resonance technology to carbon capture, utilization and storage: A review. *Journal of Rock Mechanics and Geotechnical Engineering*, *11*(4), 892-908.

Xu, Q. H., Li, Q. Y., Yu, K., Ge, Q. M., Shi, W. Q., Li, B., Liang, R. B., Lin, Q., Zhang, Y. Q., & Shao, Y. (2020). Altered brain network centrality in patients with diabetic optic neuropathy: a resting-state FMRI study. *Endocrine Practice*, *26*(12), 1399-1405. https://doi.org/10.4158/EP-2020-0045

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, *16*(3), 645-678. https://doi.org/10.1109/TNN.2005.845141

Zhang, H., Hubbard, P. L., Parker, G. J., & Alexander, D. C. (2011). Axon diameter mapping in the presence of orientation dispersion with diffusion MRI. *Neuroimage*, *56*(3), 1301-1315. https://doi.org/10.1016/j.neuroimage.2011.01.084

Zhang, H., Schneider, T., Wheeler-Kingshott, C. A., & Alexander, D. C. (2012). NODDI: practical in vivo neurite orientation dispersion and density imaging of the human brain. *Neuroimage*, *61*(4), 1000-1016. https://doi.org/10.1016/j.neuroimage.2012.03.072

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, *20*(1), 45-57. https://doi.org/10.1109/42.906424

Zhao Y, Liang Q, Jiang Z, Mei H, Zeng N, Su S, Wu S, Ge Y, Li P, Lin X, & Yuan K. (2024). Brain abnormalities in survivors of COVID-19 after 2-year recovery: a functional MRI study. *The Lancet Regional Health–Western Pacific*, *47*.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C. & Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI open*, 1, 57-81. https://doi.org/10.1016/j.aiopen.2021.01.001

Zhou, Q., Womer, F. Y., Kong, L., Wu, F., Jiang, X., Zhou, Y., Wang, D., Bai, C., Chang, M., Fan, G., Xu, K., He, Y., Tang, Y., & Wang, F. (2017). Trait-related cortical-subcortical dissociation in bipolar disorder: analysis of network degree centrality. *The Journal of Clinical Psychiatry*, *78*(5), 3831. https://doi.org/10.4088/JCP.15m10091