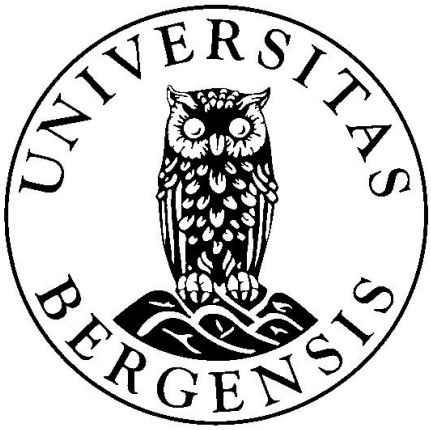
# Exploring the IBS Brain: Resting State Functional Connectivity and Machine Learning




**Peder August Gudmundsen Lillebostad**

Department of Biomedicine
University of Bergen

This thesis is submitted in partial fulfilment of the requirements for the degree of
*Master of Science in Medical Biology*

June 2019

# Acknowledgements

First and foremost I would like to express gratitude to my primary supervisor Professor Arvid Lundervold and cosupervisor Professor Trygve Hausken for providing me with the opportunity to immerse myself into such an enthralling field, at the intersection of neuroscience, computer science and clinical medicine. I am thankful for the social activities I have had the pleasure and privilege to take part in, in particular the trip to Haukelifjell with the Brain-Gut team and a two-day retreat to Solstrand hotel with Mohn Medical Imaging and Visualization centre (MMIV). A special thanks goes to Professor Arvid Lundervold for helpful comments on the thesis, and rich conversations about science in general.

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

# Summary

Germinating from the work of Watts and Strogatz twenty years ago [Watts and Strogatz, 1998], the field of network science has come to blossom [Vespignani, 2018]. Theoretical innovations from mathematicians have over the years led to parallel advancements in applied sciences, proving fruitful in areas as disparate as military intelligence [Krebs, 2002] and the molecular interactions inside cells [Barabási and Oltvai, 2004]. Network science has also been embraced by neuroscientists, exemplified by the introduction of the term 'connectome' - a road map of every neuronal connection in the brain [Sporns et al., 2005]. Systems neuroscientists have acknowledged that the brain is a network and a complete understanding is fundamentally unattainable without studying it explicitly as such Rowe [2010]. Sophisticated medical imaging technologies have enabled the investigation of the human connectome and how it is altered in disorders of the mind, like Alzheimer's disease, schizophrenia, autism and more [Heuvel and Sporns, 2019; van den Heuvel and Sporns, 2013]. The combined effort of computer- and neuroscientists have led to the application of graph theoretical metrics as network-based biomarkers of neurological conditions, often in conjunction with machine learning algorithms to achieve impressive discrimination between clinical groups [Bachmann et al., 2018; Brown and Hamarneh, 2016; Fornito et al., 2015, 2012; Hosseini-Asl et al., 2016]. In fact, the network approach to disease in general is emerging [Barabási et al., 2011; Mayer et al., 2015].

The network approach has also taken foothold in the research on irritable bowel syndrome (IBS), with particular appreciation of the brain-gut axis [Mayer, 2018; Mayer et al., 2015]. Not until recently have efforts started to integrate the different observations of diet, the gut micribiome, immune system and nervous system. Much work has been done to establish the brain involvement in IBS pathology through structural and functional neuroimaging [Bhatt et al., 2019; Bonaz, 2002; Gupta et al., 2015; Icenhour et al., 2017; Labus et al., 2015; Seminowicz et al., 2010; Wang et al., 2017] (see [Tillisch

et al., 2011] and [Mayer et al., 2015] for an overview). Although some work has been done in terms of network modelling of structural/functional brain connectivity in IBS [Labus et al., 2008, 2014, 2009, 2019], no study to date has investigated the application of machine learning classification on graph metrics derived from functional connectivity in IBS patients.

The contribution of this thesis is twofold. The first part is a methodological investigation of robustness of graph metrics through a test-retest reliability study. Developing reproducible biomarkers is of utmost importance in the imaging of neurological and psychiatric disorders with the potential to predict vulnerable individuals, prognosis and treatment response [Vieira et al., 2017; Waller et al., 2017]. By test-retest simulation on human connectome project data, evidence is presented for the relative reliability of twelve commonly used graph metrics, and how they are affected by thresholding (a common processing step). Expanding windows analysis is used to explore the effect of scan duration on graph metric estimates, demonstrating that some are highly sensitive to change even at 8 minutes. The second part is an exploration of resting state functional connectivity of IBS patients as assessed through functional magnetic resonance imaging. Anatomical segmentation in FreeSurfer is used to define network nodes to extract time courses from functional data preprocessed in AFNI. Different approaches to network modelling (Pearson correlation, partial correlation and sparse inverse covariance) are attempted from which graph metrics are estimated. These network-based biomarkers are used to train machine learning classifiers in order to discriminate IBS patients from healthy controls, with a beyond chance classification accuracy. Self-reported nausea is more readily detected than the actual diagnostic status, although these groups were not controlled for motion differences. High dimensionality data and few samples ($p \gg N$) are addressed using recursive feature elimination as a feature selection/dimensionality reduction. A network difference between the IBS and HC groups is further established with high confidence using permutation testing (network based statistic). Most of the work was conducted in Python through Jupyter Notebook, an interactive environment for the development and testing of scripts.

The introduction aims to give a broad interdisciplinary overview of IBS and magnetic resonance imaging and brain connectivity, and is followed by a theoretical section delving into graph theory and statistical concepts. An experimental section describes what was done, before presenting the results, and finally a discussion of the nuisances of the study, interpretation and clinical implications.

# Table of contents

# List of figures

# Selected abbreviations

**BGA** Brain-Gut Axis. 7

**BOLD** Blood Oxygen Level Dependent. 10

**CBF** Cerebral Blood Flow. 10

**CV** Cross Validation. 45

**FDR** False Discovery Rate. 26

**FGID** Functional Gastrointestinal Disorder. 3

**fMRI** Functional Magnetic Resonance Imaging. 10

**FWER** Family-Wise Error Rate. 26

**HC** Healthy Control. 37

**IBS** Irritable Bowel Syndrome. 3

**ICA** Independent Component Analysis. 30

**LCC** Largest Connected Component. 49

**LFP** Local Field Potential. 12

**LOGOCV** Leave-One-Group-Out Cross Validation. 46

**LOOCV** Leave-One-Out Cross Validation. 45

**MRI** Magnetic Resonance Imaging. 8

**PET** Positron Emission Tomography. 10

**QOL** Quality of Life. 3

**RBF** Radial Basis Function. 46

**RFE** Recursive Feature Elimination. 47

**rs-fMRI** Resting State fMRI. 16

**RSN** Resting State Network. 16

**SVM** Support Vector Machine. 46

# Part I

# Theoretical Background

# Chapter 1

# Introduction

## 1.1 Irritable Bowel Syndrome - a multifaceted disorder

Irritable bowel syndrome (IBS) is a heterogenous chronic gastrointestinal disorder affecting between 8% and 23% of the global population [Elsenbruch, 2011]. The condition is categorized as a functional gastrointestinal disorder (FGID), implying the absence of a well-established defining cause of biochemical, cellular, infectious or otherwise structural origin. The symptoms of IBS are manifested as recurring pain/discomfort in the gut, visceral hyperalgesia, dysfunctional GI motility, and often altered bowel habits. Patients of IBS are commonly subdivided into groups based on their predominant excretion symptom: constipation (IBS-C), diarrhea (IBS-D), a combination of the two (IBS-M), or none (IBS-U). Those who suffer from IBS also have an overall higher rate of depression, scoring significantly lower on quality of life QOL measures than controls [Gralnek et al., 2000].

Due to the ill-defined pathophysiology of IBS, it has long been a subject of controversy among medical professionals, some even dismissing it as a 'real' disorder altogether [Drossman, 2006]. Moreover, IBS has historically been a 'last resort' diagnosis, only considered after excluding other alternatives with a straight forward cause, like colonic cancer or gastroenteritis [Canavan et al., 2014]. The desire of medical doctors to identify a 'structural' cause of IBS has historically led to numerous instances of clinical mishandling, like superfluous surgery [Canavan et al., 2014]. Today, diagnostic status

is assessed independently with respect to the outcome of a standardized questionnaire based on GI symptoms (ROME criteria, for the most recent version (IV) see, Palsson et al. [2016]). The acknowledgement of psychosocial risk factors has been crucial in the modern understanding of IBS.

## 1.2 Epidemiology and the societal impact of IBS

The global incidence of IBS is alarmingly high. Reports vary, and some countries have unreliable estimates, but the number reported rarely subceeds 10% [Canavan et al., 2014]. All demographic groups are affected by IBS, but females are substantially over represented in the statistics [Elsenbruch, 2011]. The incidence of IBS also drops in those above 50 years of age. IBS patients frequently report fatigue [Gralnek et al., 2000], which may lead them to miss working hours. Together with often futile doctor visits, IBS contributes to a non-trivial loss to the global work force. So despite comparatively mild symptoms, the shear prevalence of IBS in itself poses a considerable economic burden on society. Thus, in addition to the primary detriment of reducing the QOL of patients, there is an economic incentive to increase understanding of IBS, ultimately to guide the development of effective therapies.

## 1.3 Factors involved in IBS pathophysiology

A wide array of causes are believed to contribute to IBS symptoms [Mayer et al., 2015]. A great deal of work has been done investigating the role of diet, infectious agents, the gut microbiome, immune system and enteric nervous system (ENS), epithelial permeability in the gut, genetics, sex, psychological trauma and stress, central processing and more in IBS etiology [Mayer et al., 2015]. This work has elucidated groups at elevated risk of developing IBS (e.g. people with infections of the GI tract often develop IBS later in life) and generated a wealth of significant, but relatively weak biomarkers for IBS.

### 1.3.1 Immune mechanisms in IBS

Evidence points to an involvement of the immune system in facilitating the symptoms experienced by IBS patients. The primary reason to involve the immune system is

that individuals who have suffered intestinal infection (gastroenteritis) at some point frequently proceed to develop IBS later in life (post-infectious IBS, [Barbara et al., 2011]). Secondly, mast cells (a type of white blood cell) have been observed in elevated numbers in the intestinal mucosa of a subset of IBS patients [Cremon et al., 2009]. Concurrently, mucosal biopsies reveal elevated histamine levels in patients than in controls. Being a mast cell mediator, histamine triggers activation of the immune response [Enck et al., 2016].

### 1.3.2 The genetics of IBS

With regards to genetic explanations, hereditary studies can be helpful, not to identify which genes specifically are involved, but rather to what extent genetics are involved at all. The incidence of IBS among parents and their offspring do indicate a certain degree of heredity: parents with IBS are at heightened risk of developing it themselves [Canavan et al., 2014]. Interestingly though, having a homozygous twin with IBS is actually less predictive of an individual's disease status than having parents with IBS [Canavan et al., 2014]. This supports a slightly stronger effect of environmental factors than for genetic ones [Levy et al., 2001].

One highly cited paper [Saito et al., 2009] reports the discovery of a mutation in a sodium channel (SCN5A) expressed in the interstitial cells of Cajal, cells lining the gut involved in coordinating muscle contraction. Genetic sequencing of 49 subjects diagnosed with IBS revealed a single individual with a nonsense (early terminating) mutation in SCN5A. The mutation leads to a premature stop-codon, leading to a dysfunctional sodium channel in, thereby serving as a highly plausible mechanistic explanation of abnormal bowel habits. An analysis of 1500 healthy subjects did not show a single occurrence of the same or equivalent mutation, strongly supporting its role in IBS. Despite stirring great interest among researchers, it doesn not account for more than an estimated 2% (1/49) of IBS cases [Saito et al., 2009], leaving the remaining 98% elusive.

### 1.3.3 The microbiome of IBS

With a cell count competing with our own, (about 1:1 according to a modern estimate, [Sender et al., 2016]), gut bacteria have received well-deserved attention the last years

regarding their role in health and disease, particularly regarding GI health [Mayer et al., 2014]. A study from 2016 found an altered microbiota based on molecular analysis of fecal samples from 62 IBS patients and 46 controls. The severity of symptoms have been shown to correlate positively with the lack of biodiversity [Tap et al., 2017]. The complexity of the microbiome makes it challenging to characterize the mechanisms through which they affect IBS symptoms, and most studies merely measure the composition of bacterial species without regarding things like genetics. However, indirect evidence of microbiota involvement is found in studies of how antibiotics and probiotics affect IBS symptoms; generally probiotics have a positive effect and antibiotics a negative, in accordance with observations of germ-free mice [Mohajeri et al., 2018]. While studies have reported divergent results on the effect of microbiota, the most likely mechanism is through their metabolic products (e.g. short chain fatty acids, serotonin) [Mayer et al., 2014]. While correlational links have been established, the causation is not equally clear [Martin et al., 2018]. A particularly promising gateway is the synthesis of the neurotransmitter serotonin by certain bacteria [Labus et al., 2019].

### 1.3.4 Psychosocial interactions and central processing

Observations that IBS strongly overlaps with anxiety disorders and early adverse life events [Park et al., 2016; Przekop et al., 2012] support a role of psychological mechanisms in facilitating the disorder. Inclusion of psychological and behavioural factors into the disease model of IBS has old roots in observations of the influence of mood on gastrointestinal symptoms [Drossman, 2016]: one anecdote describes how medical students who were healthy, but were told they had cancer, experienced an increase in rectal contractions [Drossman, 2016]. The concurrence of IBS with other functional disorders (Migraine, [Georgescu et al., 2017]; functional dyspepsia, [Hillilä et al., 2007]; fibromyalgia, [Przekop et al., 2012]) further underpins the notion that functional pain disorders are not best explained by a single misfolded protein or other structural aberration, but rather a dysfunction in the organ that both perceives pain and regulates bodily function - the brain. Descending pain modulation is an attractive target for a role in functional pain disorders [Wilder-Smith, 2011]. For instance the periaqueductal gray - a small nucleus situated in the brain stem - produces pain-relieving neuropeptide enkephalins [Vanegas and Schaible, 2004]. It should be noted

that it is hard to separate the psychological symptoms that result from depression from those that are from IBS *per se*.

Functional brain imaging studies have shone some light on brain alterations in patients with IBS [Holtmann et al., 2016]. A meta analysis on distension studies in IBS patients show a clear pattern of significant activation of the insular and cingular cortices compared to controls [Tillisch et al., 2011]. Furthermore, distributed structural alterations in grey matter thickness and white matter integrity, albeit subtle, have been identified in IBS patients [Holtmann et al., 2016; Seminowicz et al., 2010]. The involvement of the brain however is a chicken-and-egg type problem - alterations in the brain and ultimately in mood and psychological states may just as well be a result of peripheral processes than the other way around [Holtmann et al., 2016].

## 1.4 IBS: a brain-gut disorder

After decades of research on these apparently separate or unrelated mechanisms, they have yet to be integrated in a unified framework which considers how they interact [Mayer et al., 2015]. One theory which has taken hold among many researchers is the brain-gut theory of IBS [Mayer et al., 2015]. At the core is the brain-gut axis (BGA) - an all encompassing term for communication pathways that link central nervous function with enteric function [jones et al., 2006]. Bidirectional communication occurs through endocrine, immune and neuronal pathways [jones et al., 2006]. The hypothalamic-pituitary-adrenenal (HPA) axis is a major endocrine pathway, important in stress regulation [Liang et al., 2018]. Beginning with the secretion of corticotropin-releasing hormone (CRH) from the hypothalamus, stimulating the pituitary to release adrenocorticotropic hormone ACTH, the HPA axis terminates in the andrenal gland, where ACTH triggers cortisol secretion (cortisol being an effector of multiple stress-related functions, like increased heart rate and alertness). Upregulation of the HPA-axis has been reported in IBS patients compared to healthy controls [Chang and Glover, 2009; jones et al., 2006]. Furthermore, animal studies have shown that germ-free mice develop an dysfunctional HPA axis, effectively elevating the stress response. This effect is partly reversible by the introduction of *Bifidiobacterium infantis* to the gut [Liang et al., 2018], demonstrating the gut's influence over the mind.

The reverse route of communication has also been known for some time [Tannock and Savage, 1974]. For example, HPA activity has been shown to alter intestinal

motility and permeability [Mayer et al., 2015]. Even more, stressful events have the ability to alter the gut microbiota composition, as demonstrated in rats [O'Mahony et al., 2009]. Another intriguing phenomenon is that of microbial metabolism. The numerous bacterial species inhabiting the human intestines are capable of synthesizing a rich soup of neurotransmitters - such as gamma-aminobutyric acid (GABA), serotonin, dopamine and acetylcholine, serving as a plausible mechanism of action by direct effect on nerve cells [Cryan and Dinan, 2012].

A number of functional imaging studies have shed much light on brain networks which appear dysregulated in IBS [Mayer et al., 2015; Tillisch et al., 2011]. The emotional arousal network comprises a set of frontal regions of cortex, in addition to the amygdala and locus corelolus (noradrenalin producing center of the brain, located in the pons). Upregulation of this network has been demonstrated in people upon receiving painful stimuli. Central autonomic network is important in IBS, in part because of stress-responsiveness. The sensorimotor network role is implied from studies of both structural and functional magnetic resonance imaging (MRI), implicating structural changes in these regions in people with IBS. This is attractive because these regions (S1, S2 are early relay stations of visceral sensory input). The salience network (insula, dACC) is especially relevant. The insular cortex is primarily associated with the detection of and direction of attention towards stimuli [Uddin, 2015]. Particular attention has been directed towards the somatosensory network, salience network and default mode network. The salience network being in [Gupta et al., 2015; Icenhour et al., 2017]. IBS is associated with increased attention (salience) of visceral stimulus, thereby making the salience network of particular interest [Seeley et al., 2007], for the sake of its role in awareness on a particular stimulus. Interestingly, it has been shown that just the anticipation of pain is sufficient to upregulate this system in IBS, even in the absence of actual stimuli [Mayer et al., 2015]. The central executive network is highly coactivated with the salience network. Through the effects exerted by various neurotransmitters in these systems, gastrointestinal function is altered (e.g. noradrenalin acts through the sympathetic nervous system and inhibits motility and is associated with arousal and intense fear when released in large amounts in the CNS [Sara and Bouret, 2012]; seretonin also produced in basal brain areas (raphe nuclei) is largely implicated in mood and depression. Serotonin is also altered in the enteric nervous system in IBS, but that is a seperate matter [Beattie and Smith, 2008].

In summary, the brain holds a pivotal role in the regulation of intestinal function that is relevant for symptom generation in IBS. However, the influence is not unidirectional, but is best understood as a complex interplay between the brain and the gut, mediated by hormones, direct neural innervation, immune mechanisms and microbiome metabolism [Mayer et al., 2015]. The next section dives into how brain function can be studied using magnetic resonance imaging (MRI).

## 1.5 Magnetic resonance imaging

Owing to its ability to generate high resolution anatomical images essentially non-invasely, nuclear magnetic resonance imaging (MRI) stands out among medical imaging modalities. The technology exploits the nuclear spin properties of hydrogen atoms, which are ubiquitous in all biological tissues (primarily due to water and lipids). Application of a strong magnetic field $B_0$ (usually between one and seven Tesla (T)) forces the nuclear spins to align with $B_0$. A radio frequency (RF) pulse is emitted to excite the hydrogen nuclei, temporarily offsetting the magnetization vector's longitudinal component $M_z$ from $B_0$. Following an exponential restoration equation they regain equilibrium with $B_0$, the kinetics of which are characterized by the time constant T1. Alongside the relaxation, energy is reemitted and picked up by the scanner's receiver coil.[1] The intensity of the received signal is determined by the progression of the relaxation process; increasing as nuclear spins return to steady-state. Simultaneously, the transverse magnetization component $M_{xy}$ decays along the plane orthogonal to $B_0$, described by a similar kinetics and a relaxation constant, T2. The key to MR images is the differential tissue-dependent T1 and T2 relaxation of protons, determined specifically by their chemical environment bond. This generates a contrast in signal intensity between tissue types, and thus an image can be produced. An array of MRI modalities have been developed from these basic principles by varying scanning parameters and the pulse sequence (ordered sequence of RF emission and reception), providing a means to study the anatomical and functional properties of the brain and other organs.[2]

[1] This description is a massive simplification; for an in-depth discussion on the matter, see [Hoult, 2009, 1989]).

[2] The detailed physics underlying the nuclear magnetic resonance phenomenon and the technological advancements behind its exploitation in medical imaging is far beyond the scope of this work. Many good sources on the matter are available for the interested reader: [Ai et al., 2012; Kubo and Tomita, 1954; Lauterbur, 1973; Rabi et al., 1938; Ramsey, 1999; Rigden, 1986; Rinck et al., 2018].

## 1.6 BOLD fMRI and resting state connectivity

### 1.6.1 Discovery

Although the brain makes up only 2% of the body's mass, it consumes a staggering one fifth of its energy Clarke and Sokoloff [1999], primarily due to the activity of ion transporters perpetually maintaining a voltage across the plasma membrane. It was postulated more than a century ago that the brain coordinates oxygen delivery to regions based on local oxygen consumption and immediate needs [Roy and Sherrington, 1890]. In other words, the brain actively takes part in delegating oxygenated blood to where it is needed based on metabolic energy requirements: simple supply and demand. This knowledge was exploited by neuroscientists in the late 20th century who used positron emission tomography PET to measure cerebral blood flow (CBF) in response to neural stimulation.

This phenomenon is known as neurovascular coupling - neural stimulation elicits a response of increased cerebral blood flow CBF. While the essence of Roy and Sherrington's postulate [Roy and Sherrington, 1890] is true (i.e. neurovascular coupling is a real phenomenon, see Fig. 1.1), one of their core assumptions were challenged in the late 1900's, namely the assumed one-to-one relationship of oxygen consumption and regional CBF [Fox and Raichle, 1986; Raichle, 1998]. The work of Fox and Raichle [Fox and Raichle, 1986] demonstrated an an unanticipated phenomenon: uncoupling between $O_2$ consumption and CBF. Although at resting conditions, a near perfect correlation could be reported between metabolic rate of oxygen consumption (CMRO2) and CBF, they consistently observed a mismatch during acute neural activation - in fact the CBF increased disproportionately compared to the CMRO2, which only rose slightly. These findings were pivotal to the understanding of brain metabolism, laid the foundation of blood oxygen level dependent (BOLD) functional MRI (fMRI) and led a forefront of research in coming decades of brain imaging [Raichle and Mintun, 2006].

To the delight of many researchers, a completely non-invasive technique was developed some years later by [Ogawa et al., 1990]. Ogawa's work was based on the findings of Fox and Raichle, and exploited a phenomenon known as T2* ("T2 star") relaxation [Kwong et al., 1992]. This method, BOLD fMRI, is directly sensitive to local concentrations of deoxygenated hemoglobin, and could thereby bypass the need for any exogenous contrast agents, tracers or otherwise invasive technology. BOLD

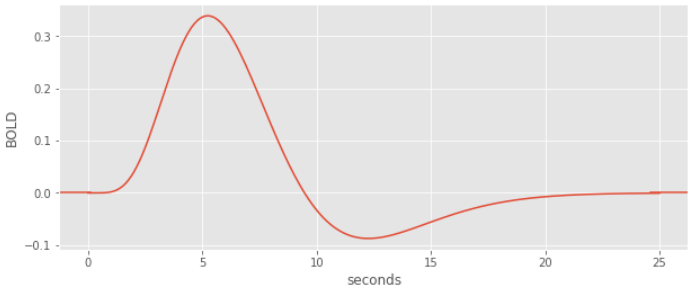



**Fig. 1.1:** Hemodynamic response function. A stimulus at time=0 elicits a response in the local cerebral blood flow, which can be observed in the BOLD signal. The onset is a few seconds delayed and peaks after about five seconds.

fMRI remains the predominant method in functional brain imaging to this day. The sensisitivity is due to the presence of four iron atoms in every molecule of hemoglobin. Being paramagnetic, iron interacts with and distorts the magnetic field generated by the scanner. This reduces the T2 and T2* constants (i.e. slows down the relaxation), resulting in a weaker signal in voxels in the vessel's proximity. Molecular oxygen ($O_2$) is diamagnetic, effectively cancelling the effect of iron. Thus, oxygenated and deoxygenated hemoglobin have differential magnetic properties, and this effect is measurable. Accordingly, when a brain region receives more oxygen due to vasodilation, the relative oxyhemoglobin/deoxyhemoglobin concentration rises, increasing the brightness of the voxel [Hillman, 2014]. This process is facilitated by an increase in rCBF to the neuronal populations as they perform work and consume energy.

The detailed mechanism of the generation of the BOLD response is quite intricate, and not fully understood [Hillman, 2014]. Activation in a region ensues with an increased consumption of oxygen, stripping oxyhemoglobin of their $O_2$. This effect initially reduces the concentration of oxyhemoglobin in favor of deoxyhemoglobin. According to theory, a rise in deoxyhemoglobin will weaken the fMRI signal, but paradoxically, the opposite effect is observed: stimulation of a region is followed by an increase of the signal. In other words, CBF does compensate for oxygen consumption, but they don't cancel out to zero. The reason is that the initial deficiency of $O_2$ triggers incoming arteries to expand (vasodilation), increasing the oxyhemoglobin/deoxyhemoglobin ratio. In fact, there is an overcompensation for the loss of $O_2$ which allows the process to be measures [Raichle, 1998]. The whole process from physiology to be outlined in two steps (see Fig. 1.2).

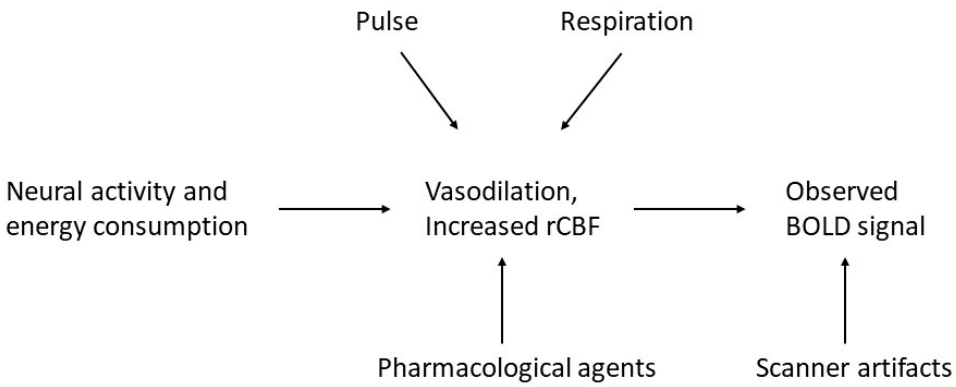



**Fig. 1.2:** Neurovascular coupling. Two main steps outline the path from neural activation to fMRI signal: 1) energy and oxygen consumption by neurons signal neighboring glial cells to facilitate vasodilation, increasing the rCBF and oxyhemoglobin concentration, and 2) the different magnetic properties of oxyhemoglobin and deoxyhemoglobin impacts the T2 and T2* relaxation constants, affecting the fMRI signal. This BOLD response occurs only a few seconds after the neuronal activation (Hillman, 2014). Researchers are ultimately only interested in the neurophysiology, but only have access to the BOLD signal. To complicate the picture even more, other nuisance variables are present and affect the signal in multiple ways (only a few are shown).

### 1.6.2 Biological interpretation of the BOLD signal

Early efforts to infer brain function from the BOLD signal were met with great skepticism in regards to whether the observed signal actually reflected neuronal activity at all, or if it was merely indicative of other physiological processes. Today, conclusive evidence has put this question to rest, through the simultaneous recording of electrophysiology and BOLD fMRI in monkeys [Logothetis et al., 2001; Shmuel and Leopold, 2008]. Similar findings with electroencephalography (EEG) have since been confirmed in humans [Laufs et al., 2003]. Exactly *how* the coupling between the BOLD signal and neural activity happens is a different matter, and much more challenging to answer. There is nonetheless no controversy today as to whether BOLD fMRI is reflective of neural activation [Fox and Raichle, 2007]. As a matter of fact, extensive investigation has revealed some details of the relationship between the the BOLD signal and neural activity demonstrating a strong concordance between local field potentials (LFPs) and low frequency oscillations observed in the BOLD response [Logothetis et al., 2001]. There are, however, an array of nuances that may affect this coupling, and the precise mechanisms are generally considered poorly characterized [Hillman, 2014]. With regards to interpreting the BOLD signal beyond "a region of activation",

another question must be answered: does the BOLD signal reflect a region's synaptic output, or does it reflect the integration of dendritic input? [Raichle and Mintun, 2006] argue for the latter, due to the correlation of BOLD signal to LFP.

### 1.6.3 Summary

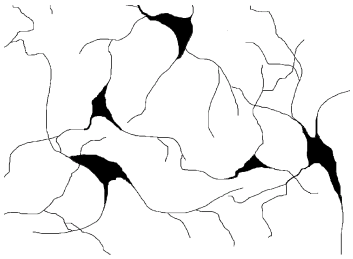
It is important to emphasize the indirect nature of the BOLD signal. The observed signal can only serve as a proxy for neural spiking, and we should be careful to automatically attribute it as otherwise. BOLD fMRI is a powerful tool to investigate the living brain of humans and other animals non-invasively. The total signal however is extremely noisy, and the BOLD response is only a minor contributor - on report stating as little as 2.9% of the total signal [Bianciardi et al., 2009]. Other nuisance variables include motion, physiological noise, and scanner instabilities. The fMRI signal is thus a complex mixture of various sources of noise and neural activity. The level of noise poses a major challenge on fMRI data analysis, and requires a multitude of preprocessing to remove or correct for variables that are not of interest.

## 1.7 Brain Connectivity

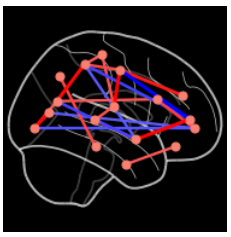
The human brain is a dauntingly complex network. At its finest resolution, it comprises around 80 billion neurons, interconnected through 100 trillion synapses. This view has been apparent since the iconic drawings by Santiago Ramon y Cajal of cells stained with silver nitrate. At a larger spatial scale, the brain is organized into functional communities/neuronal populations, which are integrated by long white matter connections [Betzel and Bassett, 2017], facilitating the flow of information between distant cortical regions as well as subcortical nuclei. Irrespective of spatial scale, this description of the brain is referred to as *brain connectivity* (Fig. 1.3). The terms *node* and *link* will be used to denote the individual units and their connections (a more formal introduction on the matter is provided in chapter 2. The term connectome was added to the heap of *-omics* by [Sporns et al., 2005], marking a new era of neuroscience [3]

Network structure and topology are crucial to study because they are determinants of network function [Boccaletti et al., 2006]. The spread of misinformation on social

[3]The term "connectivity" was in use long before the term "connectome" [Friston et al., 1993].

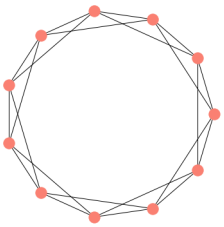


**(a)** Microscale connectivity

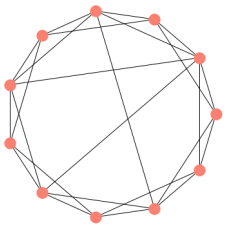


**(b)** Macroscale connectivity

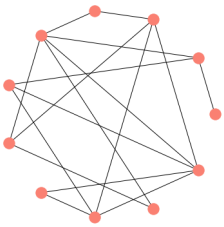
**Fig. 1.3:** Two scales of brain connectivity



**(a)** Regular grid-like network



**(b)** Small-world network



**(c)** Erdős-Rényi random network

**Fig. 1.4:** Different network topologies ranging from completely regular to fully random

media is determined the links between online profiles and pages, the ease of air travel is restricted by the structure international airport networks, and viruses spread through the physical connections between their hosts. This principle also holds true for the brain: it is not the individual neurons, but rather the connections between neurons are what facilitate learning, memory, and ultimately thoughts and behaviour.

### 1.7.1 Structural connectivity

Many have investigated the topologies of brain networks in model organisms, demonstrating far from random topologies [Bassett and Bullmore, 2006; Watts and Strogatz, 1998]. For example, a neuron-complete map of the nervous system of C. *elegans* [Watts and Strogatz, 1998], displays a property known as 'small-worldness', with similar findings for more complex animals like the macaque. Small-world networks [Milgram, 1967] have a remarkable global connectedness - like the famous "six degrees of separation" of social networks. This topology is characterized by a large number of weakly connected clustered nodes, and the presence of a few highly connected nodes (Fig. 1.4b), referred to as hubs. Structural connectivity has classically been inferred by the injection of fluorescent dyes in dissected brains, which can be traced along white matter tracts to

build a "wiring diagram" of the brain. Modern methods also allow for non-invasive imaging to build models of anatomical connectivity [Bihan et al., 2001].

Energy expenditure is an important organizing principle of brain networks [Bullmore and Sporns, 2012]. The 'wiring cost' of a connectome is defined in terms of physical space (all the connections are confined to a room just around 1.5 liters), signal conduction velocity and material (the cytoplasmic contents of a brain cell). The impact of economical efficiency is most apparent in the distance between connected neurons: neurons are primarily connected to spatially proximate neighbors. A similar pattern is observed at higher spatial scales - the probability of connection between two regions is inversely proportional to their spatial distance [Bullmore and Sporns, 2012]. Another property of structural brain networks is related to modularity - the separation of the brain into functionally coherent divisions [Bullmore and Sporns, 2012]. A module, or community can be loosely defined as a highly interconnected subset of the network, with fewer connections to other, more distant modules. Different modules are functionally specialized. Centrally positioned nodes serve as integration centers, facilitating the communication *between* modules. Such an organization has been observed in macaques, cats and humans [Bullmore and Sporns, 2012]. This supports the idea that brain topology is, at least in part, optimized in terms of economy.

### 1.7.2 Functional connectivity and resting state networks

Functional connectivity (FC) is a more abstract concept than the structural description of the brain. Bound by the restrictions set by the brain's anatomical connections, functional connectivity complements its static nature by assessing the correspondence of *activity* between different loci [Park and Friston, 2013]. This is most usually calculated through simple Pearson correlation between node time series, but many alternatives exist (e.g. partial correlation, coherence, mutual information). This allows for a continuous representation of links, based on the strength of correlation. Functional connectivity is purely descriptive - it is not concerned with cause and effect (i.e. effective connectivity), but simply statistical dependence. It is all based on the assumption that regions that coactivate also interact. Functional connectivity is not as easily interpreted as its anatomical counterpart, and different methods of estimation will produce different results. For example, the presence of a functional connection does in no means imply a structural connection, much less the causal interaction between regions. In fact, correlation-based overestimates interactions because of the transitivity

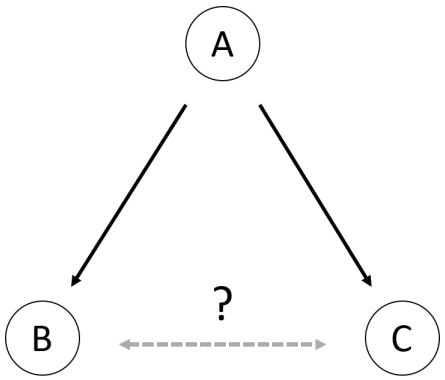



**Fig. 1.5:** Shared input connectivity. Innervation from region A to B and C will produce a correlation in the BOLD signal between A and B and between A and C. Despite the absence of an anatomical connection between B and C, they will appear correlated. Models of effective connectivity (and to a lesser extent partial correlation) are aimed at resolving this issue of Pearson correlation, so that only direct connections are identified.

of correlation: if A is correlated with B, and B is correlated with C, then A and C are also correlated [Zalesky et al., 2012], for example a shared input from a common region [Friston, 1994] (see Fig. 1.5).

Up until a quarter century ago, all fMRI studies subscribed to a task-paradigm. These traditional task-fMRI studies worked by subjecting the examinee to two or more different conditions or "tasks" (such as looking at faces versus looking at houses; [Kanwisher et al., 1997]), in an event-related or block design. Inference was made from subtracting the regional activity between the two conditions with the purpose of "isolating" the neural correspondence of a behaviour of interest [Fox and Raichle, 2007]. Yet another milestone in the field of fMRI was reached by Bharat Biswal when he demonstrated that the brain exhibited interesting patterns of activity even in the absence of active task engagement [Biswal et al., 1995], a paradigm that came to be known as resting-state fMRI (rs-fMRI). By selecting a small patch of voxels located in the left hemisphere motor cortex, Biswal correlated its temporal activity (BOLD time series) with every other voxel of the brain. From this set of correlations he was able to synthesize a correlation map detailing the topography of the somatosensory system. Known as seed-based connectivity, this approach has become the primary method to study so called resting state networks (RSN) (Fig. 1.6). Biswal's experiment initially received backlash [Biswal, 2012], and many claimed the correlation to be purely artifactual. However, deeper research into the matter showed that at least part of the signal corresponded to physiologically meaningful events as opposed to non-neuronal noise [Biswal, 2012]. These signals correspond to slowly fluctuating (0.01-0.1 Hz) changes in the BOLD signal.

Many other resting state networks have been identified all across the brain since Biswal's seminal work, relating to language, visual, auditory and other modalities

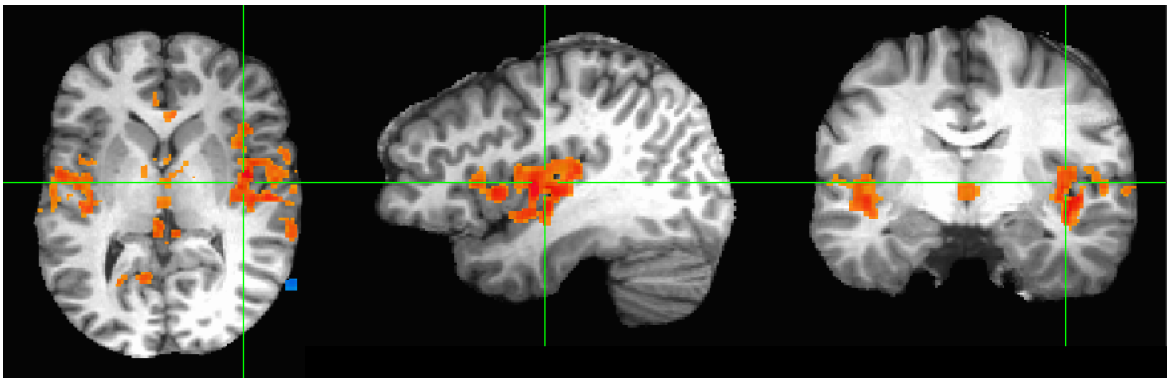


**Fig. 1.6:** Seed-based connectivity. A correlation map was computed from a seed in the insular region (green crosshairs). Stronger correlations are in red and weaker correlations in orange. Obtained from resting state data from a healthy subject. Preprocessed, analysed and visualized using AFNI.

of brain function [Fox and Raichle, 2007]. The resting state networks are highly reproducible across individuals [Biswal et al., 2010]. Most notably is the default mode network (DMN), a network which are consistently more active *during* resting conditions than task [Barkhof et al., 2014]. The approach of studying these networks under rest is supported by a generally strong correspondence with networks during task activation [Smith et al., 2009]. Moreover, the energy consumption of an "active" brain is only marginally higher (roughly 5%) than the resting state brain [Fox and Raichle, 2007].

Functional connectivity has been successfully applied in combination with graph theory and machine learning classification algorithms to predict the diagnostic status of patients of psychiatry, as well as other other clinical, cognitive or behavioural measures (Alzheimer's disease, [Bachmann et al., 2018]; epilepsy, [Zhang et al., 2012] mental maturity, [Dosenbach et al., 2010]; smoking status, [Pariyadath et al., 2014]; task preparation, Ekman et al. [2012]; major depressive disorder, [Craddock et al., 2009; Zhong et al., 2017]; schizophrenia, [Liu et al., 2008; Shen et al., 2010]). The next chapter aims to give an overview of the theoretical concepts behind the study of brain connectivity.

# Chapter 2

# Theory

## 2.1 Graph Theory

The first historical account of graph theory can be dated back to the Swiss mathematician Leonhard Euler in 1736, when he was able to show that it was impossible to cross all seven bridges of Königsberg without crossing any one bridge twice or more (see 2.1) [Euler, 1741]. The two fundamental building blocks of a network are its *vertices/nodes* and *edges/links*. Thus we may regard the landmasses as nodes, and the bridges as edges connecting the nodes. Early graph theory was focused on characterizing small graphs with precisely defined topologies, akin to the toy problem of Königsberg. These graphs are simple and can be understood visually by tracing every link connecting the nodes. The same luxury cannot be enjoyed for larger and more complex real life networks, where network edges often are continuous, vary with time or might even be uncertain. These networks challenged the field to develop novel concepts.

Graph theory had a renaissance in the 1990's [Barabási and Albert, 1999; Vespignani, 2018; Watts and Strogatz, 1998] which led to the development of many interesting ideas about networks. Of particular importance was the small-world networks, described by [Watts and Strogatz, 1998]. A small-world topology is of interest because of its trade-off between low economic cost and high efficiency at integrating information. The architecture of small-world networks is characterized by a high number of clustered nodes with a majority of short-distance connections. However, a handful of highly connected hub nodes bridge distant regions of the network, dramatically reducing the distance between any two randomly chosen nodes. Extensive work on small-

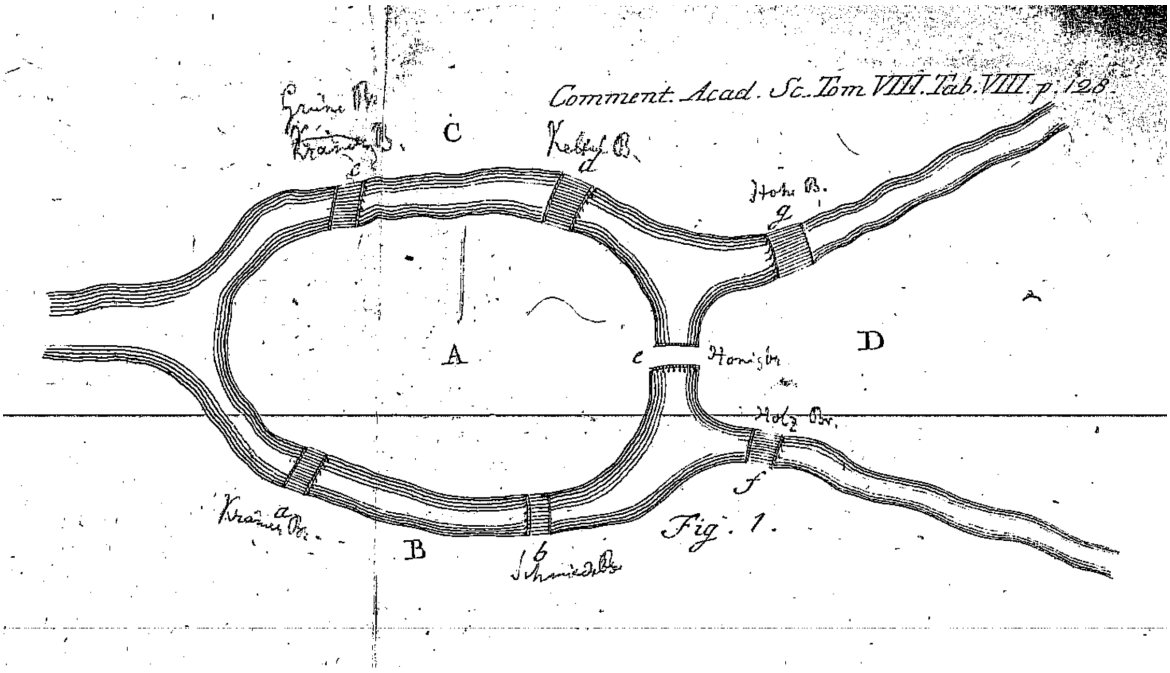
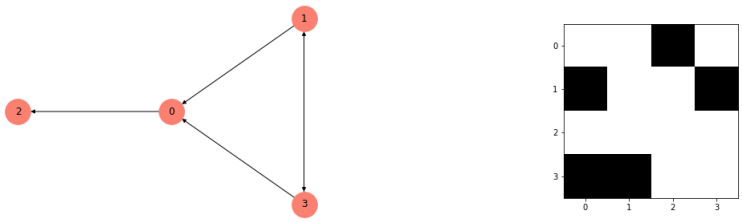



**Fig. 2.1:** The seven bridges of Königsberg, Prussia (today's Kaliningrad). The problem formulation was whether it would be possible to cross all seven bridges, but no bridge more than once. Euler showed this to be impossible, by realizing that landmasses with an odd number of bridges needed to be a starting point. Adopted from [Euler, 1741].

world networks the years following [Watts and Strogatz, 1998] helped shine light on phenomena such as the six degrees of separation and high connectedness of social networks [Milgram, 1967], its remarkable robustness to random perturbation, theories of how they arise from natural processes (preferential attachment) [Barabási and Albert, 1999], and their ubiquity in the real world [Bassett and Bullmore, 2006; Boccaletti et al., 2006; Newman, 2002]. In order to fully appreciate this and other concepts, a more formal introduction to the fundamentals of graph theory and terminology is warranted.

A minimal mathematical description of a network or graph (G) is simply its set of vertices (V) and edges (E): $G = (V, E)$. Further refinements are often made to the description, such as directionality of the edges (like rumour spreading [Boccaletti et al., 2006]) or weighted edges (indicating connection strength as opposed to binary connections). For the sake of mathematical convenience, we may represent any network of N nodes, in a $N \times N$ adjacency matrix A, whose rows and columns represent the nodes, and the matrix entries $a_{i,j}$ denote the connections between the i'th and the j'th node (fig 2.2b). A simple undirected graph will be identical to its own transpose ($\mathbf{A}^T = \mathbf{A}$), easily recognized by a symmetry axis along the main diagonal (Fig. 2.3).




**(a)** Visual representation **(b)** Adjacency matrix representation

**Fig. 2.2:** Two equivalent representations of networks. A visual representation **(a)** can be understood by drawing arrows between nodes to signify a connection. A matrix representation **(b)** ("adjacency matrix") contains the exact same information, but provides more computational convenience. The connection from the ith to the jth node in the graph (entry $a_{ij}$) is indicated by a 1 (black) where an edge is present and a 0 (white) where there is none. For example, the lower left corner in **(b)** represents the connection from node 3 to node 0 in (a). Note that the index starts at 0.

In the case of directed graphs, each entry $a_{i,j}$ represent the link to node $i$ from node $j$ (Fig.2.2b)

With such a generic theoretical framework at hand, it is not surprising that the scope of graph theory has touched a myriad of scientific areas, including but not limited to social networks [Milgram, 1967], the world wide web and network security, epidemiology (viral spread and the development of vaccination strategies) [Pastor-Satorras and Vespignani, 2002; Sattenspiel and Simon, 1988], the engineering of power grids [Pagani and Aiello, 2011], taxonomic classification of viruses [Jang et al., 2019], and inference about the brain in health and disease [Bullmore and Sporns, 2009; Fornito et al., 2015, 2012].

### 2.1.1 Graph metrics

Graph metrics are summary statistics for networks - single numbers that quantify the topological properties of a network. This subsection will cover a few central metrics used in graph theoretical analysis. The majority of the metrics are formulated in terms of discrete binary networks, because they are more intuitive to grasp. However, most of these have been generalized to the case of weighted networks.

#### Terminology

- **Path**: a sequence of steps required to reach one node j from another node i.

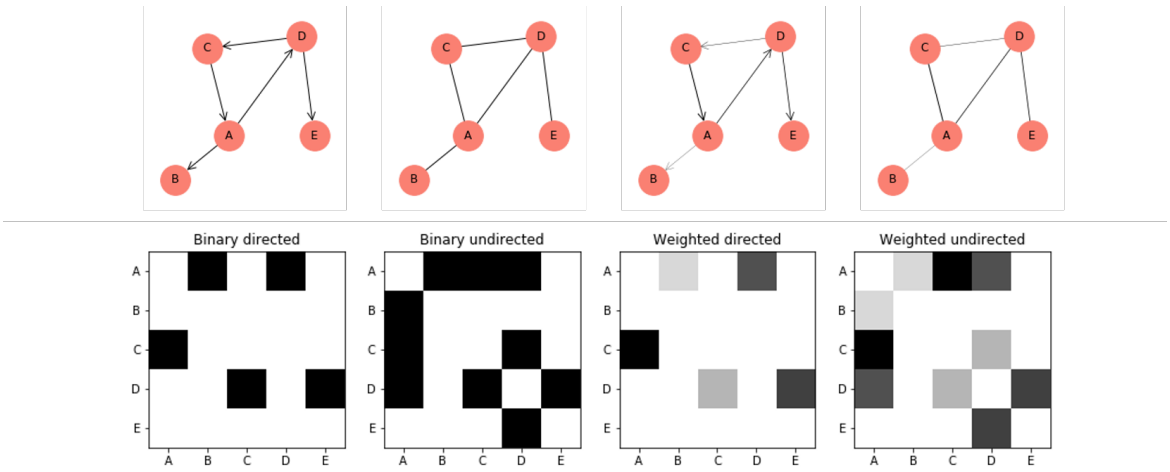



**Fig. 2.3:** Four different descriptions of networks can be interchangeably and equivalently represented both visually (top row) and in matrix form (bottom row). The most basal network description is the mere presence of edges captured in an undirected binary network (second from the left) and can be used to describe many real life networks, like the international airports network. Other networks require directionality, such as disease spread networks - infection is strictly unidirectional, and should be described as a binary directed network (far left). A weighted directed network (third from the left) can be used to describe strictly directional, but non-discrete spread of information. This is the description of effective connectivity. When no information of directionality is available, like in functional connectivity, one must settle with a weighted undirected description (far right). The visual representation is intuitively appealing, but does not provide the same mathematical convenience as the matrix representation (for example, node degree or strength is simply the sum along each axis). Note how the undirected networks are equal to their own transpose ($A^T = A$).

- **Shortest path**: Considering all possible paths connecting two nodes i and j, the path that takes the least number of steps (or hopcount).
- **Adjacency matrix**: a matrix representation of a network where the entries $a_{ij}$ denotes the connection (binary or weighted) from node i to j.
- **Length matrix**: a matrix that represents the length between nodes, commonly made by setting each entry to $\frac{1}{a_{ij}}$. For $a_{ij} = 0$, they are set to $\infty$.
- **Distance/shortest path length**: The most efficient path traversed from node i to j is based on minimizing the hopcount in binary networks. In weighted networks, it also reflects the strength of each weight.
- **Distance matrix**: a matrix which where each entry represents the shortest path between two nodes i and j.
- **Hubs**: nodes of particular importance - characterized by many connections, and central position. Measures of 'hub-ness' include degree centrality, closeness centrality and betweenness centrality.
- **Small-world network**: a network which is characterized by high clustering but a few long-reaching connections connecting distant nodes allowing for efficient communication / information transfer.
- **Module**: a sub-network characterized by a higher connectivity within a module than between different modules. Functionally similar regions tend to be organized in the same modules.

Although most terms are defined in terms of binary networks, they all can be generalized to weighted graphs.

**Node degree**: $k_i = \sum_{j \neq i \in N} a_{ij}$ The node degree is one of multiple ways to measure node importance - the number of connection it has to other nodes in the network. Equivalent of just summing the rows and/or columns of the adjacency matrix.

**Connection density**: $\rho = \frac{\sum_i \sum_{j \neq i} a_{ij}}{(N)(N-1)}$ Simply the number of edges present as a fraction of the total possible number of edges in the network.

**Shortest path length or hopcount** – The shortest possible route from node A to node B, i.e. traversing the fewest number of intermediate nodes. The number of steps corresponds to path length.

**Closeness centrality**: $L_i^{-1} = \frac{n-1}{\sum_{j \in N, j \neq i} d_{ij}}$, for $d_{ij}$ being the distance defined above. Another measure of centrality is based on its topological distance to all other nodes in the network. This can be calculated by averaging the shortest path length between node i and all other nodes, and taking the inverse of this number.

**Betweenness centrality**: $b_i = \frac{1}{(n-1)(n-2)} \sum_{h,j \in N, h \neq j, h \neq i, j \neq i} \frac{\rho_{hj}(i)}{\rho_{hj}}$, where $\rho$ is defined as the number of shortest path connecting h and j, (i) referring to the subfraction of these that pass through i.

**Eigenvector centrality**: $CE_i = \frac{1}{\lambda_1} \sum_{j=1}^{N} A_{ij} x_j$ Identified by eigendecomposition of the adjacency matrix. Nodes with the largest corresponding eigenvalues have higher eigenvector centrality. This metric also considers the centrality of its neighbors, similar to PageRank centrality. A neuronal interpretation of centrality measures must be rooted in a model of information transfer - be it serial or parallell and transferring or duplicative [Fornito et al., 2016]. On a cellular level, this metric has been found to reflect the firing rate of neurons [Fletcher and Wennekers, 2018].

**Clustering coefficient**: $cc = \frac{1}{N} \sum_{i \in N} \frac{2t_i}{k_i(k_i-1)}$, where $t_i = \frac{1}{2} \sum_{j,h \in N} a_{ij} a_{ih} a_{jh}$

The proportion of a node's immediate neighbors which are also are also neighbors of each other - the cliquishness of the network. Equivalent to the number of triangles around the node divided by the number of possible triangles. A weighted analogue is formulated as follows.

$C_i = \frac{\sum_{j \neq i} \sum_{h \neq j, h \neq i} a_{ij} a_{ih} a_{jh}}{(\sum_{j \neq i} a_{ij})(\sum_{j \neq i} a_{ij}^2)}$ is a weighted analogue, where 0<a<1.

An equivalent interpretation is the probability that two nodes i and j are connected, provided they are both connected to a third node h. (If bob and jane both are friends with emily, what is the likelihood bob and jane also are acquainted?).

**Characteristic path length**: $L = \frac{1}{N(N-1)} \sum_{i \neq j} l_{ij}$, lij is the shortest path as identified by Dijkstra's algorithm (Dijkstra 1959). A measure of functional integration - a low characteristic path length indicates a network with a high degree of functional integration.

**Global efficiency**: $E = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{l_{ij}}$ A strongly correlated measure to characteristic path length, but is regarded as being more accurate than the above. Also deals naturally with disconnected networks (if lij = inf, that only amount to adding 0 in the sum).

**Modularity**: $Q = \sum_{u \in M}[e_{uu} - (\sum_{v \in M} e_{uv})^2]$, How well a network is segregated into functionally specialized modules. $e_{uv}$ is the fraction of between-module connections.

**Assortativity**:

$$\frac{a \sum_{jk} (e_{jk} - q_j q_k)}{\sigma_q^2}$$

The assortativity describes the tendency for nodes to connect to nodes of similar degree/strength. Assortativity is a global metric, and can reflect the resilience of the network - simulation studies removing hub nodes have found that networks that have a greater assortative mixing also are less affected by node deletion [Newman, 2002; Rubinov and Sporns, 2010].

**Small-worldness**: $S = \frac{C/C_{rand}}{L/L_{rand}}$ A network architecture commonly observed in nature. It is characterized by high clustering but a few long-reaching connections connecting distant nodes allowing for efficient communication / information transfer.

## 2.2 Statistical concepts

A straight-forward approach to compare connectomes between subjects or clinical groups would be to perform statistical tests like Student's t-test on each individual edge in the graph, and from this make inference on potential disease-related aberrations. However, the number of edges in a graph scales with the square of the number of vertices, which quickly leads to hundreds or thousands of statistical tests.

### 2.2.1 Multiple testing

The widely used p-value is defined as the probability of obtaining a false positive under the null hypothesis that no effect exists: $p \triangleq P(\text{false positive}|H_0)$. Medical researchers usually set an arbitrary threshold for deeming a result as statistically significant or not (commonly p=0.05 or p=0.01). These numbers are of course completely arbitrary, and the term 'statistically significant' has been critizised by statisticans for its inherent ambiguity [Amrhein et al., 2019; Bennett et al., 2009; Ioannidis, 2005]. A significance threshold, or critical value (p*) of 0.05 only means that if no true effect exists, we will still observe false positives in one out of twenty instances (provided that experimental data is completely unbiased and the statistical assumptions such as normality are justified).

If we perform a t-test on each edge in a connectome of 80 nodes, we are faced with $80 \times 79 \div 2 = 3160$ tests, an expected $5\% = 158$ of which will produce a false positive. Thus, comparing the connectomes of two randomly selected groups of people will lead to on average 158 "significant" hits. The most obvious solution is to lower the critical value to a more strict threshold. The family-wise error rate (FWER) is the probability of observing *at least one or more* false positives. For example, we might set a new threshold $P^* = 0.05 \div N_{tests} = 0.05 \div 3160 \approx 0.000016$. This method is known as Bonferroni correction [Bonferroni, 1936], and is a strong control of the FWER. So as long as we set the FWER to 0.05, we are guaranteed to observe not a even a single false positive in 95% of cases, given a true null hypothesis. That means that we can treat the complete set of statistical tests as a single test with p-value of 0.05. Although Bonferroni correction correctly keeps the FWER below a certain threshold, it is effectively useless when the number of tests is large enough. A p-value of $1.6 \times 10^{-5}$ is so strict that small, but real, group differences in edge weights realistically never will be picked up. This introduces another problem: false negatives. Smaller data sets are especially vulnerable to this, where a small p-value only can be observed when the effect size is large. Larger data sets can yield a small p-value despite a moderate effect size. See BOX 1 for one approach to handle the weak power of the Bonferroni method.

**BOX 1. Controlling the False Discovery Rate (FDR)**

A common solution to the weak statistical power of Bonferroni correction is the Benjamini-Hochberg procedure [Benjamini and Hochberg, 1995]. This approach only weakly controls false discovery rate (FDR). The false discovery rate Q is the proportion of type I errors in a set of statistical tests (Q=FP/(FP+TP)). The procedure is as follows: perform m individual statistical tests. Rank the p-values by size, from smallest to largest, P(1) ... P(m). We define 1<=k<=m to be the largest integer for which the following criterion is satisfied:

$$P(k) \leq i\frac{m}{q*} \tag{2.1}$$

To control the FDR at level q*, we reject all null hypothesis correponding to the P-values P(1), P(2),..., P(k). Being a good trade-off between a low FDR and high statistical power, the Benjamini-Hochberg procedure is widely used.

### 2.2.2 Network based statistic

For the case of networks, there exist yet an even more more suitable alternative, further improving statistical power [Zalesky et al., 2010]. Network based statistic (NBS), takes advantage of the fact that the connections in a network are far from independent (an overarching principle in network science), or in the words of the authors: "To potentially offer a substantial gain in power, the NBS exploits the extent to which the connections comprising the contrast or effect of interest are interconnected".

## 2.3 Intraclass correlation and test-retest reliability

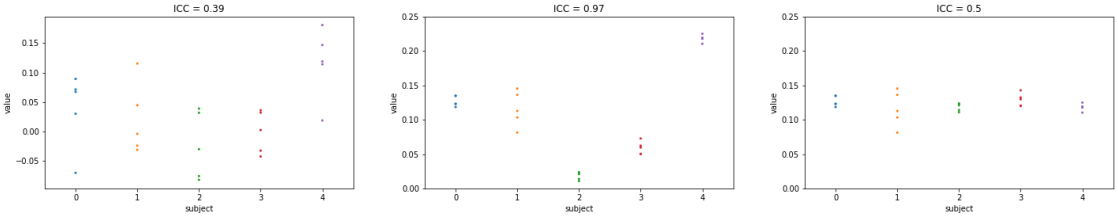
Reproducibility is at the core of empirical science. In statistics test-retest reliability can be informally defined as the degree to which any given estimate yields the same answer upon multiple measurements. This reflects the extent to with the results are consistent - serving as an index for reliability. The intra-class correlation coefficient (ICC) is just that - a coefficient of consistency [Shrout and Fleiss, 1979]. For the purpose of this work, the ICC is used to assess the appropriateness of graph metrics through a test-retest study design.

### 2.3.1 Motivation

As a motivation, consider the following example. We have obtained five scans of five individuals (25 in total), and estimated a graph metric M from each subject. To no one's surprise, the value of M varies slightly between trials, due to random scanner instabilities or other factors like the psychological state of the subject being scanned. The spread (measurement error) between each measurement can be visualized in a dot plot (Fig. 2.4). In panel (a), there is considerable spread within each subject, and little variation between subjects. A more ideal situation is painted in panel (b), in which there is little within-subject variability, and much between subjects. To quantify the difference between situations (a) and (b), we may calculate the variance for each subject, finding that indeed (b) is more reliable. However, a third situation may arise, with both a low within-subject variance and a low between-subject variance (panel (c)). The reliability in this situation is comparable to situation (a), as the graph metrics are not very informative to discriminate between subjects. Thus, both between and within

subject variability must be taken into account. ICC quantifies this by expressing the relationship of variance between subjects to the total variance of interest [Li et al., 2015; McGraw and Wong, 1996]:

ICC = between subject variance / total variance




**(a)** high within, low between **(b)** low within, high between **(c)** low within, low between

**Fig. 2.4:** ICC dotplots. Intuitively, a low between-subject variance and high within-subject variance should yield a low to moderate ICC (a). The opposite situation should yield a high ICC (b). A low within-subject variance is necessary but not sufficient to yield a high ICC score (c).

Multiple definitions of ICC have been described in the litterature [Shrout and Fleiss, 1979], but all can be understood in the context of analysis of variance (ANOVA). [Shrout and Fleiss, 1979] introduce the notation ICC(m,k), where m is the 'model' and k is the 'form'. m can take on a value 1, 2 or 3, referring to the underlying ANOVA model: one-way random, two-way random or two-way mixed respectively. The form is generalized to either k=1 or k>1, alluding to the number of measurements per data point (i.e. k>1 implies averaging multiple observations per data entry). This text only covers the k=1 case. A one-way ANOVA model can be summarized as

$$Y_{ij} = u + a_i + e_{ij} \tag{2.2}$$

where $u$ is the global mean, $a_i$ is the group effect of a particular group $i$ [1], and $e_{ij}$ is the residual term modelled as Gaussian noise: $e_{ij} \sim N(0, \sigma^2)$ This model is thus only concerned with the difference between groups/subjects, and does not attempt to model the effect from the rater. The two-way model is similar, but has an additional explanatory variable:

$$Y_{ij} = u + a_i + b_j + e_{ij} \tag{2.3}$$

[1]The term group only refers to any entity from which there are multiple measurements. Thus a single individual is considered a group in a test-retest study.

In the ICC framework, $b_j$ is the effect from the rater who is doing the measurement. The rater can for instance be the clinic, scanner, processing pipeline, or experimental measurement condition. Thus the two-way model accounts both for the group effect and the rater effect. (interactions can also be modeled, but this will not be covered here). ICC(2,1) is further distinguished from ICC(3,1) by how they treat the effect of the rater - as a random effect (case 2) or a fixed effect (case 3). In other words ICC(2,1) treats $b_j$ as a random variable, while ICC(3,1) treats it as fixed. How to calculate the coefficients is covered next.

Define the following sources of variability (mean sum of squares): MSB = between-subject variance; MSW = within-subject variance; MST = between-rater variance; MSE = irreducible error variance. MSB is often referred to as the treatment mean sum of squares (MSTr) in the ANOVA literature. The different ICCs can be computed as follows [Li et al., 2015], McGraw and Wong:

$$
\begin{aligned}
ICC(1,1) &= (MSB - MSW)/(MSB + (k-1)MSW) \\
ICC(2,1) &= (MSB - MSE)/(MSB + (k-1)MSE + k(MST - MSE)/n) \\
ICC(3,1) &= (MSB - MSE)/(MSB + (k-1)MSE)
\end{aligned}
$$

The mathematical definitions of MSB, MSW, MST and MSE plus accompanying Python code are provided in Appendix A.

### 2.3.2 Interpreting the intraclass correlation coefficient

Out of the many available ICC variants to chose from, it is not always obvious which one to use, especially since the result may differ considerably [Trevethan, 2017]. Therefore multiple attempts of varying technicality have been made to guide the selection process [Koo and Li, 2016; Li et al., 2015; Shrout and Fleiss, 1979; Trevethan, 2017]. Interpretations vary between case 1, 2 and 3. Case 1 is used to model situations in which each subject is rated by a different (randomly selected) rater - no two groups/subjects are judged by the same rater. The second case is used when the measurements are obtained by a fixed set of raters, each measuring a subset of the groups. ICC(2,1) is designed to be generalizable to a completely new set of raters. The third case, ICC(3,1) assumes all groups are measured by the same set of raters, and does not consider the

inter-rater variability (MST) [Li et al., 2015]. Thus the results from ICC(3,1), which treats the raters as fixed, are not generalizable to a new set of raters [Li et al., 2015; Trevethan, 2017]. Another interpretation of the distinction between model 2 and 3 may help clear up some confusion: ICC(2,1) represents the absolute agreement between the raters, and ICC(3,1) considers only the consistency (e.g. relative agreement) [2]. That means ICC(3,1) allows for systematic bias between the raters [Li et al., 2015], as long as it is consistent - for instance an instrument that is wrongly calibrated and always reports too high values. For test-retest studies, ICC(3,1) is often regarded as a suitable choice [Müller and Büttner, 1994] because there is ideally just a single rater [Braun et al., 2012; Trevethan, 2017]. The core idea is that metrics with higher ICC values are both robust to noise and informative, while those with low ICC values are vulnerable to noise and provide little to no value as biomarkers.

## 2.4 Connectivity modelling

To study the functional connectivity of the brain, one must extract useful information from the time series of roughly 100 000 voxels. Assuming that certain brain regions are functionally coherent, we can reduce the computational complexity by parcellating the cortex into a set of anatomically labeled regions, and extracting the BOLD time courses (usually the mean voxel signal). Alternative (functional) methods exist to define nodes, most notably through independent component analysis (ICA) (The mathematical details of ICA is beyond of the scope of this text, see [Beckmann, 2012; Hyvärinen and Oja, 2000] for good introductions to the topic). Once the nodes have been defined, and representative time courses have been extracted, the functional connectivity between every pair of nodes can be modelled using a statistical measure of choice (Fig. 2.5).

### 2.4.1 Pearson correlation

Pearson correlation is the most common way of estimating functional connectivity. Considering two time series $\mathbf{X}$ and $\mathbf{Y}$, the Pearson correlation is calculated as the dot

[2] Here, the literature is somewhat inconsistent. Some (e.g. McGraw and Wong 1996) define both consistency and agreement to be compatible with both case 2 and case 3, but as far as I can tell, the formulas are excessive: case 3 consistency is numerically identical to case 2 agreement, and case 3 agreement is identical to case 2 consistency [Koo and Li, 2016]. In my opinion, this only causes confusion, so the convention of [Koo and Li, 2016] is used.

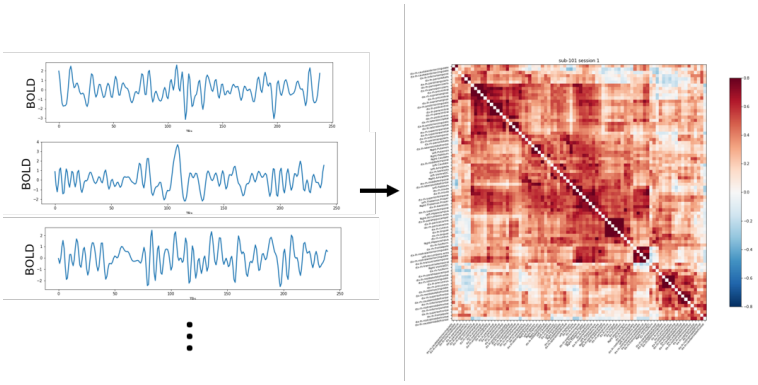



**Fig. 2.5:** Connectivity modelling. A functional connectome can be estimated by computing some measure of similarity between node time courses, typically correlation or partial correlation. The result is visualized as an undirected matrix.

product of the normalized time series. Let **a** and **b** denote the normalized **X** and **Y**, respectively. Then

$$\rho_{XY} = \frac{\frac{1}{n}\sum_{i=1}^{N}(X_i - \mu_{X,i})(Y_i - \mu_{Y,i})}{SD(X_i)SD(Y_i)} = \mathbf{a} \cdot \mathbf{b}$$

This represents the linear coactivation of regional BOLD time courses, and is a natural indicator of regional interaction. However, it tends to overestimate the connectedness of networks due to the transitive property of correlation (see 1.5). Despite its drawbacks, it is most frequently used due to its ease of computation and straight-forward interpretation [Zalesky et al., 2012].

### 2.4.2 Partial Correlation

A trade-off between complexity and simple calculation is partial correlation. Consider a brain with 100 defined cortical regions/nodes and their associated BOLD time series. One caveat is that correlation-based networks display a property known as transitivity: if node X and Y are both strongly correlated to Z, then X is also correlated to Y, irrespective of the actual underlying connectivity between X and Y (see Fig. 1.5). This introduces bias to graph metrics, such as node strength calculation [Power et al., 2012]. This begs for/motivates another calculation of connectivity: partial correlation.

In clinical terms: if we wish to answer any question about the direct functional connectivity of region X and Y, we need to control for the possible involvement of the remaining 98 regions $\mathbf{Z} = [z_1, z_2, z_3 \cdots z_{97}, z_{98}]$, where $z_i$ is the time series of the $i^{th}$ region.

Mathematically, the partial correlation between two normally independently identically distributed (i.i.d.) variables X and Y controlled by **Z** is denoted $\rho_{XY \cdot \mathbf{Z}}$. It is computed as the Pearson correlation between the residuals $e_{X,i}$ and $e_{Y,i}$ after regressing out the effects of **Z**:

$$\rho_{XY \cdot \mathbf{Z}} = \frac{\frac{1}{n}\sum_{i=1}^{N}(e_{X,i} - \mu_{e_X,i})(e_{Y,i} - \mu_{e_Y,i})}{SD(e_{X,i})SD(e_{Y,i})},$$

where the error term is the residual of the model estimate, $e_{X,i} = X_i - \hat{X}_i$, for the linear model:

$$\hat{X} = A\beta$$

In practical terms, the residuals can be interpreted as the remaining activity of regions X and Y after removing any influence from **Z**. Partial correlation has been recommended as a good intermediate between high model complexity and performance for larger graphs. Another common approach which shows promise is sparse inverse covariance, or precision [Smith et al., 2011; Varoquaux et al., 2010]. Like partial correlation, precision is only concerned with direct connections, but it returns a sparsely connected graph, which is often favoured when acquisition times are shorter [Smith et al., 2011].

# Chapter 3

# Aims

This thesis has in part been a contribution to the ongoing FRIMEDBIO-funded project, "Brain-Gut Microbiota Interaction in Irritable Bowel Syndrome: A multidimentional Approach". (https://braingut.no/). The work took place at the Neuroinformatics and Image Analysis Laboratory, Neural Networks Research Group, Department of Biomedicine and the Mohn Medical Imaging and Visualization Centre in collaboration with The Norwegian National Center for Functional Gastrointestinal Disorders, Department of Medicine, Haukeland University Hospital. The thesis aims to explore machine learning as a tool in characterizing the functional connectivity of IBS patients from fMRI data, through the lens of network science. A secondary aim is to assess the reproducibility of graph metrics as clinical biomarkers in general.

# Part II

# Experimental

# Chapter 4

# Experimental

## 4.1 Methodological experiments

Neuroimaging based biomarkers of psychiatric and neurological disorders is a promising area of research [Bachmann et al., 2018; Birn et al., 2013; Fornito et al., 2015]. This approach warrants a rigorous methodologically justified practice, especially in a field as controversial as fMRI in terms of reproducibility [Bennett et al., 2009; Kriegeskorte et al., 2009; Vul et al., 2009]. A simulated test-retest study was performed on HCP data using twelve common graph metrics to provide a relative ranking from least to most robust to random noise. Another alleyway was explored with regards to the effect of scan length: what is the time needed for convergence? This question was explored through expanding window analysis on clinical data from IBS patients and healthy controls (HCs). The answer to this question has important implications for clinical practice.

### 4.1.1 HCP network matrices

Methodological analysis was performed on adjacency matrices provided by the Human Connectome Project [Glasser et al., 2013], available from `db.humanconnectome.org`. 15 minutes of minimally preprocessed rs-fMRI data from healthy adults (HCP1200, 4800 time points) was the basis for connectome estimation. Nodes were defined using group-ICA (MELODIC), at 15, 25, 50 and 100 components. The final matrices had been generated using Pearson correlation between each component. The matrices,

which were provided as Z-scores, were transformed back to correlation prior to analysis, using the equivalence: $\rho = tanh(Z)$.

Two experiments were performed with the purpose of investigating the robustness and reproducibility of graph metrics applied to rs-fMRI data. First, a test-retest reliability simulation study was conducted by adding Gaussian noise to the network edges and recalculating graph metrics at each iteration. The obtained results were used to compute the ICC of each metric as a measure of reliability between the measurements. Secondly, graph metric stability over time was evaluated qualitatively by calculating the metrics using an expanding window framework.

### 4.1.2 Test-retest reliability of graph metrics

Seven subjects were selected at random from the HCP1200 data set. Four level of Gaussian noise ($\mu = 0$, $\sigma = 0.05$, 0.1, 0.2 and 0.4) was simulated and applied to the correlation matrix entries. The noise was applied additively (that is, uniformly across all pairs of nodes). Following the noise addition, the matrices were corrected for asymmetry and thresholded at five levels (p = 0.1 ,0.2, 0.4, 0.6, 0.8, where (1-p) is the final density of the graph) to remove the weakest edges. A small set of graph metrics were calculated from the thresholded networks (for details, see4.2.5). The retest was performed eight times per subject, simulating eight different 'raters'. This was used to compute the intraclass correlation coefficient ICC(3,1) for each metric at 20 noise/threshold combinations. The metrics were finally ranked according to their mean ICC from the 20 simulations. This whole procedure was done at 15, 25, 50 and 100 ICA dimensions. A 'consensus' rank was then determined as a final assessment of the relative graph metric reliabilities.

### 4.1.3 Expanding windows and graph metric convergence

To the extent that functional connectomes can be considered fixed in time, their estimates converge to the ground truth as scan duration is increased. What is the minimum time sufficient to obtain good graph metric estimates? This question was investigated using expanding windows analysis. Preprocessed data from a healthy subject (see below for details on preprocessing and network modelling) was truncated at different lengths ranging from 20 TRs to the full scan length of 240 TRs (TR=2s).

With 20 TR increments, a total of 20 adjacency matrices (Pearson correlation) were estimated. Global graph metrics were calculated from each adjacency matrix. The change in each (normalized) metric was recorded between each increment, serving as an index for convergence (a change of zero means convergence). The analysis was performed at multiple noise and threshold levels, and repeated using two brain atlases (see below).

## 4.2 Clinical experiments

Functional connectivity has been implicated in IBS with emphasis on the BGA and gut microbiota [Labus et al., 2019; Mayer et al., 2015]. The purpose of the study was to investigate the difference in functional connectivity between IBS patients and controls, to test whether IBS can in part be explained by dysregulated brain connectivity, and if disease status could be predicted from connectivity signatures using machine learning.

### 4.2.1 Experimental setup and MRI protocol

MRI scans were obtained from IBS patients and healthy controls in order to investigate functional alterations in the brains of IBS patients. All measurements took place at Haukeland University Hospiatl using a GE Sigma HDxt 3.0T system (8-channel standard head coils). The experimental setup was as follows: first a high resolution anatomical image was acquired, followed by eight minutes of resting state fMRI. As an intervention to elicit clinical symptoms, the subjects were instructed to drink 3 cups of Toro meat soup. Another round of anatomical and resting state scans was repeated. The study design is summarized in Fig. 4.1.

Fifteen individuals diagnosed with IBS (11 female, 4 male) and fifteen healthy controls (10 female, 5 male) were selected for the study, in accordance with the ROME II and III criteria. Mean ages of participants were $38.6 \pm 12.4$ and $35.8 \pm 13.2$ for IBS patients and healthy controls, respectively. The genders were also balanced in terms of the age ($37.2 \pm 11.4$ and $37.2 \pm 12.9$ for male and female). Each subject reported zero history of psychiatric illness. The control group was otherwise free of any other functional gastrointestinal disorder than IBS, like functional dyspepsia, and had no severe gut related problems. To control for the role of food intake, patients were instructed to avoid eating for a minimum of 4 hours prior to being scanned.

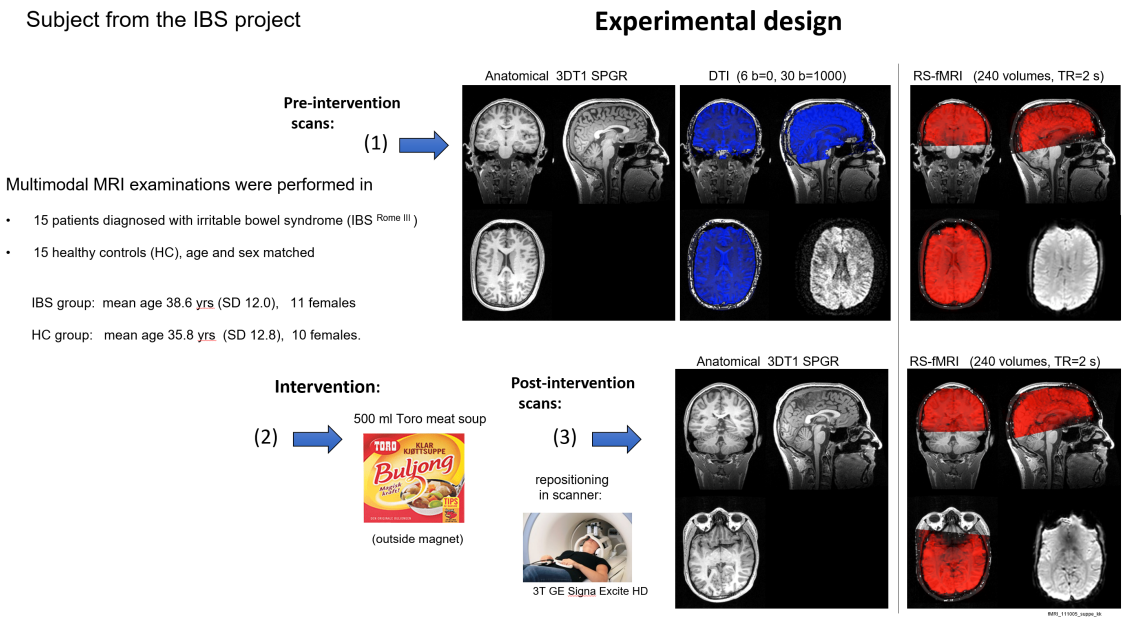



**Fig. 4.1:** IBS imaging: Experimental design

### Anatomical imaging

A high-resolution anatomical T1-weighted volume was obtained using a spoiled gradient-echo sequence (GRE) (TR=7.76 ms, TE=2.95 ms, TI=500 ms, flip angle=14°, and FOV=256x256 mm2, 188 1.0mm sagittal slices), yielding approximately isotropic voxels with resolution 1.00mm x 1.0156mm x 1.0156mm.

### Functional imaging

Following the anatomical acquisition, a BOLD fMRI scan was carried out using an echo-planar imaging (EPI) sequence with the following parameters: flip angle = 90, TE = 0.03, acquisition matrix PE = 96, voxel dimensions = 1.72x1.72x3.50$mm^3$. Both the scan before and after soup intervention lasted for a total of four minutes.

## 4.2.2 File Management

DICOM images were converted to the NIfTI file format with an accompanying JSON file using the Python library `dcm2niix` [Li et al., 2016]. A particular bug in the software

required an extra step for converting slice timing from milliseconds to seconds. The NIfTI files were organized into the BIDS (Brain Imaging Data Structure) directory structure for ease of automating processing pipelines [Gorgolewski et al., 2016]. Proper formatting was ensured using the BIDS-validator available from `https://github.com/bids-standard/bids-validator`.

### 4.2.3 Anatomical segmentation

A much discussed question in brain connectivity is what is the appropriate way to delineate the cortex into meaningful nodes of a network? This question is easy to answer at the microscopic scale, but is controversial at the mesoscopic and macroscopic scale Fornito et al. [2016]. Two principled ways exist to define nodes: anatomically and functionally. For this analysis, the former was used due to its straight forward interpretation and the lack of need for manually labeling nodes and removing elements corresponding to noise, as is required with ICA. Freesurfer [Reuter et al., 2012] offers individualized partitioning of gray matter into cortical and subcortical regions. Two default atlases were used in this work: the coarse Desikan-Killiany atlas [Desikan et al., 2006] and the finer Destrieux atlas [Destrieux et al., 2010], allowing for analysis at two levels of granularity. Anatomical parcellation was carried out in FreeSurfer V6.0.0 (`freesurfer-Linux-centos6_x86_64-stable-pub-v6.0.0-2beb96c`), `http://surfer.nmr.mgh.harvard.edu`. A single T1-weighted NIfTI image (pre-soup) was used as input in FreeSurfer's standard recon-all pipeline, as exemplified below.

```
recon-all -i subj1 -s subj101 autorecon-1 -3T
```

The 3T option was added to correct for intensity inhomogeneities specific to 3 Tesla scanners. No further refinements or manual intervention (like correcting the skull-strip) were made, accepting outputs as is. The resulting segmentations were later used as masks for extracting BOLD time series. FreeSurfer's pipeline is advantageous in terms of accuracy compared to other methods of template registration (like FLIRT or FNIRT in FSL), but requires significantly more computing resources.

### 4.2.4 Functional preprocessing

By correcting for and removing various sources of noise, functional preprocessing serves to highlight signal fluctuations that correspond to neural activity. All preprocessing steps were executed out in AFNI (`AFNI_19.0.24 'Tiberius'`) [Cox, 1996]. A T1 weighted image (intensity corrected in FreeSurfer) was input as anatomical reference for coregistration between structural and functional modalities. The final processing script was generated using the AFNI script afni_proc.py (representative script is provided in Appendix A.

A minimal processing pipeline was initiated as follows. The two first TRs were removed to let the scanner reach steady-state. TRs with too much head motion were disposed of, setting the motion limit to 0.2 (censoring). Timeseries exhibiting marked spikes (an artifact arising due to motion) were truncated. Slice timing correction was performed. Volume registration (EPI-EPI alignment) was performed as a means of motion correction, and coregistration with a T1 weighted volume was done from the most stable TR. Gaussian smoothing (fwhm=4) was performed to reduce spatial noise. Motion parameters and their first-order derivatives were regressed out using a linear model. Bandpass filtering was performed (in the same step) to remove signals outside the range of 0.01-0.1 Hz.

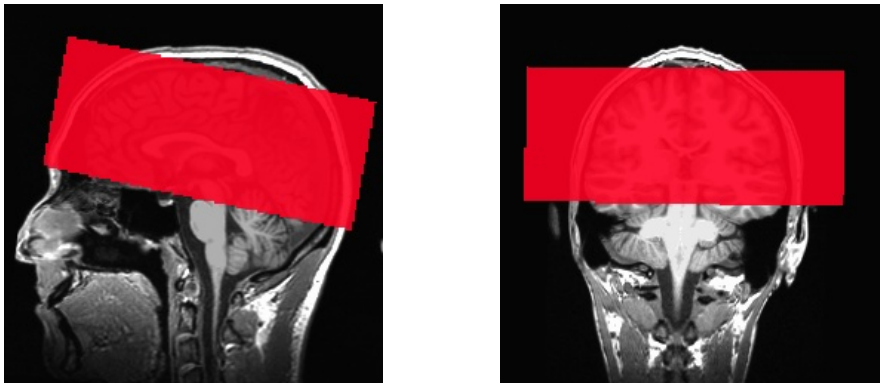
### 4.2.5 Network modelling

#### Defining the nodes

Functional connectivity matrices were estimated for each subject, pre- and post soup intervention, yielding $2 \times 30 = 60$ adjacency matrices. FreeSurfer segmentations were used as define the network nodes. Both the Desikan-Killiany atlas [Desikan et al., 2006] and Destrieux atlas [Destrieux et al., 2010] were investigated. Representative time courses from the anatomically defined nodes were estimated using the mean time course for each voxel within the region of interest.

Functional connectivity was estimated between every N(N-1)/2 pair of time series. All steps were performed using functions from the `nilearn` Python library [Abraham

et al., 2014]. Some regions were excluded from the analysis due to a fully absent or incomplete overlap between the T1 weighted image and the BOLD data, corresponding mainly to the temporal lobe and the cerebellum (see 4.2). The exclusion criterion was that at least one subject/session had zero overlap. The regions (FreeSurfer labels) that met the criterion are provided in Appendix A. The resulting adjacency matrices were of sizes 80x80 (Desikan-Killiany) and 154x154 (Destrieux).



**(a)** Saggital view **(b)** Coronal view

**Fig. 4.2:** Extent of overlap between functional and anatomical modalities. The red square marks the extent of coverage from the EPI data superimposed on the T1 weighted image. Large areas are lost, including significant regions of temporal cortex and almost the complete cerebellum.

### Connectivity measures

Functional connectivity was estimated using three measures: Pearsson correlation, partial correlation and sparse inverse covariance (precision). The latter was estimated using the graphical lasso algorithm implemented in `scikit-learn` [Pedregosa et al., 2011].

### Network thresholding

Graph thresholding is canonical practice when dealing with weighted graphs, reasoned to remove spurious connections. There is however, no globally optimal numerical rule governing which connections are to be deemed spurious. Therefore it is common to use a range of thresholds, even though it imposes computational costs on the analysis

[De Vico Fallani et al., 2014]. Thresholding can be performed either globally, by setting all values below a certain threshold to zero, or proportionally, in which only the top $p{\cdot}100\%$ percent strongest edges are kept (density preserving). The former fixes the network to a certain strength range, while the latter keeps the network density fixed. Because overall connectivity strengths often vary greatly between (and within) subjects, the network density will be artificially increased for high-strength networks. Network density in itself affects many graph metrics, so global thresholding has been discouraged in favor of proportional thresholding. Despite its caveats [van den Heuvel et al., 2017], proportional thresholding was used in the current work. Another unresolved issue originates from the fact that adjacency measures often allow for negative correlations. It is hard to interpret the physiological meaning [Poldrack et al., 2011], although one study reported significant correlation between anti-correlations and shortest path length [Chen et al., 2011a]. While some [Labus et al., 2019] choose to ignore below zero entries altogether, they were included for the purpose of this work.

### Calculating graph metrics

A selection of graph metrics were computed from the acquired network matrices. The analysis was carried out in Python using the Python implementation of the brain connectivity toolbox, `bctpy` [Rubinov and Sporns, 2010] and `pynets` (`https://github.com/dPys/PyNets`). A total of twelve graph metrics were selected: modularity quotient Q, node strength (total positive, total negative, and nodewise positive and negative), assortativity, betweenness centrality, eigenvector centrality, clustering coefficient, transitivity, global efficiency, characteristic path length, and small-worldness. Some of these metrics (node strength, betweenness centrality, eigenvector centrality and clustering coefficient) were computed on a node-by-node basis (e.g. each node in the network has an associated clustering coefficient). Other metrics like characteristic path length were calculated for the whole network, referred to as a "global" metric. Two approaches were implemented: one in which the nodewise metrics were averaged across all nodes (producing 12 features) and one using the full nodewise calculations, yielding feature vectors of 328 and 624 elements for the Desikan-Killiany and Destrieux atlas, respectively.

### 4.2.6 Machine learning

Unlike statistical measures like correlation, the success of a predictive model is evaluated on how well it performs on unseen data. Therefore machine learning can be used in a completely unbiased way to test the viability of imaging-based biomarkers. To this end, multiple machine learning approaches were attempted, aiming to find discriminatory features between IBS patients and controls based on multivariate network-derived features.

#### Evaluating the model

Training and testing a machine learning model on the same data set is methodologically flawed, because one can always find a mathematical function which gets arbitrarily close to a desired decision boundary. Such overmodelling is known as *overfitting* and leads to poor performance on unseen data. The canonical scheme of training a machine learning classifier is therefore to split the data into two independent sets: the training set and test set. The model is fit to the training set, and evaluated by its performance on the test set.

**Cross validation** Provided enough data, one may choose to perform cross validation (CV) as a step to optimize model hyperparameters (like regularization strength). CV consists of further dividing the training set into multiple "folds" of a given size. The classifier is fit on all folds but one, and validated on the remaining fold. This procedure is repeated across all $k$ folds, such that every fold is used exactly *once* as a validation set. If the folds are single observations, it is referred to as leave-one-out cross validation (LOOCV). CV is tested across multiple model hyperparameters, the best performing of which (by a given performance measure) are selected to train the final model, which is evaluated on the hitherto untouched test set. This complete procedure would usually require hundreds to thousands of samples to perform well for challenging problems. Considering the complex problem of classifying diagnostic status and symptomatic measures from functional imaging data, and the small sample size (N=30 subjects),

this approach was deemed unfeasible. Therefore, the models were evaluated based on their average performance by CV, using accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

As such the whole pipeline was incorporated into a leave-one-group-out cross validation (LOGOCV) scheme. As opposed to regular LOOCV, LOGOCV ensures that no single subject is used both in the test set and validation set, which would violate the criterion of independence, considering there are two scans per subject [Pereira et al., 2009]. A drawback to CV is that no final model is produced: the results only reflect the degree to which using a machine learning approach is viable to begin with.

### Feature scaling

Many machine learning classifiers are sensitive to the absolute value of its inputs. For instance k-nearest neighbors is based on the Euclidean distance between vectors in high dimensional feature space. As such, it will underestimate the importance of features with small absolute values as these contribute disproportionately little to the Euclidian distance between a pair of points. Feature scaling is thus an essential step for many classifiers. The graph metrics were therefore scaled to zero mean and unit variance prior to training any algorithm:

$$\mathrm{M}_{norm} = \frac{M - \mu_M}{\sigma_M}$$

### Prediction from global graph metrics

Discrimination between IBS patients and HCs was attempted using the global metrics as feature vectors. The following classifiers were attempted, available from the `scikit-learn` Python library [Pedregosa et al., 2011]: k-nearest neighbors (3 neighbors), random forest, support vector machine (SVM) - linear and with a radial basis function (RBF) kernel, Gaussian naive Bayes and logistic regression (with slight L1 regularization). The classifiers had been selected based on criteria of simplicity and the *status quo* in related problems. Simple classifier with regularization were favored as these deal well with multiple noisy features [Pereira et al., 2009]. The linear support vector classifier does implicit regularization, determined by regularization parameter C.

Logistic regression can also incorporate a regularization. All hyperparameters were set to default (`scikit-learn` version: 0.20.1). The approach was tested for graph metrics computed from Pearson correlation, partial correlation and precision, at three threshold: 0.3, 0.5 and 0.8.

Patient reports of pain and nausea were also attempted as targets for the same machine learning classifiers. Patients had been asked to score their experience of visceral pain and nausea on a scale from 0 to 100, before and after soup intervention. The groups were discretized into two classes (high versus low pain/nausea), by splitting the observations at the median. No explicit feature selection or dimensionality reduction was done for the global graph measures.

### Prediction from nodewise graph metrics

**Feature selection and dimensionality reduction** A recurring obstacle in neuroimaging is when the number of features far exceed the number of samples ($p \gg N$). The primary concern with this is related to the curse of dimensionality: an ever-increasingly small proportion of the feature space is covered by the sample when the number of features is increased. The most common solution is to perform some sort of feature selection or another dimensionality reduction technique prior to training a classifier, effectively reducing the number of features per sample (see BOX 2).

**BOX 2. Common methods of feature selection**

Feature selection is the act of choosing a subset of the (best) available features to train the algorithm, while other dimensionality reduction techniques transform the original data set into a new space (e.g. PCA, ICA). This may not only removes redundant, correlated and noisy features, but it can also improve classifier performance and speed. Feature selection approaches are divided into *filter* methods, *wrapper* methods, and *embedded* methods.

Filter methods are performed prior to and are completely independent of the classification algorithm. Examples are univariate statistical tests such as the t-test and ANOVA. Wrapper methods are based on the actual classification performance of the classifier itself, using different combination of features and accounts for multivariate relationships in the data [Guyon and Elisseeff, 2003]. Recursive feature elimination (RFE) is an example of this, and works by training a classifier,

then removing feature of least importance (e.g. lowest beta weight). The process is repeated, until we are left with the desired number of 'best' features. This is only possible to do for models which have an inherent way of ranking feature importance [Pereira et al., 2009]. Embedded methods are part of the classifier itself, which performs a sort of implicit feature selection. Decision trees are an example of this, which select the features which maximizes information gain (the change in entropy), while L1 linear regression applies a penalty to having too many and too high beta weights. For the current work, RFE was attempted in combination with linear support vector machine, random forest and logistic regression (with lasso regularization).

#### Recursive feature elimination and nodewise graph metrics

The $p \gg N$ problem was issued using recursive feature elimination. This was done in combination with linear SVM, random forest and logistic regression, to select the top 5, 10, 15 and 20 features. The final model was trained on the top features. The whole pipeline was incorporated into a LOGOCV scheme. The process was repeated for three thresholds (0.3, 0.5 and 0.8) using Pearson correlation, partial correlation and precision. More complex models like neural networks were omitted, as they are more prone to overfitting than simple classifiers. Simple classifiers often outperform complex ones when data sets are small with many featues [Pereira et al., 2009], and are easier to interpret.

## 4.3 Controlling for head movement

Head motion during resting state fMRI scans has a profound impact on measures of connectivity, for example by introducing spuriously edge strengths [Power et al., 2012]. Although most preprocessing steps are specifically concerned with minimizing the effect of motion, there exists no way to completely alleviate all its effects. It is therefore imperative to leave out motion as a possible explanation for any observed group differences in brain connectivity. For the sake of simplicity the overall head motion was summarized by averaging all six degrees of freedom divided by the length of each run. IBS/control group differences were tested for using a two-sample t-test for

both pre-intervention, post-intervention and between sessions (to test if there is an increase or decrease of movement).

## 4.4 Network based statistic

The potential network difference between clinical groups were also assessed with a purely statistically motivated, graph metric-independent approach. Correlation matrices from the Desikan-Killiany atlas were used with NBS at different t-value thresholds as recommended by [Fornito et al., 2016], between 1 and 3. The nodes comprising the resulting resulting pseudonetwork (largest connected component; LCC) were extracted to identify regions of interest.

# Chapter 5

# Methodological Results

Graph metrics are increasingly being explored as biomarkers in neurological and psychiatric disorders, which too often rely on subjective assessment of behavioural traits rather than physiological biomarkers. Graph metrics can potentially fill the hole left by clinicians, as well as increase understanding of the physiological underpinnings of psychological or behavioural states in disease. Although graph metrics offer a theoretically interesting and plausible alleyway, its practicality ultimately comes down to the extent to which graph metrics are consistent between measurements for a given individual at different times and/or experimental nuisances (various sources of noise). The same concern applies to IBS in which the brain, besides being the ultimate arbiter of the subjective experience of pain, partakes in the bidirectional communication with gastrointestinal processes. To increase understanding of graph metric reliability I performed a simulated a test-retest experiment on HCP data and assessed robustness with the intraclass correlation coefficient. The effect on scan duration was also assessed using clinical data from IBS patients and healthy controls using an expanding window analysis.

## 5.1 Test-retest reliability of graph metrics

To investigate the invariance of graph metrics to noise, a simulated test-retest study was conducted on HCP1200 network matrices at various thresholds and noise levels. Only global metrics were explored. Eight raters were simulated, and the reliability was calculated as ICC(3,1) (Python code is provided in Appendix A).

The experiment was repeated at 20 threshold/noise combinations in order to see if thresholding affects reproducibility and to sample a large number of scenarios to get a representative average. The resulting ICC values were plotted in Fig. 5.1 for 15 ICA dimesions. The overall ICC varied substantially not only between different metrics, but also as a function of threshold and noise. As expected, the ICC dropped with increased noise. More interestingly was the observation that strict thresholding was consistently linked to a reduced ICC. The analysis was repeated at 25, 50 and 100 dimensions, and yielded a similar pattern as with 15 nodes (plots not shown).

By averaging the ICC across all threshold/noise combinations, a total ICC was yielded for each metric. This mean score was used to rank the metrics according to their relative robustness. The order was highly consistent between 15, 25, 50 and 100 ICA dimensions, with some exceptions. A score was assigned to each metric based on its rank at each dimensionality. The sum of the scores was used to determine a "consensus" rank, visualized in Fig. 5.2. Furthermore, the ICC was consistently higher for higher dimensions of ICA, a pattern applying to all graph metrics tested.

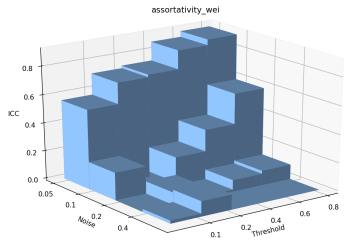
For some threshold-noise combinations the ICC evaluated to negative numbers (Fig. 5.3). This is a well known issue in the ICC literature occurring when the within-group variance exceeds the between-group variance [Müller and Büttner, 1994], imposing interpretability difficulties (what is negative reliability?). These values were set to zero, as suggested by [Bartko, 1976], although others have criticized the practice, calling for alternative approaches completely, like Bland-Altmann plots (see [Müller and Büttner, 1994]). These alternatives were not investigated. As a quality assurance, a random sample of the data yielding negative ICC values were plotted and inspected (a representative example is shown in Fig. 5.3). All were characterized by a low between-subject variability and high within-subject variability, suggesting that setting the ICC to zero was justified.

## 5.2 Sliding windows correlation

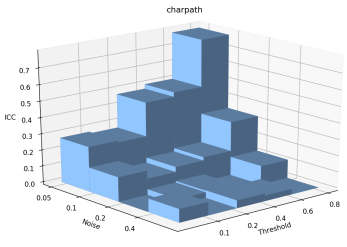
Another aspect of consideration was the graph metric dependency on scan duration. Understanding the convergence time of graph metrics has great importance for experimental design for any study investigating graph metrics. Without invoking the dynamic evolution of graph metrics [Zalesky and Breakspear, 2015], graph metrics are considered fixed parameters for any individual. Under this assumption, correlation

matrices were calculated from the time series at varying window lengths using 20 TR increments. The change between each 20 TR increment was calculated and plotted in Fig. 5.4.

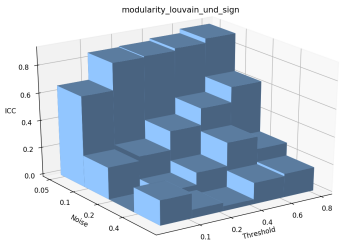
Considerable variation was observed across subjects, but most graph metrics had settled in the vast majority of subjects at eight minutes in the scanner. Some metrics, however, were frequently observed to change more than a full standard deviation between increments even towards the end of the scan. Small-worldness was particularly ill-behaved. The change was larger in the beginning where each increment comprised a larger proportion of the total window. In accordance with what was observed in terms of ICC analysis, stricter thresholds were associated with a weaker graph metric reliability (compare panel (a) to (b) and (c) to (d) in Fig. 5.4). There was also a less pronounced, but clear tendency for the Destrieux atlas to yield more stable graph metrics, which is in line with the above observation that finer granularity is associated with higher ICC.



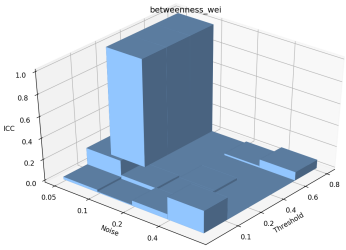

**(a)** Assortativity



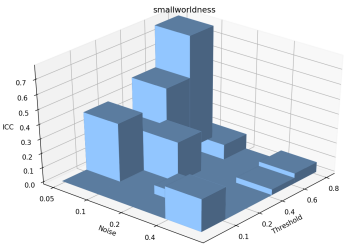

**(b)** Characteristic path length



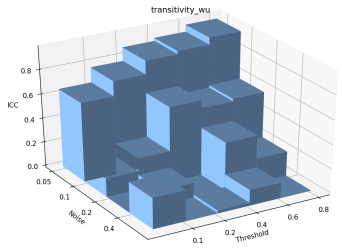

**(c)** Modularity



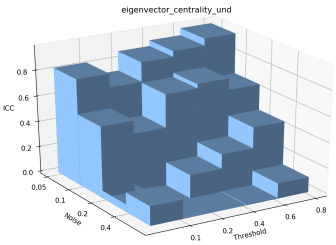

**(d)** betweenness



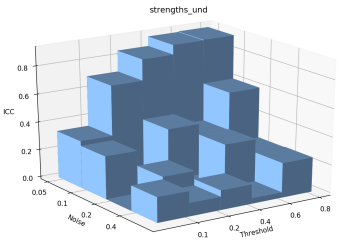

**(e)** Small-worldness



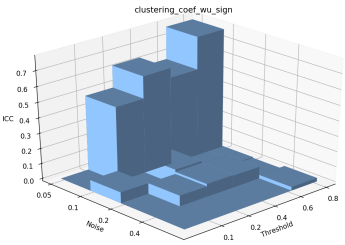

**(f)** Transitivity



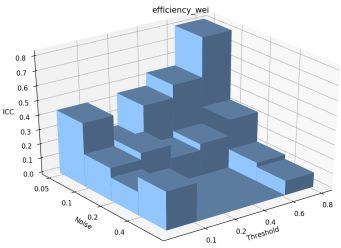

**(g)** Eigenvector centrality




**(h)** Strength




**(i)** Clustering coefficient




**(j)** Global efficiency

**Fig. 5.1:** Test-retest reliability of global graph metrics. The ICC(3,1) was calculated for a set of global graph metrics across a range of thresholds and noise levels. The noise is reported as $\sigma$ from a Gaussian distribution, and the threshold p represents the proportion of edges kept in the graph after thresholding. All metrics were obtained from the HCP1200 data set, ICA (n=15), Pearson correlation.

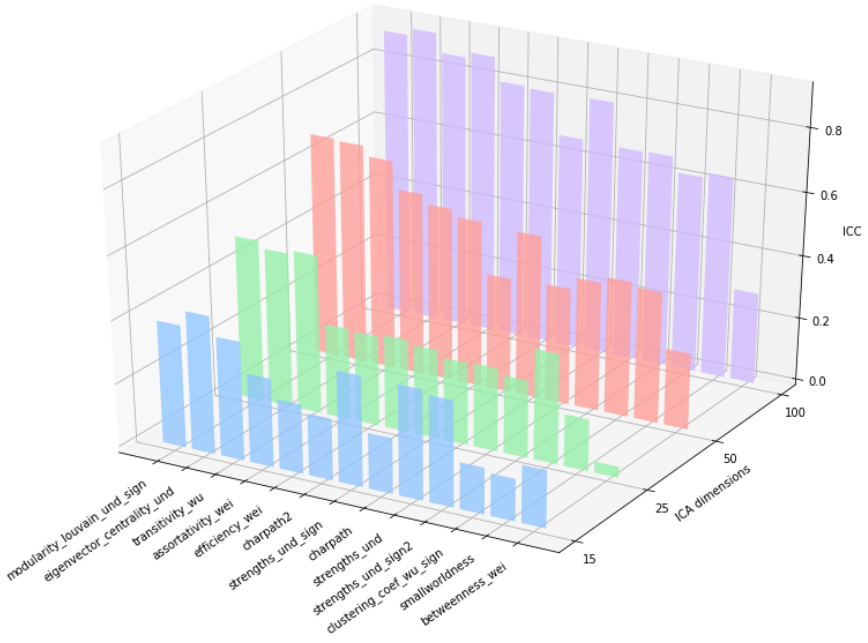



**Fig. 5.2:** Relative graph metric robustness from simulated test-retest experiments of 13 graph metrics, averaged across 4 levels of noise and 5 levels of thresholds. ICC(3,1) was used to quantify the reliability of each metric. The metrics are ordered according to a consensus rank from largest to smallest ICC: Modularity metric Q, eigenvector centrality, transitivity, assortativity, global efficiency (duplicated), strength, characteristic path length, strength (only positive), strength (only negative), clustering coefficient, small-worldness, betweenness centrality.

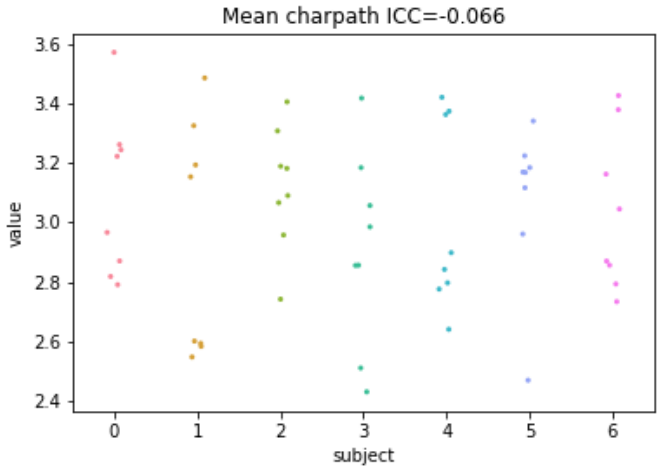
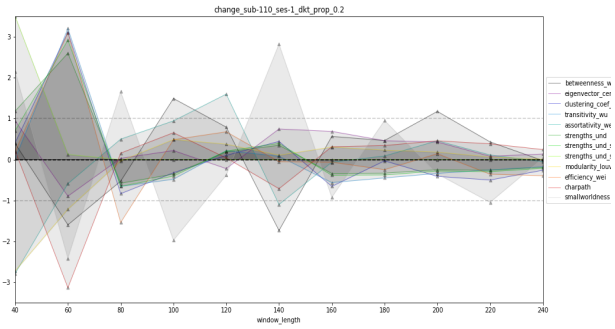
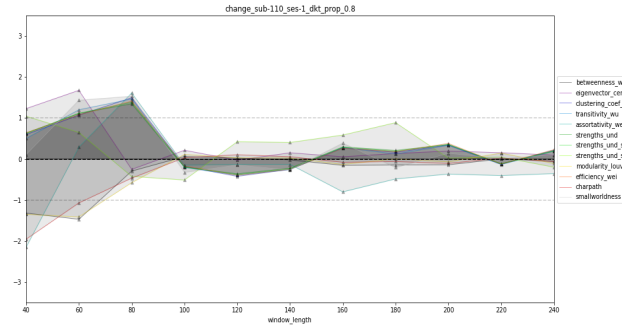



**Fig. 5.3:** Negative intraclass correlation coefficient. Mean characteristic path length was calculated from simulated test-retest experiment on seven subjects from HCP1200, by the addition of Gaussian noise, yielding a negative ICC of -0.066. The dotplot reveals a low between-subject variability, exceeded by a markedly high within-subject variability. Thus, the characteristic path length calculated at this threshold and noise level (threshold=0.1, $\sigma$=0.005, 25-dimensional ICA) is not informative, and serves as a useless biomarker.

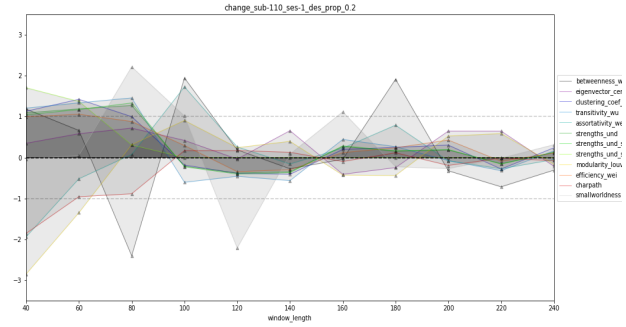
**Fig. 5.4:** Global graph metric dynamics from expanding window analysis of a single subject/scan, with varying parcellation granularity and graph density. Graph metrics were calculated from connectomes estimated by varying time series window length in 20 TR (40 second) increments. The change in the each graph metric (normalized to zero mean and unit variance) between each increment is plotted on the y-axis. The Destrieux atlas (154 regions) yielded slightly more stable graph metrics than the Desikan-Killiany atlas (80 regions). A much more pronounced effect was seen for graph density, favouring high graph density. The same patterns were observed across all subjects, but at varying magnitude. Different subjects were highly variable in the time to converge, some of which did not settle altogether even after the full length of the scan for certain metrics. The y-axis is in units of standard deviation. DKT, Desikan-Killiany; DES, Destrieux.



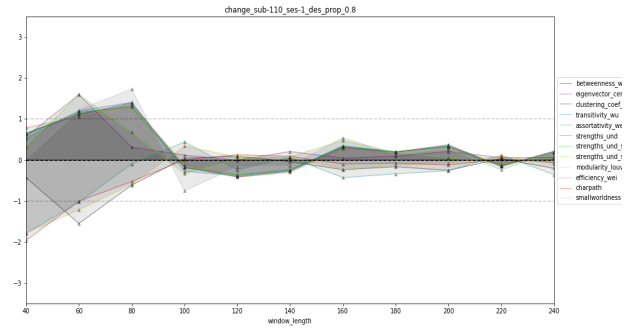

**(a)** DKT, threshold=0.2




**(b)** DKT, threshold=0.8




**(c)** DES, threshold=0.2




**(d)** DES, threshold=0.8

# Chapter 6

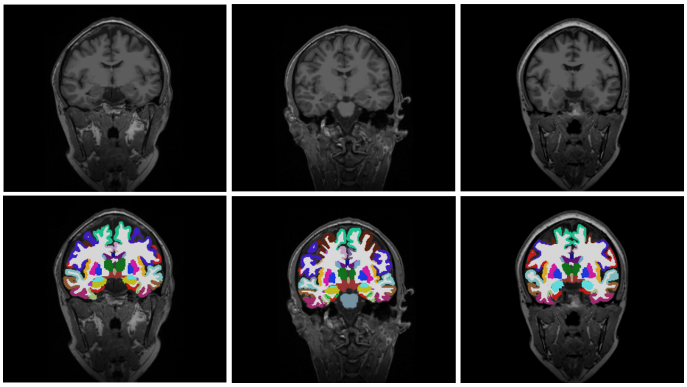
# Clinical Results

60 rs-fMRI scans from 30 subjects with or without IBS (two scans per subject, before and after soup intervention, eight minutes each) were used as the basis for functional connectome modelling. Twelve common graph metrics were calculated from the estimated networks, both global and nodewise metrics. The viability of using machine learning classifiers to predict diagnostic status from these biomarkers were explored. The same strategy was later applied to predicting self-reported behavioural measures.

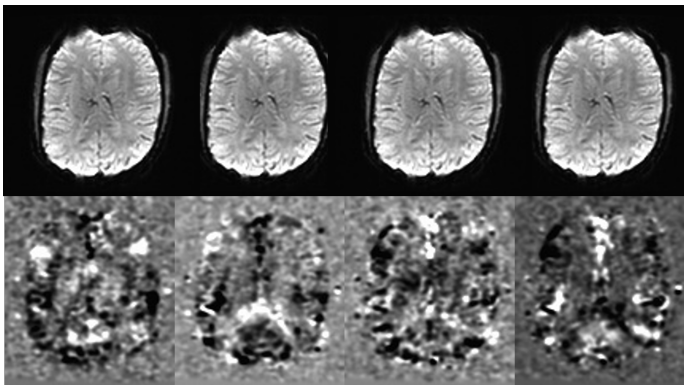
## 6.1 Anatomical segmentation and functional preprocessing

Individualized partioning of gray matter regions were obtained from T1 weighted images using FreeSurfer (Fig. 6.1 panel (a)).

Functional preprocessing was conducted in AFNI. The motion limit (for censoring TRs) was set rather strictly to 0.2, as motion artifacts are more problematic in resting state than task fMRI (0.3 is conventional in the task paradigm). However, two subjects moved substantially more than the others, to the extent that the regression model initially ended up with negative degrees of freedom (more regressors than data points), a meaningless situation. In order to not having to discard two observations from the already meager data set, the motion limit was expanded to 0.3 and 0.35 for the two subjects. While this could in principle introduce bias to the data, the subjects belonged to different clinical groups. The effect of functional preprocessing is illustrated in Fig.



**(a)** Structural preprocessing in FreeSurfer



**(b)** Functional preprocessing in AFNI

**Fig. 6.1:** Anatomical and functional preprocessing using FreeSurfer and AFNI. **a)** T1 weighted images (top row) were partitioned into anatomical regions based on a standard atlas (Destrieux), yielding personalized maps to use as network nodes (bottom row). **b)** EPI data before and after functional preprocessing. Four time points from the original BOLD timeseries (top row) and their corresponding timepoints after preprocessing in AFNI (bottom row). The two image modalities is later aligned in order to extract the BOLD time courses, by using the anatomical regions as network nodes.

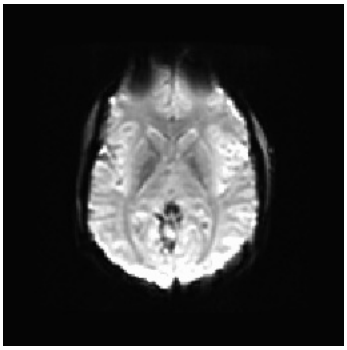
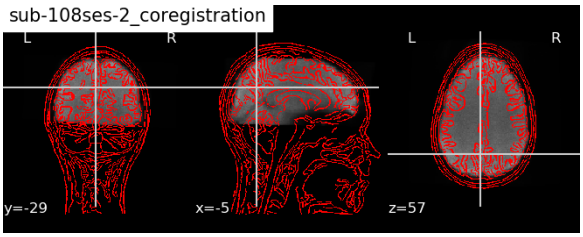


**Fig. 6.2:** Dropout and warping artifacts in the inferior frontal lobe can be seen in the EPI data from an axial slice. These artifacts arises due to interference between the air-tissue interface near the sinuses, and may cause trouble in terms of coregistration with an anatomical image. The effects can be partly overcome by unwarping the image using a field map.

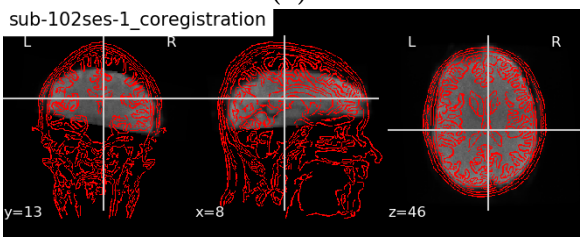
6.1 (panel b) on a contiguous series of TRs from an EPI data set, highlighting small fluctuations in the BOLD time series not visible in the raw data.

### 6.1.1 Coregistration

The EPI volumes were subject to dropout (signal loss) and distortion in frontal regions due to the tissue/air interface close to the sinuses (Fig. 6.2, upper center). No field maps were obtained during the scans, so the EPI data remained slightly warped. This introduced major difficulties during the anatomical-functional coregistration in approximately half of the 60 scans (see Fig. 6.3). By substituting AFNI's default alignment cost function ('hel', Hellinger metric), the improperly aligned runs were screened for a number of alternative cost functions. This was achieved using the AFNI program `align_epi_anat.py`, with the '-multi_cost' option and 'giant_move' option, to expand the search space. The alignment quality was inspected visually, and the data was realigned using the optimal cost function for each individual run. The cost function 'lpc+ZZ' (weighted sum of cost functions, but primarily the local Pearson correlation) was consistently superior to the other cost functions in the vast majority of subjects, with a few exceptions where it was outperformed by mutual information ('mi'). The final alignments were inspected and confirmed to be of acceptable quality.




(a)




(b)

**Fig. 6.3:** Coregistration of functional and anatomical data. Example of a successful (a) and failed (b) attempt of coregistration. Both alignments were produced using AFNI's default cost function ('hel'). Visualized using nilearn.

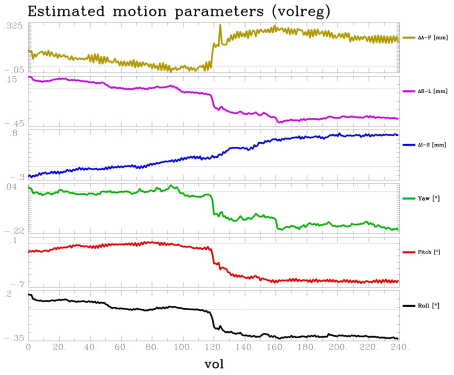
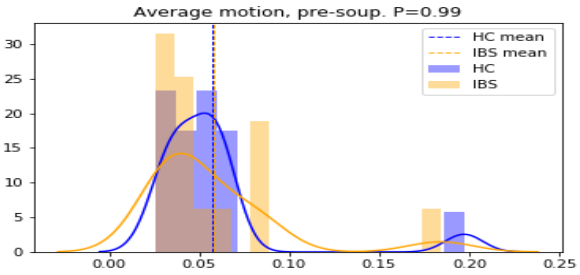



**Fig. 6.4:** Head motion throughout a single scan, described by three translations (yaw, pitch and roll) and three rotations (A-P, R-L and I-S)). These values and their first-order derivatives were used as regressors in the linear model during preprocessing. A-P, anterior-posterior; R-L, right-left; I-S, inferior-superior. Estimated and visualized using AFNI.

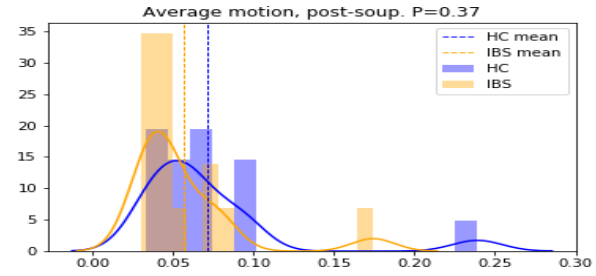
## 6.2 Controlling for motion differences

Head movement is arguably the primary confound in functional connectivity analysis. Head motion during an fMRI scan can be described with six motion parameters: three translations and three rotations as shown in Fig. 6.4.

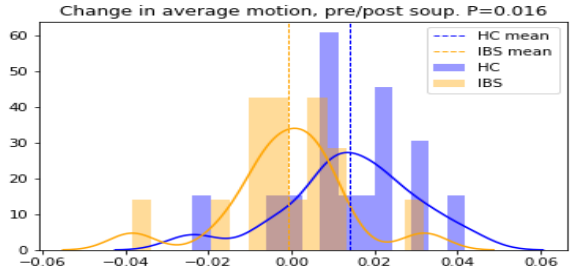
In order to exclude head motion as a possible explanatory variable for any group differences in connectivity, the average motion was compared across both clinical groups, in a similar fashion as [Pariyadath et al., 2014]. To simplify analysis, the average motion across all six motion parameters were used. Particularly considering many IBS patients experience nausea after food intake, it is reasonable to suspect such a difference. However, a two-sample t-test excluded this suspicion (see Fig. 6.5).




**(a)** Pre-intervention




**(b)** Post-intervention




**(c)** inter-session change

**Fig. 6.5:** Distribution of head motion in IBS patients and controls. Six motion parameters from each scan was averaged to represent total subject motion, both before **(a)**, after **(b)** and between **(c)** soup intervention (yellow=IBS, blue=HC). No significant group differences were found for either session (before, P=0.99; after, P=0.37; two-sample t-test). However, the change in subject motion from before to after intervention **(c)** demonstrated a slight increase in the control group (P=0.016, one-sample t-test). The patient group did not differ between the two sessions (P=1.0, one-sample t-test). Due to the confounding effect of motion on functional connectivity measures, equal distributions of motion is a prerequisite for making inference about true neurological differences with high confidence. Any observed group differences are very unlikely to be motion-induced artifacts based on these results. However, any inference based on inter-session change on the other hand, should be interpreted with care.

The overall motion of IBS patients and HCs were essentially the same, both pre- and post-intervention (P=0.99, P=0.37 respectively, 0.64 pooled, one-tailed t-test), with a slight tendency for the HC group moving more. Although no significant differences were observed in total motion, the change in motion from pre- to post-intervention suggested a mild group difference, in favor of the HC increasing (P=0.016). The IBS group remained essentially unchanged after the intervention. This implies that although the overall group motion distributions were more or less equal, any measure obtained by comparing pre- and post-intervention statistics should be interpreted with caution. In summary, there could not be established a group level difference in head motion between IBS patients and controls. Moreover, the data set is balanced both in terms of age and sex. Thus, any difference observed between the functional connectomes of IBS patients and HCs are unlikely to be attributed to differences of motion artefacts, but rather as being due to biologically considerable effects.

## 6.3 Network modelling

Average voxel time courses were extracted using FreeSurfer regions as masks to represent a single node time series, visualized for three Destrioeux regions in Fig. 6.6).

The adjacency matrices were thresholded prior to graph analysis. A proportional thresholding scheme was used, forcing the graphs to a fixed density by discarding the (1-p)×100% weakest entries in the adjacency matrix. As no globally optimal threshold exists, multiple values of p were set in the range from 0.1 to 0.9, as is common practice. Correlation, partial correlation and sparse inverse covariance (precision) were used in the analysis. Fig. 6.7 illustrates the three different connectivity measures before and after being thresholded. Because of the transitivity of Pearson correlation, this graph was impacted much more than the other two by a 20% threshold. The precision based network was initially so sparse that a 20% threshold had no effect. The final products were weighted undirected graphs, with both positive and negative weights preserved.

## 6.4 Graph metrics

Twelve graph metrics (nodewise and global) were computed from the thresholded graphs. No notable differences were observed for any single global graph metric between the

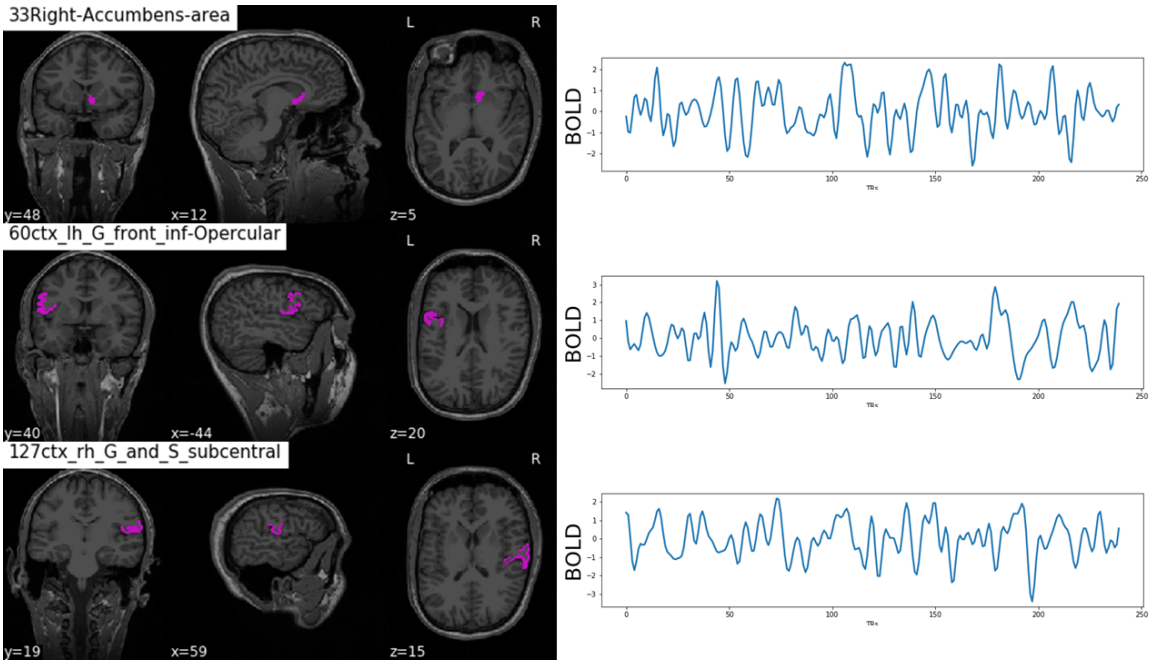
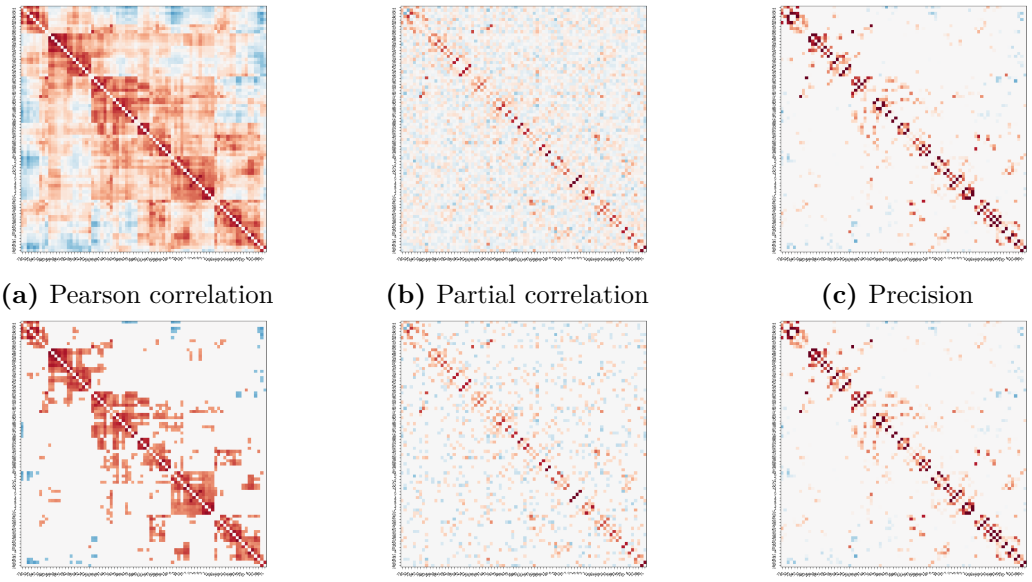



**Fig. 6.6:** BOLD time courses extracted from FreeSurfer regions. By aligning T1 weighted volumes with preprocessed 4D BOLD data, the individual FreeSurfer segmentations were used as masks from which the BOLD signal is extracted (represented by the average voxel signal). Adjacency matrices were built from correlating all pairwise time series to estimate the functional connectomes. The final product is a set of time series which can be correlated. Functional connectivity was modelled from the statistical coactivation of these time courses. Visualized using nilearn.

IBS group and controls. The joint pair-wise metrics were largely overlapping between the groups, not surprising considering the heterogeneity of the subjects. As seen in Fig. 6.8, most graph metrics were moderately to highly correlated with at least one other metric (Fig. 6.8). The correlations were in generally higher for lenient thresholds (results not shown).

## 6.5 Machine learning and classification

The primary goal of this work was to make a predictive model of IBS based on individual connectivity profiles. The machine learning classifiers were trained on all subjects but one, leaving out the remaining for prediction (LOGOCV), presenting the average prediction accuracy across all permutations as the final performance. Because of the small data set and large number of features, simple (linear) classifiers with regularization were emphasized, although a few non-linear also were investigated.



**(a)** Pearson correlation **(b)** Partial correlation **(c)** Precision

**(d)** Pearson correlation **(e)** Partial correlation **(f)** Precision

**Fig. 6.7:** Adjacency matrices with and without thresholding. Pearson correlation, partial correlation and precision matrices from a single subject (before intervention, Desikan-Killiany). A 20% proportional threshold was applied (bottom row), affecting the matrices to a varying degree. Visualized using nilearn.

### 6.5.1 Global features

Twelve global graph metrics were used as input to train six machine learning classifiers with the purpose of predicting the diagnostic status from graph metric biomarkers. By training each algorithm on all but one subject, it was evaluated by its success at predicting the held-out individual (LOGOCV). The average accuracy across all 30 folds served as the index of performance. A representative selection of thresholds (0.3, 0.5 and 0.8) was applied to the matrices prior to calculating the metrics, from which the average score is reported.

Fig. 6.9 displays the average performance from each classifier, based on graph metrics from Pearson correlation, precision and partial correlation matrices, averaged over the thresholds. The overall performance from Pearson correlation-based metrics fluctuated around what would be expected by random guessing in a binary classificaiton task (50%). However, precision and partial correlation deviated considerably from chance. Interestingly however, the prediction accuracy was consistently *below* chance (Fig. 6.9). Considering the inherent symmetry of a binary classification task, this paradoxical situation does indeed support a multivariate separability of IBS patients from HCs based on global graph metrics (simply reverse the prediction). that diagnostic

status is separable from IBS and HCs. This phenomenon, known as 'anti-learning' [Kowalczyk and Chapelle, 2005] is discussed more below.

Similarly, patient reports of experienced nausea and visceral pain were used as targets for the classifiers. The average LOGOCV accuracies are shown in Fig. 6.10 and Fig. 6.11. Again, the anti-learning phenomenon was encountered, for both pain and nausea. The magnitude of anti-learning was greatest for nausea.

Slight imbalances in the numbers of representatives in each class were imposed when stratifying the observations into low and high pain/nausea. [1] However, this was adjusted for visually by finding the expected accuracy (under the strategy of simply predicting the majority class) for the classes (labeled C1 and C2):

$$\max\left(\frac{N_{C1}}{N_{C1}+N_{C2}}, \frac{N_{C2}}{N_{C1}+N_{C2}}\right)$$

The expected accuracy was plotted, as a benchmark performance.

Prediction of pain and nausea was similarly paradoxically below chance. Nausea in particular deviated greatly, with the most extreme value at 28% (kernel SVC, partial correlation, the average of three thresholds). The phenomenon was initially suspected to be an artifact of the way the model evaluation. LOGOCV by design introduces a slight bias in the training data, by offsetting the number of observations belonging to each group: at each permutation, the training set will have 28 observations of one class and 30 of the other. This possibility was investigated by changing the CV strategy such that the groups always were balanced in the training set (group K-fold CV). This had effectively no impact on the accuracy, which still exhibited the peculiar below chance accuracy.

### 6.5.2 Nodewise metrics

Nodewise graph metrics provide more detailed information than their averaged global counterparts, but require further processing to account for the $p \gg N$ situation. RFE

[1] The binarization of pain and nausea scores was based on the median value, which in theory would yield two equal classes. However, it led to slight unevenness in the number of observations of each class due to a large number of patients reporting the same level of pain - it would be nonsensical to regard some 0's as 'low pain' and others as 'high pain'. Furthermore, the removal of NaN-values from precision-based adjacency matrices led to a different class imbalance than for Pearson and partial correlation, and is not directly comparable to the other two.

was used in conjunction with linear SVM, random forest and logistic regression in a LOGOCV framework. The most discriminative features, determined by the RFE were used to train the classifier, which was tested on the held-out subject. The problem of anti-learning ceased with this strategy, but in return most tests fluctuated around 50% accuracy. Fig. 6.13 displays the mean LOGOCV score from each classifier on Pearson and partial correlation and precision (average of three thresholds). The global average across all tests was just above 50% for SVC and logistic regression, the latter being marginally better. However, random forest yielded a slight below-chance accuracy of about the same magnitude. Pearson correlation and partial correlation performed better overall than precision using nodewise graph metrics. The nodewise features were not attempted for symptom scores. Classifiers based on precision generally performed worse than Pearson and partial correlation. The single best performing model had an average LOGOCV score of 80% (partial correlation, see 6.13 panel (b)) and the worst performing model around 30% (Pearson correlation, RFE-RF).

#### Recursive feature elimination

The three most informative features, decided by RFE-SVM are plotted in Fig. 6.12. It should be noted that the randomness in the computation produced slightly different results upon multiple trials, but computing the top 20 features revealed a large and consistent overlap between different trials. Different connectivity measures and thresholds were also largely selecting the same features. Importantly, almost every single of the top 20 features were betweenness centrality (about 25% would be expected by chance if all metrics were equally (un)informative).

## 6.6 Network based statistic

The statistical significance of the machine learning classifiers was not explored. However, to settle the question, another alleyway was attempted through NBS. This network-specific permutation test was performed on correlation matrices from the Desikan-Killiany atlas. A range of t-value thresholds between one and three were tested as recommended in [Fornito et al., 2016].

With a multiple t-value thresholds between 1 and 3, a NBS permutation was carried out in `bctpy`, with K=1000 permutations. The results are summarized in Fig. 6.14.

Small t-values yielded suitably small p-values (0.001). Selecting a threshold at 3.0 yielded a just-above significant p-value. The threshold was lowered slightly to 2.8 to find the smallest pseudonetwork that would still give a reasonably small p-value. This resulted in a p-value of 0.008, with observed 30 nodes from the Desikan-Killiany atlas, provided in Appendix A.


**Fig. 6.8:** Pairwise joint distribution of common global graph metrics, from sparse inverse covariance. The univariate distribution of any particular metric (diagonal) did not imply a substantial difference between IBS patients (red) and controls (blue). The bivariate distributions (dot plots) display a great overlap between the diagnostic grops. Notice the high degree of correlation between a number of metrics. Similar patterns was observed for partial and full correlation. Graph metrics were calculated from precision matrices using the Desikan-Killiany atlas.

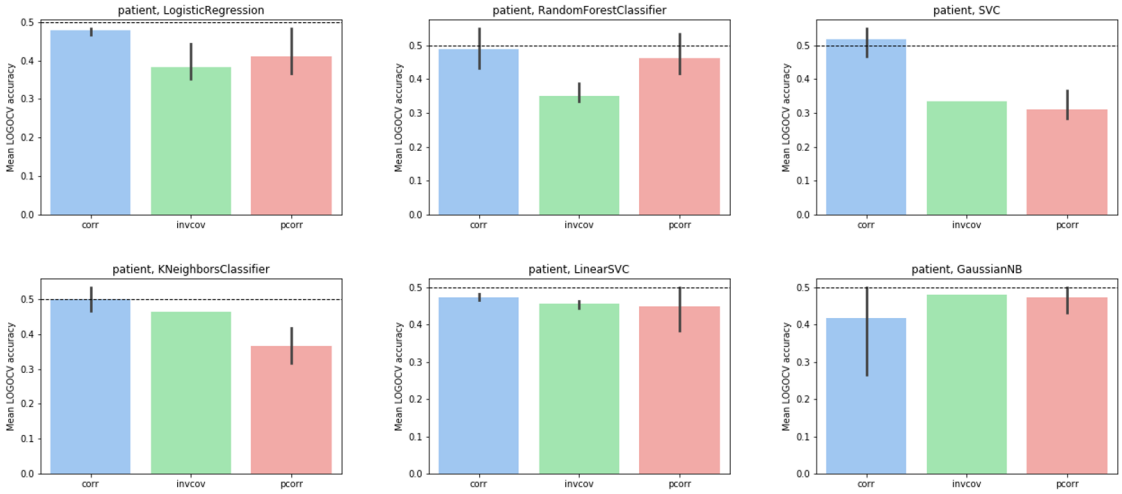



**Fig. 6.9:** Average LOGOCV score on global graph metrics, IBS vs HC. The black dotted line is the expected accuracy by random guessing, corrected for by any imbalance in the classes ($\frac{N_{c1}}{N_{c1}+N_{c2}}$). The bar heights were computed as the mean prediction accuracy from LOGOCV (30 folds) of three threholds (0.3 ,0.5, 0.8) from correlation (blue), precision (green) and partial correlation (red). Uncertainty is indicated by error bars. All classifiers performed consistently below chance for most connectivity measures. Top, from left: logistic regression (with slight L1 regularization), random forest classifier, support vector classifier (radial basis function kernel); bottom, from left: K-nearest neighbors, linear support vector classifier, Gaussian naive Bayes.

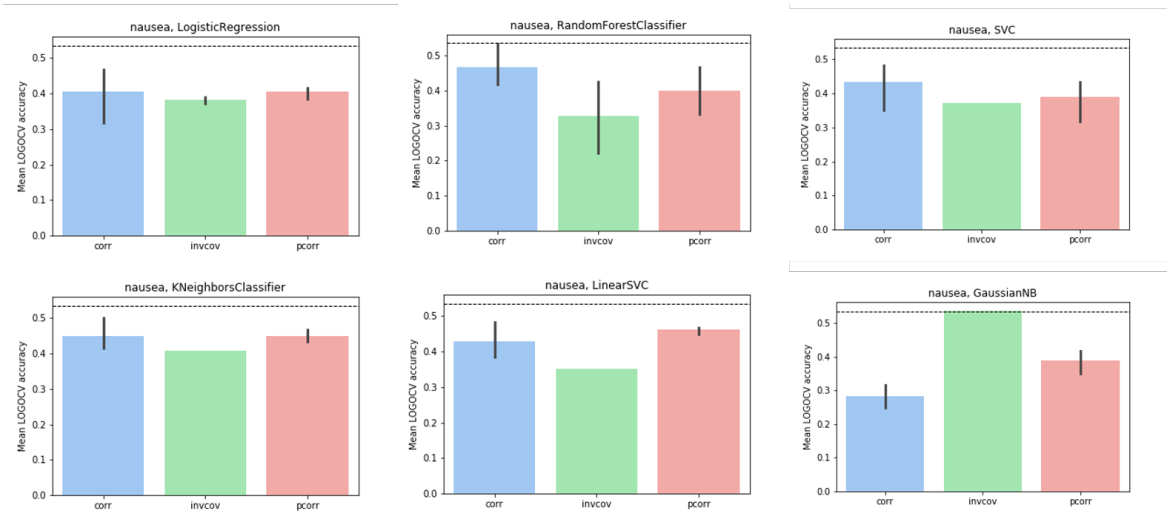



**Fig. 6.10:** Average LOGOCV score on global graph metrics, low vs high nausea. The black dotted line is the expected accuracy by random guessing, corrected for by any imbalance in the classes ($\frac{N_{c1}}{N_{c1}+N_{c2}}$). The bar heights were computed as the mean prediction accuracy from LOGOCV (30 folds) of three threholds (0.3 ,0.5, 0.8) from correlation (blue), precision (green) and partial correlation (red). Uncertainty is indicated by error bars. The accuracy from nausea deviated more from chance than pain and clinical status. Top, from left: logistic regression (with slight L1 regularization), random forest classifier, support vector classifier (radial basis function kernel); bottom, from left: K-nearest neighbors, linear support vector classifier, Gaussian naive Bayes.

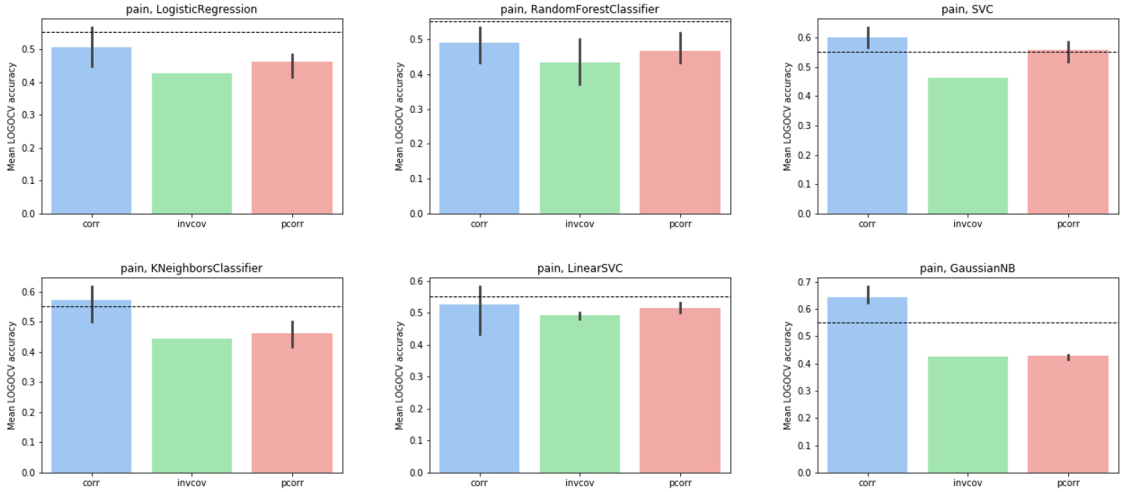
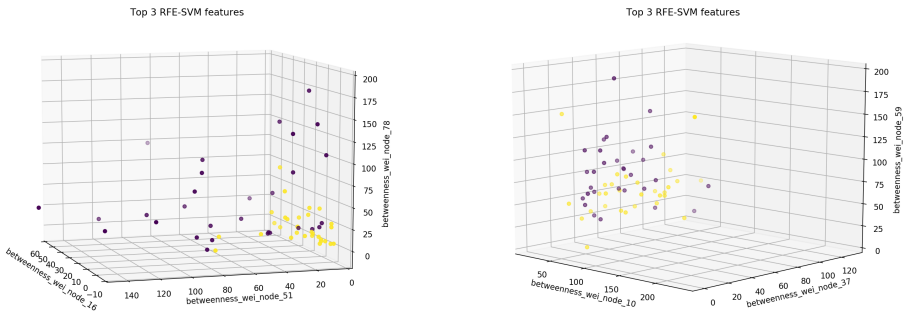
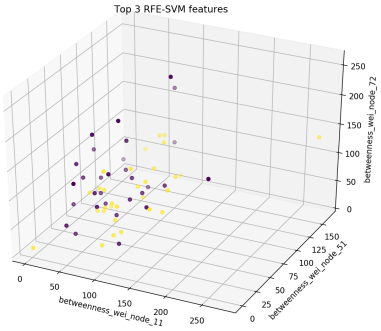



**Fig. 6.11:** Average LOGOCV score on global graph metrics, low vs high pain. The black dotted line is the expected accuracy by random guessing, corrected for by any imbalance in the classes ($\frac{N_{c1}}{N_{c1}+N_{c2}}$). The bar heights were computed as the mean prediction accuracy from LOGOCV (30 folds) of three threholds (0.3 ,0.5, 0.8) from correlation (blue), precision (green) and partial correlation (red). Uncertainty is indicated by error bars. The classifier performance was comparable to that obtained from clinical status. Top, from left: logistic regression (with slight L1 regularization), random forest classifier, support vector classifier (radial basis function kernel); bottom, from left: K-nearest neighbors, linear support vector classifier, Gaussian naive Bayes.

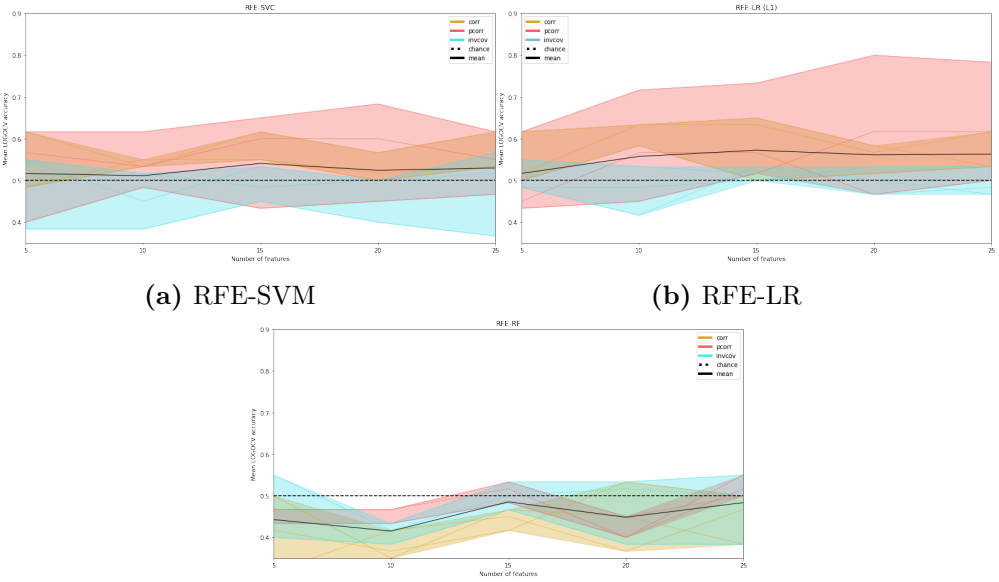



**(a)** Pearson correlation **(b)** Partial correlation




**(c)** Precision

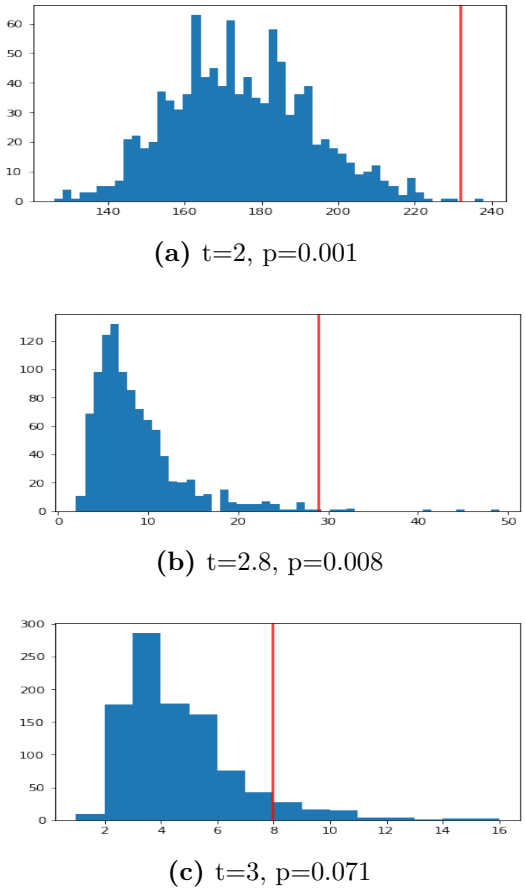
**Fig. 6.12:** RFE-SVM: the top three most informative features as determined by RFE-SVM from Pearson correlation **(a)**, partial correlation **(b)** and precision **(c)** matrices. HC (yellow) and IBS (purple). The clinical groups appear moderately separable, but with considerable overlap. Note that all features correspond to betweenness centrality. Threshold=0.5, nodewise graph metrics, Destrieux atlas.




**(a)** RFE-SVM **(b)** RFE-LR

**(c)** RFE-RF

**Fig. 6.13:** Accuracy from recursive feature elimination ranging from 5 to 20 features, using three different classifiers. The performance accuracy was assessed with leave-one-group-out cross validation on the complete pipeline from Pearson correlation (yellow), partial correlation (red) and precision (turquoise). Performance was largely equal between feature counts. Precision generally performed worse than the other two measures. SVC, support vector classifier; LR, logistic regression; RIDGE, linear regression with l1 regularization; RF, random forest. Hyperparameters were set to scikit-learn's default values. Three thresholds (0.3, 0.5, 0.8) were used per connectivity measure. The mean performance is plotted in black, barely surpassing chance guessing in (a) and (b). RFE-RF performed below chance.




**(a)** t=2, p=0.001

**(b)** t=2.8, p=0.008

**(c)** t=3, p=0.071

**Fig. 6.14:** Network based statistic histograms from 1000 permutations at three thresholds. The null distribution (in blue) and the observed statistic (in red). The x-axis represents the largest connected component (in terms of number of edges) of the pseudo-network which represents the edges remaining after being thresholded from a specified t-value. Taking advantage of the distributed effect of network aberrations, NBS holds an advantage to other statistical tests like the Benjamini-Hochberg procedure. high t-value thresholds indicate strong localized network alterations, while low values unravel more distributed effects. Three t-values thresholds revealed confident network differences between the IBS group and the HC group.

# Chapter 7

# Discussion

## 7.1 Graph metric reliability

Establishing robust biomarkers is important in order to reliably predict vulnerable individuals, prognosis and treatment response [Waller et al., 2017]. This requires thorough analysis of reliability under different circumstances, which can be used to guide practitioners. To this end, different graph metrics were ranked according to reproducibility on a simulated test-retest study, and evaluated in terms of dynamic stability (time to converge).

### 7.1.1 On the relative reliability of graph metrics

The reliability rank, as determined by ICC(3,1) on simulated test-retest data, was largely concordant between different ICA dimensionality levels (15, 25, 50 and 100) (see Fig. 5.2). Betweenness centrality, small-worldness and clustering coefficient were ranked as least reliable, while the modularity metric Q, eigenvector centrality and transitivity scored the highest. An interesting pattern that emerged was how clearly reliability increased with more granular brain parcellations. This was observed both in the simulated test-retest experiments on HCP1200 data and from clinical IBS data, strongly favouring high node parcellation schemes, wether functionally defined (ICA) or anatomically defined (FreeSurfer).

### Graph thresholding

Two distinct patterns were observed across all metrics: unsurprisingly, a higher noise level weakened reliability, but less anticipated was the tendency for stricter thresholds to reduce the ICC (Fig. 5.1). In isolation, this would advocate keeping thresholding at a minimum, or quitting the practice altogether. The observation that low graph density led to weaker reliability coincides with an earlier non-simulation study Braun et al. [2012]. Yet, Braun et al. still advocates using low density graphs (stricter thresholds) because they "preserve network-specific information", but this will necessarily have a negative impact on reproducibility, which is ill-advised for developing robust biomarkers. Another observation was that lenient thresholds were associated with stronger correlations between metrics. This may be problematic for machine learning classifiers, due to multicollinearity, and would support using stricter thresholds, contradicting the above conclusion. However, this effect needs to be weighted against the negative effect noisy features. It should be noted that [Telesford et al., 2010], who used another thresholding scheme did not observe a systematic difference in reliability between thresholds, suggesting it is inherent to proportional thresholding only.

### Agreement with other studies

Small-worldness is arguably the most studied graph theoretical measure, being thoroughly investigated in the early days of brain connectomics [Bassett and Bullmore, 2006; Sporns et al., 2002]. Although the results presented here deemed small-worldness as a poor biomarker, this is only in terms of rs-fMRI. Moreover, it may still not even be generalizable to anything but the specific circumstances in this work. In fact, [Braun et al., 2012] concluded small-worldness to be a highly reliable metric. Braun et al., however used anatomical automatic labeling (AAL), while this study relied primarily on ICA. Other potential explanations include different preprocessing practices, or the assumptions behind the simulation-study. Nonetheless, a systematic review on graph metric reliability also reported small-worldness to be moderately to highly robust, although there were some discrepancies between studies [Welton et al., 2015]. This suggests the simulation was too simplistic. However, the other ranks corresponded moderately well with the present results, with betweenness centrality and clustering coefficient as the least reliable metrics, placing some trust in the simulation approach. The low score of betweenness centrality could still be partially influenced

by the method of adding noise. Betweenness centrality favors nodes which connect topologically distant regions. This property will be substantially altered when adding weights with a uniform probability across all edges, without considering anatomy or physiology. By weighting the probability across edges (i.e. with multiplicative noise) this could potentially account for that.

### On the characteristic path length

A technical problem frequently arose when calculating characteristic path length (the average distance, $l_{ij}$). The distance between any two nodes in a network can be considered the path of 'least resistance'. In fragmented networks, when no such route exists, the path length is undefined (or may be set to infinity). This introduces trouble for calculating the characteristic path length, with no guidelines as how to treat infinite entries. Some solve this problem by avoiding thresholds that lead to fragmenting the network. Another related metric, the global efficiency, can be an alternative option. This metric deals smoothly with fragmented networks, by taking the average of the reciprocal of the distance, $\frac{1}{l_{ij}}$, which is simply set to zero if $l_{ij} = \infty$.

Both characteristic path length and global efficiency were used in the analysis. In the case of fragmented networks, I used the LCC for the calculation of characteristic path length. This was done primarily to avoid NaN-entries in the calculations. In fact, it is reasonable to speculate that this approach artificially inflated the characteristic path length, for the following reason. The LCC corresponds to the strongest edges of the network. By definition, this 'core' will exhibit a shorter path length, as path length is negatively related to edge strength. Another aspect to consider is the sudden point at which a network fragments. When adding noise before thresholding in a test-retest situation, it is conceivable that some iterations disconnect the network substantially, while others leave it in tact. This can be argued to induce much variability in the measures, or in ICC terms: increase the within-subject variability. Accordingly, the characteristic path length was slightly less robust to noise than the global efficiency, ranking $6^{th}$ and $8^{th}$ in the consensus order, respectively. A similar observation about the effect of network fragmentation was suggested by [Braun et al., 2012] to explain why low density graphs on average performed worse than higher density graphs.

### Limitations to the test-retest experiment

It must be noted that the absolute values of the ICC reported here were inherently meaningless, given that the noise was simulated (and arbitrarily so). Even more, assuming the noise to be Gaussian was neither theoretically nor experimentally justified or based on the actual propagation of error through the processing pipeline, from noise during acquisition to noise in the processed connectome product. Gaussian noise was chosen only due to simplicity. This could be criticized for its lack biological realism, but as a proof of concept, it can be shown that Gaussian noise added directly to the BOLD time series propagates to the correlation matrices. This proposition was simulated in Python, and confirmed that the resulting edge weight distribution turned out to closely follow a Gaussian distribution. This does however not take into consideration the regional variation. On the counter side, the results do indicate the relative robustness of each graph metric, in addition to revealing interesting patterns related to network dimensionality and thresholding stringency - with important implications for processing practice.

### Expanding window analysis

The temporal dependence of graph metrics was looked into through the lens of expanding window analysis. How metrics changed across 20 TR increments was used as an indicator of convergence. The overall results were quite similar for both atlases, keeping everything else fixed. However, there was a mild but clear tendency for metrics estimated from the Destrieux atlas to be more stable across time. This aligns well with the above finding that higher ICA dimensionalitites improved test-retest reliability. Although no quantitative criteria was set to rank the temporal stability, it was noted that some metrics changed more than others (by analyzing which graph metrics tended to change more than a full standard deviation during the last increments). Small-worldness was consistently more time sensitive than the other measures. From these results it would be generally be advised to not keep scan length below eight minutes.

Another observation was again that low graph densities yielded less stable metrics in time, further supporting the practice of using dense graphs. By agreement with the simulation experiment, this further supports that simulation lends itself to generalizion to observational test-retest data. In summary, the main results seem to be consistent across studies, like the advantage of using fine-grained parcellations [Welton et al.,

2015]. The finding that density also systematically affected reproducibility was not considered in the meta analysis by Welton et al., but was replicated by [Braun et al., 2012].

### 7.1.2 Machine learning results

Machine learning was applied to graph metrics to discriminate IBS patients from controls as well as symptom intensity scores. Anti-learning (consistent worse than chance classification) was encountered when using global graph metrics, with accuracy fluctuating around 40% for the prediction of nausea (nausea being the most distinguished target). Nodewise graph metrics in combination with recursive feature elimination was also attempted. The performance however was very poor: two out of three classifiers performed only marginally above chance (overall accuracy around 55%), and the last performed below chance of equal magnitude. Looking aside from the anti-learning paradox, there was a stronger relationship between graph metrics and symptom scores (nausea) than with diagnostic status. One limitation however is that the symptom based classes were not tested for differences in head motion.

#### Anti-learning

Anti-learning was observed for all six classifiers trained on global graph metrics from Pearson correlation, partial correlation and precision, suggesting it was inherent to the data set rather than unstable behaviour of any particular classifier. The results were consistent across different graph thresholds. The inherent symmetry of a binary classification task confirms the existence a difference in the connectomes of IBS patients and controls, as well as between high and low symptom scores. However, for an unknown reason, the classifiers consistently predicted the wrong label. It should be stressed that this peculiar behaviour is *not* a special case of overfitting [Kowalczyk and Chapelle, 2005]. The problem of anti-learning generally comes up in cases with very high dimensions and small sample sizes ($p \gg N$), even though the current feature vectors were smaller than the number of observations.

The consistent below chance performance was initially suspected to be an artifact of the LOGOCV design, which introduces slight imbalances in the classes when removing a subject for validation. This possibility was debunked however by changing the CV strategy to account for class imbalance (assuring equal representation of IBS and

HC in the training and validation data). This did not, however, change the puzzling results substantially, demonstrating the anti-learning phenomenon was of another origin. While not necessarily always the cause of anti-learning, the phenomenon can be evoked when using a linear (or sometimes non-linear) classifier on the exclusive-or (XOR) problem [Kowalczyk, 2007; Kowalczyk and Chapelle, 2005]. Although not attempted in the current work, anti-learning can and has been exploited in the past to achieve good performance in a clinical application [Roadknight et al., 2012]. This is promising for future work on machine learning classification applied to IBS functional connectivity.

### Nodewise metrics

Expecting the whole brain to display a pronounced overall altered connectivity is optimistic for comparably mild disorders like IBS. This would of course entail that the alteration is expressed on a global level of the brain, contrary to only within a particular subnetwork. It was therefore surprising to see a weaker discriminatory ability for nodewise metrics. This does however by no means indicate that the maximum possible performance was achieved; rather it indicates that more work should be done to refine feature selection and algorithmic decisions. In fact, the nodewise graph metrics contain all the same information as the global metrics.

Recursive feature elimination was used along with SVM, logistic regression and random forest, both as an exploration of machine learning prediction, as well as an exploratory effort of identifying a subset of brain regions implicated in IBS. Determining the 20 strongest nodewise features revealed an interesting pattern: nearly every feature corresponded to betweenness centrality. Notably, betweenness centrality was found to be the least reproducible metric among the twelve tested with test-retest simulation (Fig. 5.2), and has been reproduced in many studies [Welton et al., 2015]. Assuming that the global metric test-retest reliability (ICC) is reflective of the nodewise reliability, it follows that nodewise betweenness centrality is highly variable between measurements. Considering that each node of the network (154 (Destrieux) or 80 (Desikan-Killiany) nodes) has an associated betweenness centrality, this opens for a large body of noisy features, some of which are likely to correlate abritrarily with the target classes. This is in fact quite likely to have been detrimental to the performance of the classifiers. Therefore an idea for future work could be to remove the betweenness centrality altogether.

### 7.1.3 Network based statistic

It was established with fairly high confidence that there exists a group difference in the functional connectomes of IBS patients and HCs (p=0.001). However, the interpretation of this result is a little tricky. First of all, NBS works on a network level; it does by no means localize the aberration in question. Although there might exist a statistically strong difference on the network level, it doesn't evaluate to a strong statistical difference on any particular edge [Fornito et al., 2016]. This severely limits the application of such results in any machine learning context, although they inspire further investigation.

The size of the pseudonetwork (i.e. the number of edges in its LCC) is contingent upon the choice of t-value threshold. It is important to stress that although we are free in the choice of t-value threshold, this does not affect the statistical validity of final p-value provided from NBS [Fornito et al., 2016]. For example, selecting a t-value threshold that would not be considered statistically significant by itself for any particular edge, may still yield a significant p-value by means of NBS at the *network* level. The LCC was identified from a t-value threshold of 2.8, revealing about 29 regions of the Desikan-Killiany atlas. An interesting, but speculative, follow-up experiment could be to use the NBS as a form of feature selection: doing inference only on the nodes/edges corresponding to the observed pseudo network. Unfortunately, this would easily fall prey to circular reasoning (also known as data leakage): choosing a set of features based on their correlation with the classes (IBS and HC), and then resume to use the very same features to predict the class [Kriegeskorte et al., 2009].

#### Agreement with the literature

Previous studies of IBS have implicated regions involved in salience processing and emotional arousal, with particular emphasis on the insula and anterior cingulate cortex (although a number of other regions have been reported) [Mayer et al., 2015]. A similar study investigating brain morphometry in IBS highlighted the following regions: bilateral insula, bilateral amygdala, bilateral hippocampus, bilateral middle orbital frontal gyrus, left cingulate, left gyrus rectus, brainstem, and left putamen [Labus et al., 2014]. The regions of the pseudonetwork overlapped to a large extent with the aforementioned: right insula, bilateral amygdala, right hippocampus, left putamen, but not by much more than what is expected by chance. The gyrus rectus and orbitofrontal

gyrus were not identified by the NBS, but these two regions are also among the most strongly affected by warping artifacts (both are localized in the inferior frontal cortex). This may have potentially interfered with analysis, but it should be stressed that this is pure speculation. [Chen et al., 2011b] notably reported an increase in white matter integrity (determined by fractional anisotropy) specifically by the right insula in IBS patients.

### 7.1.4 What to expect from an IBS classifier

The machine learning approach to predict disease status from rs-fMRI data is most frequently applied to moderate to severe disorders like Alzheimer's disease, Parkinson's disease, temporal lobe epilepsy or schizophrenia [Brown and Hamarneh, 2016; Vieira et al., 2017]. Many studies report prediction accuracies up to 90% from brain imaging signatures (a comprehensive table of studies with their associated accuracy score is available from `http://connectomelearning.cs.sfu.ca/table.html`; pay particular attention to the "# Scans" tab). It should be noted that these disorders affect every aspect of the patient's life, and are as follows expected to show profound alterations in the brains of patients. Moreover, the vast majority of the studies use cohorts of 100 subjects or more. Needless to say, this benchmark is unrealistically high for an IBS classifier. Only a minority of studies on the brain signatures of IBS have used predictive modelling [Gupta et al., 2015]. However, one study from 2015 used brain morphometry to discriminate between IBS patients and controls, and reported an overall accuracy of 70%, with a comparable sensitivity and specificity [Labus et al., 2015]. However, the sample size was sevenfold greater than in the current work (N=212). The involvement of the brain in IBS symptoms is not well characterized, although its presence has been established. It should therefore in principle be possible to train a machine learning classifier to predict disease status using functional connectivity and graph metrics with a comparable or better performance.

## 7.2 Conclusive remarks

Psychiatric and functionally defined disorders are in a disadvantageous position by relying on subjective reports and questionnaires rather than physiological measures. Developing objective biomarkers for disorders like IBS are therefore of high priority for

aiding diagnosis, predict prognosis, subtypes or treatment response [Waller et al., 2017]. Machine learning can further work as a possible window into disease pathophysiology, particularly in conjunction with simultaneous measurements of gut microbiome or other aberrations implemented in IBS [Labus et al., 2019].

In conclusion, this thesis has explored network theory as a tool in the functional neuroimaging of IBS, with particular focus on network-based biomarkers in a machine learning framework. Furthermore, some methodological aspects of graph metric estimation from rs-fMRI have been investigated in detail, agreeing largely with other studies as well as reporting novel results. Evidence was provided for the possibility of discriminating patients from controls, despite in the form of anti-learning. Although the anti-learning paradox was not solved, it confirms that an equal but opposite performance score is in principle attainable [Kowalczyk and Chapelle, 2005; Roadknight et al., 2012], strongly encouraging further work on the matter.

# References


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

Ai, T., Morelli, J. N., Hu, X., Hao, D., Goerner, F. L., Ager, B., and Runge, V. M. (2012). A Historical Overview of Magnetic Resonance Imaging, Focusing on Technological Innovations:. *Investigative Radiology*, 47(12):725–741.

Amrhein, V., Greenland, S., and McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, 567(7748):305.

Bachmann, C., Jacobs, H. I. L., Porta Mana, P., Dillen, K., Richter, N., von Reutern, B., Dronse, J., Onur, O. A., Langen, K.-J., Fink, G. R., Kukolja, J., and Morrison, A. (2018). On the Extraction and Analysis of Graphs From Resting-State fMRI to Support a Correct and Robust Diagnostic Tool for Alzheimer's Disease. *Frontiers in Neuroscience*, 12.

Barabási, A.-L. and Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512.

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68.

Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101.

Barbara, G., Cremon, C., Carini, G., Bellacosa, L., Zecchi, L., De Giorgio, R., Corinaldesi, R., and Stanghellini, V. (2011). The Immune System in Irritable Bowel Syndrome. *Journal of Neurogastroenterology and Motility*, 17(4):349–359.

Barkhof, F., Haller, S., and Rombouts, S. A. R. B. (2014). Resting-State Functional MR Imaging: A New Window to the Brain. *Radiology*, 272(1):29–49.

Bartko, J. J. (1976). On various intraclass correlation reliability coefficients. *Psychological Bulletin*, 83(5):762–765.

Bassett, D. S. and Bullmore, E. (2006). Small-World Brain Networks. *The Neuroscientist*, 12(6):512–523.

Beattie, D. T. and Smith, J. A. M. (2008). Serotonin pharmacology in the gastrointestinal tract: a review. *Naunyn-Schmiedeberg's Archives of Pharmacology*, 377(3):181–203.

Beckmann, C. F. (2012). Modelling with independent components. *NeuroImage*, 62(2):891–901.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical And Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, 57(1):289–300.

Bennett, C., Miller, M., and Wolford, G. (2009). Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: an argument for multiple comparisons correction. *NeuroImage*, 47:S125.

Betzel, R. F. and Bassett, D. S. (2017). Multi-scale brain networks. *NeuroImage*, 160:73–83.

Bhatt, R. R., Gupta, A., Labus, J. S., Zeltzer, L. K., Tsao, J. C., Shulman, R. J., and Tillisch, K. (2019). Altered Brain Structure and Functional Connectivity and Its Relation to Pain Perception in Girls With Irritable Bowel Syndrome. *Psychosomatic Medicine*, 81(2):146–154.

Bianciardi, M., Fukunaga, M., van Gelderen, P., Horovitz, S. G., de Zwart, J. A., Shmueli, K., and Duyn, J. H. (2009). Sources of functional magnetic resonance imaging signal fluctuations in the human brain at rest: a 7 T study. *Magnetic Resonance Imaging*, 27(8):1019–1029.

Bihan, D. L., Mangin, J.-F., Poupon, C., Clark, C. A., Pappata, S., Molko, N., and Chabriat, H. (2001). Diffusion tensor imaging: Concepts and applications. *Journal of Magnetic Resonance Imaging*, 13(4):534–546.

Birn, R. M., Molloy, E. K., Patriat, R., Parker, T., Meier, T. B., Kirk, G. R., Nair, V. A., Meyerand, M. E., and Prabhakaran, V. (2013). The effect of scan length on the reliability of resting-state fMRI connectivity estimates. *NeuroImage*, 83:550–558.

Biswal, B., Zerrin Yetkin, F., Haughton, V. M., and Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic Resonance in Medicine*, 34(4):537–541.

Biswal, B. B. (2012). Resting state fMRI: A personal history. *NeuroImage*, 62(2):938–944.

Biswal, B. B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S. M., Beckmann, C. F., Adelstein, J. S., Buckner, R. L., Colcombe, S., Dogonowski, A.-M., Ernst, M., Fair, D., Hampson, M., Hoptman, M. J., Hyde, J. S., Kiviniemi, V. J., Kötter, R., Li, S.-J., Lin, C.-P., Lowe, M. J., Mackay, C., Madden, D. J., Madsen, K. H., Margulies, D. S., Mayberg, H. S., McMahon, K., Monk, C. S., Mostofsky, S. H., Nagel, B. J., Pekar, J. J., Peltier, S. J., Petersen, S. E., Riedl, V., Rombouts, S. A. R. B., Rypma, B., Schlaggar, B. L., Schmidt, S., Seidler, R. D., Siegle, G. J., Sorg, C., Teng, G.-J., Veijola, J., Villringer, A., Walter, M., Wang, L., Weng, X.-C., Whitfield-Gabrieli,

S., Williamson, P., Windischberger, C., Zang, Y.-F., Zhang, H.-Y., Castellanos, F. X., and Milham, M. P. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences of the United States of America*, 107(10):4734–4739.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D. (2006). Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308.

Bonaz, B. (2002). Central Processing of Rectal Pain in Patients With Irritable Bowel Syndrome: An fMRI Study. *THE AMERICAN JOURNAL OF GASTROENTEROLOGY*, 97(3):8.

Bonferroni, C. (1936). *Teoria Statistica Delle Classi e Calcolo Delle Probabilità*, volume 8. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze.

Braun, U., Plichta, M. M., Esslinger, C., Sauer, C., Haddad, L., Grimm, O., Mier, D., Mohnke, S., Heinz, A., Erk, S., Walter, H., Seiferth, N., Kirsch, P., and Meyer-Lindenberg, A. (2012). Test–retest reliability of resting-state connectivity network characteristics using fMRI and graph theoretical measures. *NeuroImage*, 59(2):1404–1412.

Brown, C. J. and Hamarneh, G. (2016). Machine Learning on Human Connectome Data from MRI. *arXiv:1611.08699 [cs, q-bio, stat]*. arXiv: 1611.08699.

Bullmore, E. and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198.

Bullmore, E. and Sporns, O. (2012). The economy of brain network organization. *Nature Reviews Neuroscience*, 13(5):336–349.

Canavan, C., West, J., and Card, T. (2014). The epidemiology of irritable bowel syndrome. *Clinical Epidemiology*, 6:71–80.

Chang, C. and Glover, G. H. (2009). Effects of model-based physiological noise correction on default mode network anti-correlations and correlations. *NeuroImage*, 47(4):1448–1459.

Chen, G., Chen, G., Xie, C., and Li, S.-J. (2011a). Negative Functional Connectivity and Its Dependence on the Shortest Path Length of Positive Network in the Resting-State Human Brain. *Brain Connectivity*, 1(3):195–206.

Chen, J. Y.-W., Blankstein, U., Diamant, N. E., and Davis, K. D. (2011b). White matter abnormalities in irritable bowel syndrome and relation to individual factors. *Brain Research*, 1392:121–131.

Clarke, D. D. and Sokoloff, L. (1999). Regulation of Cerebral Metabolic Rate. *Basic Neurochemistry: Molecular, Cellular and Medical Aspects. 6th edition.*

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

Craddock, R. C., Holtzheimer, P. E., Hu, X. P., and Mayberg, H. S. (2009). Disease state prediction from resting state functional connectivity. *Magnetic Resonance in Medicine*, 62(6):1619–1628.

Cremon, C., Gargano, L., Morselli-Labate, A. M., Santini, D., Cogliandro, R. F., De Giorgio, R., Stanghellini, V., Corinaldesi, R., and Barbara, G. (2009). Mucosal immune activation in irritable bowel syndrome: gender-dependence and association with digestive symptoms. *The American Journal of Gastroenterology*, 104(2):392–400.

Cryan, J. F. and Dinan, T. G. (2012). Mind-altering microorganisms: the impact of the gut microbiota on brain and behaviour. *Nature Reviews. Neuroscience*, 13(10):701–712.

De Vico Fallani, F., Richiardi, J., Chavez, M., and Achard, S. (2014). Graph analysis of functional brain networks: practical issues in translational neuroscience. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1653):20130521–20130521.

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

Destrieux, C., Fischl, B., Dale, A., and Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, 53(1):1–15.

Dosenbach, N. U. F., Nardos, B., Cohen, A. L., Fair, D. A., Power, J. D., Church, J. A., Nelson, S. M., Wig, G. S., Vogel, A. C., Lessov-Schlaggar, C. N., Barnes, K. A., Dubis, J. W., Feczko, E., Coalson, R. S., Pruett, J. R., Barch, D. M., Petersen, S. E., and Schlaggar, B. L. (2010). Prediction of Individual Brain Maturity Using fMRI. *Science*, 329(5997):1358–1361.

Drossman, D. A. (2006). The Functional Gastrointestinal Disorders and the Rome III Process. *Gastroenterology*, 130(5):1377–1390.

Drossman, D. A. (2016). Functional Gastrointestinal Disorders: History, Pathophysiology, Clinical Features, and Rome IV. *Gastroenterology*, 150(6):1262–1279.e2.

Ekman, M., Derrfuss, J., Tittgemeyer, M., and Fiebach, C. J. (2012). Predicting errors from reconfiguration patterns in human brain networks. *Proceedings of the National Academy of Sciences*, 109(41):16714–16719.

Elsenbruch, S. (2011). Abdominal pain in Irritable Bowel Syndrome: A review of putative psychological, neural and neuro-immune mechanisms. *Brain, Behavior, and Immunity*, 25(3):386–394.

Enck, P., Aziz, Q., Barbara, G., Farmer, A. D., Fukudo, S., Mayer, E. A., Niesler, B., Quigley, E. M. M., Rajilić-Stojanović, M., Schemann, M., Schwille-Kiuntke, J., Simren, M., Zipfel, S., and Spiller, R. C. (2016). Irritable bowel syndrome. *Nature Reviews Disease Primers*, 2:16014.

Euler, L. (1741). Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, 8:15.

Fletcher, J. M. and Wennekers, T. (2018). From Structure to Activity: Using Centrality Measures to Predict Neuronal Activity. *International Journal of Neural Systems*, 28(02):1750013.

Fornito, A., Zalesky, A., and Breakspear, M. (2015). The connectomics of brain disorders. *Nature Reviews Neuroscience*, 16(3):159–172.

Fornito, A., Zalesky, A., and Bullmore, E. T. (2016). *Fundamentals of brain network analysis*. Elsevier/Academic Press, Amsterdam ; Boston. OCLC: ocn943431396.

Fornito, A., Zalesky, A., Pantelis, C., and Bullmore, E. T. (2012). Schizophrenia, neuroimaging and connectomics. *NeuroImage*, 62(4):2296–2314.

Fox, M. D. and Raichle, M. E. (2007). Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature Reviews Neuroscience*, 8(9):700–711.

Fox, P. T. and Raichle, M. E. (1986). Focal physiological uncoupling of cerebral blood flow and oxidative metabolism during somatosensory stimulation in human subjects. *Proceedings of the National Academy of Sciences of the United States of America*, 83(4):1140–1144.

Friston, K. J. (1994). Functional and effective connectivity in neuroimaging: A synthesis. *Human Brain Mapping*, 2(1-2):56–78.

Friston, K. J., Frith, C. D., Liddle, P. F., and Frackowiak, R. S. J. (1993). Functional Connectivity: The Principal-Component Analysis of Large (PET) Data Sets. *Journal of Cerebral Blood Flow & Metabolism*, 13(1):5–14.

Georgescu, D., Reisz, D., Gurban, C. V., Georgescu, L., Ionita, I., Ancusa, O. E., and Lighezan, D. (2017). Migraine in young females with irritable bowel syndrome: still a challenge. *Neuropsychiatric Disease and Treatment*, Volume 14:21–28.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., and Jenkinson, M. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, 80:105–124.

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., Handwerker, D. A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B. N., Nichols, T. E., Pellman, J., Poline, J.-B., Rokem, A., Schaefer, G., Sochat, V., Triplett, W., Turner, J. A., Varoquaux, G., and Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, 3:160044.

Gralnek, I. M., Hays, R. D., Kilbourne, A., Naliboff, B., and Mayer, E. A. (2000). The impact of irritable bowel syndrome on health-related quality of life. *Gastroenterology*, 119(3):654–660.

Gupta, A., Rapkin, A. J., Gill, Z., Kilpatrick, L., Fling, C., Stains, J., Masghati, S., Tillisch, K., Mayer, E. A., and Labus, J. S. (2015). Disease-Related Differences in Resting State Networks: A Comparison between Localized Provoked Vulvodynia, Irritable Bowel Syndrome, and Healthy Control Subjects. *Pain*, 156(5):809–819.

Guyon, I. and Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3(Mar):1157–1182.

Heuvel, M. P. v. d. and Sporns, O. (2019). A cross-disorder connectome landscape of brain dysconnectivity. *Nature Reviews Neuroscience*, page 1.

Hillilä, M. T., Siivola, M. T., and Färkkilä, M. A. (2007). Comorbidity and use of health-care services among irritable bowel syndrome sufferers. *Scandinavian Journal of Gastroenterology*, 42(7):799–806.

Hillman, E. M. (2014). Coupling Mechanism and Significance of the BOLD Signal: A Status Report. *Annual review of neuroscience*, 37:161–181.

Holtmann, G. J., Ford, A. C., and Talley, N. J. (2016). Pathophysiology of irritable bowel syndrome. *The Lancet Gastroenterology & Hepatology*, 1(2):133–146.

Hosseini-Asl, E., Gimel'farb, G., and El-Baz, A. (2016). Alzheimer's Disease Diagnostics by a Deeply Supervised Adaptable 3d Convolutional Network. *arXiv:1607.00556 [cs, q-bio, stat]*. arXiv: 1607.00556.

Hoult, D. (2009). The origins and present status of the radio wave controversy in NMR. *Concepts in Magnetic Resonance Part A*, 34A(4):193–216.

Hoult, D. I. (1989). The magnetic resonance myth of radio waves: The Magnetic Resonance Myth of Radio Waves. *Concepts in Magnetic Resonance*, 1(1):1–5.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430.

Icenhour, A., Witt, S. T., Elsenbruch, S., Lowén, M., Engström, M., Tillisch, K., Mayer, E. A., and Walter, S. (2017). Brain functional connectivity is associated with visceral sensitivity in women with Irritable Bowel Syndrome. *NeuroImage: Clinical*, 15:449–457.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, 2(8).

Jang, H. B., Bolduc, B., Zablocki, O., Kuhn, J. H., Roux, S., Adriaenssens, E. M., Brister, J. R., Kropinski, A. M., Krupovic, M., Lavigne, R., Turner, D., and Sullivan, M. B. (2019). Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nature Biotechnology*, page 1.

jones, m. p., dilley, j. b., drossman, d., and crowell, m. d. (2006). Brain-gut connections in functional GI disorders: anatomic and physiologic relationships. *Neurogastroenterology and Motility*, 18(2):91–103.

Kanwisher, N., McDermott, J., and Chun, M. M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *Journal of Neuroscience*, 17(11):4302–4311.

Koo, T. K. and Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15(2):155–163.

Kowalczyk, A. (2007). Classification of Anti-learnable Biological and Synthetic Data. In Kok, J. N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., and Skowron, A., editors, *Knowledge Discovery in Databases: PKDD 2007*, Lecture Notes in Computer Science, pages 176–187. Springer Berlin Heidelberg.

Kowalczyk, A. and Chapelle, O. (2005). An Analysis of the Anti-learning Phenomenon for the Class Symmetric Polyhedron. In *Algorithmic Learning Theory*, volume 3734, pages 78–91. Springer Berlin Heidelberg, Berlin, Heidelberg.

Krebs, V. (2002). Uncloaking Terrorist Networks. *First Monday*, 7(4).

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., and Baker, C. I. (2009). Circular analysis in systems neuroscience – the dangers of double dipping. *Nature neuroscience*, 12(5):535–540.

Kubo, R. and Tomita, K. (1954). A General Theory of Magnetic Resonance Absorption. *Journal of the Physical Society of Japan*, 9(6):888–919.

Kwong, K. K., Belliveau, J. W., Chesler, D. A., Goldberg, I. E., Weisskoff, R. M., Poncelet, B. P., Kennedy, D. N., Hoppel, B. E., Cohen, M. S., and Turner, R. (1992). Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proceedings of the National Academy of Sciences of the United States of America*, 89(12):5675–5679.

Labus, J., Naliboff, B., Fallon, J., Berman, S., Suyenobu, B., Bueller, J., Mandelkern, M., and Mayer, E. (2008). Sex differences in brain activity during aversive visceral stimulation and its expectation in patients with chronic abdominal pain: A network analysis. *NeuroImage*, 41(3):1032–1043.

Labus, J. S., Dinov, I. D., Jiang, Z., Ashe-McNalley, C., Zamanyan, A., Shi, Y., Hong, J.-Y., Gupta, A., Tillisch, K., Ebrat, B., Hobel, S., Gutman, B. A., Joshi, S., Thompson, P. M., Toga, A. W., and Mayer, E. A. (2014). Irritable bowel syndrome in female patients is associated with alterations in structural brain networks:. *Pain*, 155(1):137–149.

Labus, J. S., Naliboff, B. D., Berman, S. M., Suyenobu, B., Vianna, E. P., Tillisch, K., and Mayer, E. A. (2009). Brain networks underlying perceptual habituation to repeated aversive visceral stimuli in patients with irritable bowel syndrome. *NeuroImage*, 47(3):952–960.

Labus, J. S., Osadchiy, V., Hsiao, E. Y., Tap, J., Derrien, M., Gupta, A., Tillisch, K., Le Nevé, B., Grinsvall, C., Ljungberg, M., Öhman, L., Törnblom, H., Simren, M., and Mayer, E. A. (2019). Evidence for an association of gut microbial Clostridia with

brain functional connectivity and gastrointestinal sensorimotor function in patients with irritable bowel syndrome, based on tripartite network analysis. *Microbiome*, 7.

Labus, J. S., Van Horn, J. D., Gupta, A., Alaverdyan, M., Torgerson, C., Ashe-McNalley, C., Irimia, A., Hong, J.-Y., Naliboff, B., Tillisch, K., and Mayer, E. A. (2015). Multivariate morphological brain signatures predict patients with chronic abdominal pain from healthy control subjects:. *PAIN*, 156(8):1545–1554.

Laufs, H., Krakow, K., Sterzer, P., Eger, E., Beyerle, A., Salek-Haddadi, A., and Kleinschmidt, A. (2003). Electroencephalographic signatures of attentional and cognitive default modes in spontaneous brain activity fluctuations at rest. *Proceedings of the National Academy of Sciences*, 100(19):11053–11058.

Lauterbur, P. C. (1973). Image Formation by Induced Local Interactions: Examples Employing Nuclear Magnetic Resonance. *Nature*, 242(5394):190–191.

Levy, R. L., Jones, K. R., Whitehead, W. E., Feld, S. I., Talley, N. J., and Corey, L. A. (2001). Irritable bowel syndrome in twins: Heredity and social learning both contribute to etiology. *Gastroenterology*, 121(4):799–804.

Li, L., Zeng, L., Lin, Z.-J., Cazzell, M., and Liu, H. (2015). Tutorial on use of intraclass correlation coefficients for assessing intertest reliability and its application in functional near-infrared spectroscopy–based brain imaging. *Journal of Biomedical Optics*, 20(5):050801.

Li, X., Morgan, P. S., Ashburner, J., Smith, J., and Rorden, C. (2016). The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *Journal of Neuroscience Methods*, 264:47–56.

Liang, S., Wu, X., and Jin, F. (2018). Gut-Brain Psychology: Rethinking Psychology From the Microbiota–Gut–Brain Axis. *Frontiers in Integrative Neuroscience*, 12.

Liu, Y., Liang, M., Zhou, Y., He, Y., Hao, Y., Song, M., Yu, C., Liu, H., Liu, Z., and Jiang, T. (2008). Disrupted small-world networks in schizophrenia. *Brain*, 131(4):945–961.

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412(6843):150–157.

Martin, C. R., Osadchiy, V., Kalani, A., and Mayer, E. A. (2018). The Brain-Gut-Microbiome Axis. *Cellular and Molecular Gastroenterology and Hepatology*, 6(2):133–148.

Mayer, E. A. (2018). The Role of Gut-Brain Interactions in Influencing Symptoms of Irritable Bowel Syndrome. *Gastroenterology & Hepatology*, 14(1):3.

Mayer, E. A., Labus, J. S., Tillisch, K., Cole, S. W., and Baldi, P. (2015). Towards a systems view of IBS. *Nature reviews. Gastroenterology & hepatology*, 12(10):592–605.

Mayer, E. A., Savidge, T., and Shulman, R. J. (2014). Brain–Gut Microbiome Interactions and Functional Bowel Disorders. *Gastroenterology*, 146(6):1500–1512.

McGraw, K. O. and Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1):30–46.

Milgram, S. (1967). The small-world problem: (400002009-005). Technical report, American Psychological Association. type: dataset.

Mohajeri, M. H., Brummer, R. J. M., Rastall, R. A., Weersma, R. K., Harmsen, H. J. M., Faas, M., and Eggersdorfer, M. (2018). The role of the microbiome for human health: from basic science to clinical applications. *European Journal of Nutrition*, 57(Suppl 1):1–14.

Müller, R. and Büttner, P. (1994). A critical discussion of intraclass correlation coefficients. *Statistics in Medicine*, 13(23-24):2465–2476.

Newman, M. E. J. (2002). Assortative Mixing in Networks. *Physical Review Letters*, 89(20):208701.

Ogawa, S., Lee, T. M., Kay, A. R., and Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences of the United States of America*, 87(24):9868–9872.

O'Mahony, S. M., Marchesi, J. R., Scully, P., Codling, C., Ceolho, A.-M., Quigley, E. M., Cryan, J. F., and Dinan, T. G. (2009). Early Life Stress Alters Behavior, Immunity, and Microbiota in Rats: Implications for Irritable Bowel Syndrome and Psychiatric Illnesses. *Biological Psychiatry*, 65(3):263–267.

Pagani, G. A. and Aiello, M. (2011). The Power Grid as a Complex Network: a Survey. *arXiv:1105.3338 [physics]*. arXiv: 1105.3338.

Palsson, O. S., Whitehead, W. E., van Tilburg, M. A. L., Chang, L., Chey, W., Crowell, M. D., Keefer, L., Lembo, A. J., Parkman, H. P., Rao, S. S., Sperber, A., Spiegel, B., Tack, J., Vanner, S., Walker, L. S., Whorwell, P., and Yang, Y. (2016). Rome IV Diagnostic Questionnaires and Tables for Investigators and Clinicians. *Gastroenterology*.

Pariyadath, V., Stein, E. A., and Ross, T. J. (2014). Machine learning classification of resting state functional connectivity predicts smoking status. *Frontiers in Human Neuroscience*, 8.

Park, H.-J. and Friston, K. (2013). Structural and Functional Brain Networks: From Connections to Cognition. *Science*, 342(6158):1238411–1238411.

Park, S. H., Videlock, E. J., Shih, W., Presson, A. P., Mayer, E. A., and Chang, L. (2016). Adverse childhood experiences are associated with irritable bowel syndrome and gastrointestinal symptom severity. *Neurogastroenterology & Motility*, 28(8):1252–1260.

Pastor-Satorras, R. and Vespignani, A. (2002). Immunization of complex networks. *Physical Review E*, 65(3):036104.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., and Cournapeau, D. (2011). Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*, page 6.

Pereira, F., Mitchell, T., and Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage*, 45(1):S199–S209.

Poldrack, R. A., Mumford, J. A., and Nichols, T. E. (2011). *Handbook of functional MRI data analysis*. Cambridge University Press, Cambridge New York Melbourne Madrid. OCLC: 753167009.

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage*, 59(3):2142–2154.

Przekop, P., Haviland, M. G., Zhao, Y., Oda, K., Morton, K. R., and Fraser, G. E. (2012). Self-reported physical health, mental health, and comorbid diseases among women with irritable bowel syndrome, fibromyalgia, or both compared with healthy control respondents. *The Journal of the American Osteopathic Association*, 112(11):726–735.

Rabi, I. I., Zacharias, J. R., Millman, S., and Kusch, P. (1938). A New Method of Measuring Nuclear Magnetic Moment. *Physical Review*, 53(4):318–318.

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

Raichle, M. E. and Mintun, M. A. (2006). Brain work and brain imaging. *Annual Review of Neuroscience*, 29:449–476.

Ramsey, N. F. (1999). Early History of Magnetic Resonance. *Physics in Perspective*, 1(2):123–135.

Reuter, M., Schmansky, N. J., Rosas, H. D., and Fischl, B. (2012). Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage*, 61(4):1402–1418.

Rigden, J. S. (1986). Quantum states and precession: The two discoveries of NMR. *Reviews of Modern Physics*, 58(2):433–448.

Rinck, P. A., Books on Demand, and European Magnetic Resonance Forum (2018). *Magnetic resonance in medicine: a critical introduction : the basic text book of the European Magnetic Resonance Forum*. Books on Demand, Norderstedt. OCLC: 1084559323.

Roadknight, C., Aickelin, U., Qiu, G., Scholefield, J., and Durrant, L. (2012). Supervised learning and anti-learning of colorectal cancer classes and survival rates from cellular biology parameters. In *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 797–802, Seoul, Korea (South). IEEE.

Rowe, J. (2010). Connectivity Analysis is Essential to Understand Neurological Disorders. *Frontiers in Systems Neuroscience*, 4.

Roy, C. S. and Sherrington, C. S. (1890). On the Regulation of the Blood-supply of the Brain. *The Journal of Physiology*, 11(1-2):85–158.17.

Rubinov, M. and Sporns, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, 52(3):1059–1069.

Saito, Y. A., Strege, P. R., Tester, D. J., Locke, G. R., Talley, N. J., Bernard, C. E., Rae, J. L., Makielski, J. C., Ackerman, M. J., and Farrugia, G. (2009). Sodium channel mutation in irritable bowel syndrome: evidence for an ion channelopathy. *American Journal of Physiology - Gastrointestinal and Liver Physiology*, 296(2):G211–G218.

Sara, S. J. and Bouret, S. (2012). Orienting and reorienting: the locus coeruleus mediates cognition through arousal. *Neuron*, 76(1):130–141.

Sattenspiel, L. and Simon, C. P. (1988). The spread and persistence of infectious diseases in structured populations. *Mathematical Biosciences*, 90(1-2):341–366.

Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., Reiss, A. L., and Greicius, M. D. (2007). Dissociable Intrinsic Connectivity Networks for Salience Processing and Executive Control. *Journal of Neuroscience*, 27(9):2349–2356.

Seminowicz, D. A., Labus, J. S., Bueller, J. A., Tillisch, K., Naliboff, B. D., Bushnell, M. C., and Mayer, E. A. (2010). Regional gray matter density changes in brains of patients with irritable bowel syndrome. *Gastroenterology*, 139(1):48–57.e2.

Sender, R., Fuchs, S., and Milo, R. (2016). Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS biology*, 14(8):e1002533.

Shen, H., Wang, L., Liu, Y., and Hu, D. (2010). Discriminative analysis of resting-state functional connectivity patterns of schizophrenia using low dimensional embedding of fMRI. *NeuroImage*, 49(4):3110–3121.

Shmuel, A. and Leopold, D. A. (2008). Neuronal correlates of spontaneous fluctuations in fMRI signals in monkey visual cortex: Implications for functional connectivity at rest. *Human Brain Mapping*, 29(7):751–761.

Shrout, P. E. and Fleiss, J. L. (1979). Intraclass Correlations : Uses in Assessing Rater Reliability. *Psychological Bulletin*, 86(2):9.

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., Filippini, N., Watkins, K. E., Toro, R., Laird, A. R., and Beckmann, C. F. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences*, 106(31):13040–13045.

Smith, S. M., Miller, K. L., Salimi-Khorshidi, G., Webster, M., Beckmann, C. F., Nichols, T. E., Ramsey, J. D., and Woolrich, M. W. (2011). Network modelling methods for FMRI. *NeuroImage*, 54(2):875–891.

Sporns, O., Tononi, G., and Edelman, G. (2002). Theoretical neuroanatomy and the connectivity of the cerebral cortex. *Behavioural Brain Research*, 135(1-2):69–74.

Sporns, O., Tononi, G., and Kötter, R. (2005). The Human Connectome: A Structural Description of the Human Brain. *PLoS Computational Biology*, 1(4):e42.

Tannock, G. W. and Savage, D. C. (1974). Influences of Dietary and Environmental Stress on Microbial Populations in the Murine Gastrointestinal Tract. *Infection and Immunity*, 9(3):8.

Tap, J., Derrien, M., Törnblom, H., Brazeilles, R., Cools-Portier, S., Doré, J., Störsrud, S., Le Nevé, B., Öhman, L., and Simrén, M. (2017). Identification of an Intestinal Microbiota Signature Associated With Severity of Irritable Bowel Syndrome. *Gastroenterology*, 152(1):111–123.e8.

Telesford, Q. K., Morgan, A. R., Hayasaka, S., Simpson, S. L., Barret, W., Kraft, R. A., Mozolic, J. L., and Laurienti, P. J. (2010). Reproducibility of Graph Metrics in fMRI Networks. *Frontiers in Neuroinformatics*, 4.

Tillisch, K., Mayer, E. A., and Labus, J. S. (2011). Quantitative Meta-analysis Identifies Brain Regions Activated During Rectal Distension in Irritable Bowel Syndrome. *Gastroenterology*, 140(1):91–100.

Trevethan, R. (2017). Intraclass correlation coefficients: clearing the air, extending some cautions, and making some requests. *Health Services and Outcomes Research Methodology*, 17(2):127–143.

Uddin, L. Q. (2015). Salience processing and insular cortical function and dysfunction. *Nature Reviews Neuroscience*, 16(1):55–61.

van den Heuvel, M. P., de Lange, S. C., Zalesky, A., Seguin, C., Yeo, B. T., and Schmidt, R. (2017). Proportional thresholding in resting-state fMRI functional connectivity networks and consequences for patient-control connectome studies: Issues and recommendations. *NeuroImage*, 152:437–449.

van den Heuvel, M. P. and Sporns, O. (2013). Network hubs in the human brain. *Trends in Cognitive Sciences*, 17(12):683–696.

Vanegas, H. and Schaible, H.-G. (2004). Descending control of persistent pain: inhibitory or facilitatory? *Brain Research Reviews*, 46(3):295–309.

Varoquaux, G., Sadaghiani, S., Pinel, P., Kleinschmidt, A., Poline, J., and Thirion, B. (2010). A group model for stable multi-subject ICA on fMRI datasets. *NeuroImage*, 51(1):288–299.

Vespignani, A. (2018). Twenty years of network science. *Nature*, 558(7711):528–529.

Vieira, S., Pinaya, W. H., and Mechelli, A. (2017). Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience & Biobehavioral Reviews*, 74:58–75.

Vul, E., Harris, C., Winkielman, P., and Pashler, H. (2009). Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. *Perspectives on Psychological Science*, 4(3):274–290.

Waller, L., Walter, H., Kruschwitz, J. D., Reuter, L., Müller, S., Erk, S., and Veer, I. M. (2017). Evaluating the replicability, specificity, and generalizability of connectome fingerprints. *NeuroImage*, 158:371–377.

Wang, D., Zhang, X., Zhang, X., Huang, Z., and Song, Y. (2017). Magnetic resonance imaging analysis of brain function in patients with irritable bowel syndrome. *BMC Gastroenterology*, 17.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393:3.

Welton, T., Kent, D. A., Auer, D. P., and Dineen, R. A. (2015). Reproducibility of Graph-Theoretic Brain Network Metrics: A Systematic Review. *Brain Connectivity*, 5(4):193–202.

Wilder-Smith, C. H. (2011). The balancing act: endogenous modulation of pain in functional gastrointestinal disorders. *Gut*, 60(11):1589–1599.

Zalesky, A. and Breakspear, M. (2015). Towards a statistical test for functional connectivity dynamics. *NeuroImage*, 114:466–470.

Zalesky, A., Fornito, A., and Bullmore, E. (2012). On the use of correlation as a measure of network connectivity. *NeuroImage*, 60(4):2096–2106.

Zalesky, A., Fornito, A., and Bullmore, E. T. (2010). Network-based statistic: Identifying differences in brain networks. *NeuroImage*, 53(4):1197–1207.

Zhang, J., Cheng, W., Wang, Z., Zhang, Z., Lu, W., Lu, G., and Feng, J. (2012). Pattern Classification of Large-Scale Functional Brain Networks: Identification of Informative Neuroimaging Markers for Epilepsy. *PLoS ONE*, 7(5).

Zhong, X., Shi, H., Ming, Q., Dong, D., Zhang, X., Zeng, L.-L., and Yao, S. (2017). Whole-brain resting-state functional connectivity identified major depressive disorder: A multivariate pattern analysis in two independent samples. *Journal of Affective Disorders*, 218:346–352.