Ollscoil na Gaillimhe
University of Galway

Investigating the genetics of deep learning derived neuroimaging phenotypes of brain disorders

by

Shane O'Connell, M.Sc.

A thesis submitted for the Degree of Doctor of Philosophy to the Discipline of Bioinformatics, School of Mathematical and Statistical Sciences, University of Galway.

Supervisor: Dr. Pilib Ó Broin
Discipline of Bioinformatics, School of Mathematical and Statistical Sciences, College of Science and Engineering, University of Galway

Co-Supervisor: Prof. Dara Cannon
Discipline of Psychiatry, School of Medicine, College of Medicine, Nursing, and Health Sciences, University of Galway

Submitted November, 2023

Abstract

Brain disorders are collections of debilitating phenotypes that can affect cognition and general life quality via a myriad of symptoms, including mood swings, memory loss, altered thought processes, and psychosis. Despite their common area of action in the brain, few biomarkers have been characterised. Further, the causal relationship between neuroimaging measures and brain disorders remains largely unexplored. Understanding the biological manifestations of these conditions could help to inform improved diagnostic, prognostic, and treatment systems. To this end, we identified neuroimaging biomarkers of Alzheimer's disease using a convolutional neural network, and found 7 genome-wide significant loci associated with the resultant quantity. These findings were consistent with previously observed genetic results of Alzheimer's disease and further implicated impaired cellular homeostasis as a molecular association of Alzheimer's disease-related neuroanatomical variation. We also trained an autoencoder on participant tabular neuroimaging data from the same dataset and highlighted a latent space node significantly associated with Alzheimer's participants, finding three genome-wide significant loci mapping to non-coding RNA transcripts associated with its value. Across both studies, we also demonstrate evidence of tissue-specific expression in clinically relevant brain regions, such as the substantia nigra. We finally queried the causal relationship between neuroimaging measures and bipolar disorder using graph-based Mendelian randomization methods, finding that white matter microstructural phenotypes exert greater effects in a network context than gray matter structural phenotypes. Specifically, we find evidence of bidirectional causality between bipolar disorder and the area of the lateral orbitofrontal cortex and several components of the limbic system involved in emotional regulation. Taken together, our results provide novel avenues of enquiry into the derivation of neuroanatomical biomarkers and the investigation of causal dynamics between neuroimaging and brain disorders.

Declaration

I declare that this thesis has not been submitted as an exercise at this or any other university. I declare that this thesis is entirely my own work.

Signed: Shane O'Connell

Funding

This work was conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6214.

Dedication

This thesis is dedicated to my mother, my father, my brothers, my grandfather, my uncle, and my late grandmother. Their effect on my life cannot be quantified. To borrow from Seamus Heaney, I am "steeped in luck".

Acknowledgments

This work would not have been possible without the support I have received from a great number of people. My sincerest thanks to Dr. Pilib Ó Broin who has been instrumental in shaping my scientific career. Your mentorship, patience, and advice has been invaluable to my research endeavours. To my co-supervisor, Prof. Dara Cannon – thank you wholeheartedly for your enthusiasm, curiosity, and passion, which has been a constant source of inspiration and motivation for me during my time as a PhD candidate. Our many discourses on the neuroanatomy of brain disorders have been formative and have left me with a deep passion for research. To Dr. Niamh Mullins and her lab group – thank you for fostering an overwhelmingly positive research experience during my placement in New York.

I owe a massive thanks to my graduate research committee who have humbly suffered my lengthy yearly progress review presentations with great patience. To Dr. Aaron Golden, my favourite (and only) informal academic mentor – thank you for the many words of wisdom over the last five years. To Dr. Declan Bennett – thank you for the many useful discussions on current trends in genetic research in ADB-1018. To the wider bioinformatics community in Galway – thank you all for the thoughtful research you carry out which has informed so much of what I do. To the members of the Ó Broin Lab group and the Clinical Neuroimaging Laboratories – thank you for your constant support during my time in Galway. I am lucky to have been part of such a wonderful research environment in the bioinformatics community at the University of Galway – I owe a debt of gratitude to the students and faculty at the School of Mathematical and Statistical Sciences. I would also like to thank the doctoral candidates of the SFI Centre for Research Training whose work has been a continually inspiring. In particular, I would like to extend thanks to Prof. Cathal Seoighe and Dr. Sandra Healy for their tireless efforts steering a graduate training program of approximately 100 researchers.

I would like to thank Lydia King, Sarah Ennis, and Shane Crinion for their support during the most difficult periods of my research. Our time spent during our masters in ADB-1019 has forged lasting friendships which have grown as we have all pursued our doctoral degrees. I would also like to thank my housemates Liam and Ciara for enduring my neuroses at the height of writing this work – there are not enough words to express my gratitude for your friendship.

To my friends Seán, Andrew, Cormac, Tadhg, and Eoin – thank you for the gift of music which has been a constant source of joy for me throughout my time researching this topic. I will tell myself in retrospect that the time I spent playing music when I should have been writing was part of the 'process' (whatever that means). Thank you for being part of that process.

Finally, my deepest thanks to my family who are the most important people in my life. Their significance cannot be overstated.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Psychiatry, as defined by the American Psychiatric Association, is a "branch of medicine focused on the diagnosis, treatment and prevention of mental, emotional and behavioral disorders" [1]. These conditions can include bipolar disorder (BD), schizophrenia, major depressive disorder, and attention-deficit hyperactivity disorder. Despite comprising multiple symptoms, all psychiatric disorders are linked by their primary tissue of action in the brain [2]. Similarly, neurological disorders are defined as "any condition that affects the brain, spinal cord and/or nerves", including Alzheimer's disease (AD) and Parkinson's disease [3, 4]. Psychiatric and neurological conditions can be considered brain disorders owing to this shared characteristic. The public health burden of brain-based conditions is considerable, contributing significantly to global estimates of disability-adjusted life years (DALYs) [5–7]. For example, mental and neurological conditions accounted for 10% of global DALY estimates in 2019 [8]. As a result, understanding their pathologies to inform their effective management has become of great importance to the field of brain disorder research. The previous two centuries have resulted in numerous advances, such as improved diagnostic/prognostic systems and improved therapeutic applications [9–11]. However, their modern medical treatment is complicated by an apparent lack of biomarkers associated with disorder diagnosis and prognosis. In particular, diagnostic criteria are often defined primarily by behavioural factors such as cognitive symptoms, meaning that there is no guarantee there are strong biological features underlying these phenotypes. Additionally, while certain brain disorders are marked by manifest biological changes, such as neuroanatomical variation in AD, biomarkers are often not recognised as primary clinical factors in practice, instead providing context in the presence of other diagnostic features [4]. Consequently, much recent research has focused on understanding the biology of brain disorders. As such, the fields of neuroanatomy and psychiatric/neurological genetics are of particular interest owing to observations of brain region variation in various conditions [12, 13].

1.1 Psychiatric/neurological genetics

Some of the most significant genetic developments over the last century have included the description of the structure of DNA, the subsequent characterisation of a human draft genome, and the advent of low-cost high-throughput sequencing technologies [14–16]. These factors have given rise to a wealth of research into the genetic bases of multiple human diseases, including brain disorders, yielding unprecedented amounts of data. As such, new data-driven methods to describe genetic architectures have been developed. Twin studies were initially popular for measuring the extent to which a phenotype was heritable [17], but the breadth of data available facilitated direct statistical testing of genetic associations in large populations. Most notably, genome-wide association (GWA) studies became a popular framework [18]. This methodology provided a hypothesis-free framework by which to measure the association between single nucleotide polymorphisms (SNPs) and phenotypes at the population level. As the number of genetic markers that can be tested at once is large, the discovery rate is limited by stringent multiple testing corrections which control the false positive rate [19].

These developments allowed researchers to explore novel lines of inquiry related to brain disorders. Previously, twin studies suggested that psychiatric and neurological conditions had high heritability [20]. Linkage studies were the primary means by which to measure estimate genetic marker cosegregation with illnesses. For example, early studies of BD identified several genetic factors of large effects, such as areas of chromosome 11 cosegregating in Amish pedigrees [21]. However, this result and others did not generalise to other families. In general, linkage studies designed to capture loci of large effects appeared unsuited to phenotypes such as BD. In contrast, linkage studies of other disorders were promising, evidenced by the discovery of the APOe4 allele in various pedigrees [22]. This discovery informed subsequent molecular hypotheses of AD progression, including the APOe4 cascade hypothesis [23]. Despite this, researchers were unable to explain heritability estimates from twin studies using existing genetic results.

Brain disorders are common in the general population, with 165 million Europeans estimated to be currently affected [24]. This fact in tandem with the lack of linkage markers of large effects cosegregating in family pedigrees meant that such conditions likely had more complex genetic architectures. The GWAS experimental framework was well-suited to the problem of detecting disease-associated variants of small to moderate effect in populations [25]. Early studies of BD yielded novel genetic results, including significant associations of SNPs near *NAP5* and *NTRK2* genes [26]. These genes are involved in microtubule formation in cell differentiation respectively in brain cells. Other studies suggested a role for *ANK3* and *CACNA1C* genes [27], involved in axonal membrane function and calcium channel signalling respectively. Other BD GWAS replicated the *CACNA1C* finding and identified several novel loci, including *ODZ1* with

larger sample sizes [28, 29]. The first late-onset AD GWAS identified significant SNPs in LD with the APOe4 allele as well as SNPs in the *GALP* gene, which is involved in neuronal survival and upregulated in AD brain tissues [30]. Subsequent studies increased the number of SNPs and samples tested, replicating APOe4 findings and highlighting novel significant associations [31, 32].

In general, larger sample sizes lead to an increased number of discovered genetic associations as a function of study power. The advent of biobanking consortia such as the UK Biobank were instrumental in facilitating this trend [33]. However, such consortia were not specifically designed to genotype individuals with brain disorders. Instead, data-sharing groups like the Psychiatric Genomics Consortium carried out large-scale GWAS using data from multiple research groups, providing public access to summary statistic results [34]. These efforts resulted in GWAS of increasingly large sample sizes. For example, recent GWAS of BD, major depressive disorder, AD, and schizophrenia have featured total sample sizes of 413,466, 807,553, 1,126,563, and 320,404 respectively [35–38]. For BD, regulation of insulin secretion systems and endocannabinoid signalling were identified as molecular associations [39]. More recent results results implicated ion-channel-encoding genes as significant molecular correlates through the association of *CACNB2* and *KCNB1* [38]. Additionally, significant loci identified in smaller samples were replicated in these larger cohorts, including *CACNA1C*. Recent significant genetic hits were found to be enriched in neuronal cell types, suggesting that brain development pathways are associated with BD [38]. Larger sample sizes have also allowed researchers to better estimate the degree to which genetic variation accounts for phenotypic variation in BD. For example, 20% of BD's phenotypic variance can be explained by common genetic variants [39]. Previous twin studies have suggested that BD has a heritability of .7 [20]. Additionally, fine-mapping approaches building on results from several well-powered genetic studies have prioritised several genes for follow-up investigation, such as *FURIN*, establishing a further relationship between neuronal cell type function and BD [40].

Similarly for AD, GWAS of large sample sizes in the PGC have indicated that significant genetic associations are involved in beta-amyloid clearance pathways [35]. Additionally, APOe4 has been consistently identified as a variant of interest as sample sizes have increased. Recent results have also suggested that microglia and other immune cells are associated with AD pathology [41]. SNP-based heritability estimated from the most recent large GWAS of AD was 0.03 [35].

Concurrent with molecular findings, the field of neuroanatomy has been instrumental in identifying other biological correlates of brain disorders.

1.2 The neuroanatomy of psychiatric and neurological disorders

The field of neuroanatomy has origins in Greek antiquity, with the first recorded experiments by Herophilus describing the cerebrum, cerebellum, and the ventricles in detail [42]. While this event was of great general interest, no further neuroanatomical dissections were recorded until Andreus Vesalius' activities during the Renaissance period [43]. Neuroanatomical measurements until the 20th century were mainly reliant on post-mortem tissues, which are difficult to obtain and standardise across individuals. The development of magnetic resonance imaging technologies was instrumental in transforming measurement capabilities in the 1970s [44]. Broadly, the technology manipulates the realignment of protons to a uniform magnetic signal after the application and subsequent removal of a radiofrequency wave to determine the characteristics of the tissue of origin. The speed at which energy is released during the realignment process can be used to infer the fat or water content of the area imaged, thus allowing researchers to compile structural tissue images via transformations of the output signal.

There are multiple modalities of neuroimaging that highlight presentation of different tissue types based on the frequency of pulse application. Broadly, these include T1-weighted and diffusion tensor magnetic resonance imaging procedures. T1-weighted imaging schemes usually highlight fat-heavy regions – whose protons realign quickly – as high-intensity bright regions on output images, as distinguished from proton-heavy regions such as the cerebrospinal fluid which usually appear darker. This distinction in brightness given by alternating pulse sequences allows us to image tissue structure. Diffusion tensor imaging captures the directional diffusivity or movement potential of water molecules as constrained by extra – or intra – cellular structure.

The ability to non-invasively measure the structural features of the human brain was useful to several research disciplines. Early studies were successful in detecting neuroanatomical patterns in brain disorder phenotypes such as bd. A review by [45] of 37 studies between 1987 and 1999 described consistent findings of decreased cerebellar volume and increased white matter hyperintensities. These findings allowed researchers to hypothesise about cognitive systems impacted by BD. A similar review by [46] found that medial temporal lobe atrophy was consistently predictive of AD status and correlated with mental diagnostic tests. Despite these findings, the measurement of volumetric variation was technically limited by issues in producing robust surface models of neuroimaging data. Frequently, patterns in MRI data relied upon clinicians and made formal statistical testing difficult [47]. New algorithms such as those provided by FreeSurfer provided empirical descriptions of neuroanatomical measurements [48]. This allowed researchers to compare brain structural variation across various phenotypes empirically. A meta -analysis of such studies in BD found that enlarged right lateral ventricles was a significant effect with 32 studies

included [49]. Another review by [50] summarised the functional neuroimaging findings related to BD, also noting increased volume of right lateral ventricles as well as abnormalities in prefrontal cortex regions. These studies suggested that striatal–thalamic–prefrontal systems were likely important in the context of disease progression. Further, they demonstrated the utility of neuroimaging studies in identifying biological correlates of BD. Similarly, a review carried out by [51] of early AD neuroimaging studies found that decreased hippocampal volumes correlated well with cognitive decline. They also underscored the clinical potential of neuroimaging modalities for AD diagnostic and prognostic applications. A separate review found that structural alterations in the entorhinal cortex were discriminative for early-onset AD [52].

Concurrent with these studies, large-scale imaging consortia began to proliferate, resulting in the formation of the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the launch of the Enhancing Neuro-Imaging Genetics through Meta-Analysis (ENIGMA) Bipolar Disorder Working Group [53, 54]. These standardised collections, combined with tools such as FreeSurfer, allowed for meta and mega analyses of brain disorder phenotypes and studies of general neuroanatomical variation, yielding valuable biological insights [55–57]. For example, cortical thinning was observed in BD patients [58]. Reduced amygdala and hippocampal volumes were also described [55]. Additionally, differences in white matter microstructure in the corpus callosum were evaluated as potential endophenotypes for BD [59]. In AD, widespread atrophy in hippocampal regions has been consistently observed [60], in tandem with observations of increased sulcal depth [61]. Further, a progress report by ADNI described the importance of the data resource from an exploratory standpoint [62]. The authors noted that multi-site standardised data collection protocols had resulted in a valuable research resource that was already being utilised for clinical studies.

Other general neuroimaging consortia of interest included the release of the UK Biobank imaging collection [63]. While not enriched for brain disorder phenotypes specifically, it served as a similarly valuable resource for researchers, providing large sample sizes of standardised imaging collections. Notably, the first imaging release was used to estimate brain age from structural MRI data, with differences in biological age and predictions correlating well with adverse health outcomes [64]. Additionally, other studies have made use of this cohort to identify functional MRI differences associated with depression [65]. In general, the availability of large standardised imaging collections also brought the potential the link multiple data modalities together. For instance, individuals in the ENIGMA, UK Biobank, and ADNI consortia were also genotyped. This allowed for studies investigating the genetic architecture of brain imaging measures.

1.3 Genetics of Brain Disorders and Endophenotypes

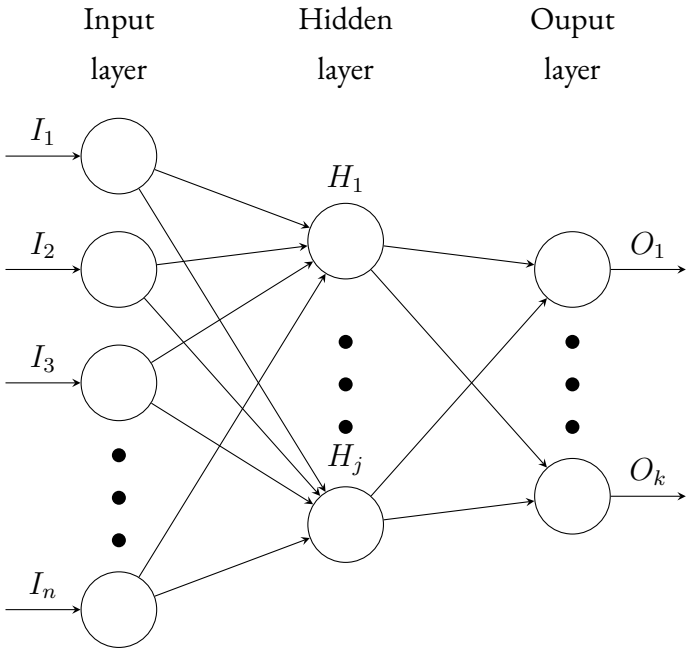
Imaging releases in the UK Biobank facilitated Elliot *et al.*'s GWAS of over three thousand measures of general brain structure and function [66], which found evidence of additive genetic effects associated with neuroanatomical variation. Their study found that structural neuroimgaing measures had high heritabilities and that significant genetic associations were enriched for brain development pathways. They also found that brain development and plasticity genes had been previously implicated in GWAS of psychiatric disorders. Concurrently, the ENIGMA consortium also captured the genetic architectures of various neuroimaging phenotypes including BD across different population samples [67]. These studies were related to the concept of endophenotype derivation. Endophenotypes in psychiatry are defined as 'internal or molecular' features of behavioural conditions that are independent of disease state [68]. The concept was introduced in 1972 as a means by which to better understand psychiatric conditions [69]. Theoretically, their description can reframe GWAS experiments as studies of quantifiable biological features of broader phenotypes, which are considered more direct expressions of gene effects. As the genetics of brain structural variation became better understood, several candidate endophenotypes of BD were considered. For instance, [70] suggested that hippocampal structural variation was consistently observed and satisfied all qualities of a strong endophenotype. Later, [59] reviewed several other candidate endophenotypes, including circadian rhythm disturbances and executive function decline. They further posited that white matter abnormalities were strong candidate endophenotypes of BD.

Similarly, a review by [71] found that medial temporal lobe atrophy was a strong candidate endophenotype of AD. A separate review by [72] discussed the potential of other reproducible structural neuroimaging signatures associated with AD as endophenotypes. The advent of neuroimaing consortia has meant that the qualities of such endophenotypes could be tested appropriately in large sample sizes. For example, GWAS by [66] allowed the SNP heritability of neuroimaging phenotypes to be better estimated. Endophenotypes in general are an attractive concept; their definition does not center genetic effects as the primary causal determinant owing to the fact that such traits must occur in the population at large. Instead, they imply a "non-hierarchical" disease model of brain disorders wherein social, environnmental, and genetic components all contribute to an outcome [73]. Data availability through large imaging collections has lead to increased focus on defining neuroimaging endophenotypes. This includes recent research into gene-environment correlates of brain structure endophenotypes of BD [74], the identification of transcriptomic signatures associated with decreased cortical thickness in mild cognitive impairment patients [75], and amygdala structural variation as an endophenotype of schizophrenia [76]. Efforts to

identify neuroimaging endophenotypes are reliant in part on establishing a statistical association between the overall phenotype and the candidate trait, which may be carried out by a variety of means.

1.4 Methodological approaches to endophenotype classification and neural networks

Large imaging collections of neuroanatomical variation has been instrumental in the study of endophenotypes. Usually, the association between individual regions and brain disorders is described in a discrete terms, such as in the case of volumetric reductions of medial frontal cortex areas in Alzheimer's disease. It may be the case that structural variation occurring in multiple regions – rather than just one – is an associated endophenotype of a condition. For example, the review by [77] details several studies reporting widespread functional connectivity disruption in Alzheimer's disease, suggesting that symptoms can manifest biologically in the form of brain network changes. Similarly, disruptions in key emotional processing networks have been previously described in BD, which can involve variation in multiple regions [78]. Furthermore, the relationship between condition and regions may not be linear. Other modelling strategies to capture complex data patterns include deep learning approaches such as neural networks [79]. Neural networks can model non-linear data relationships via a number of sum-weighted transformations of input data mapped to the following layer via nodes; each layer can contain multiple nodes and this process can be repeated. Here, the value of any node is determined by the input values multiplied by their respective node weights, summed together, and passed through a function which can impart non-linearity on the output signal (termed an activation function). This operation can then be repeated multiple times, resulting in user-specified outputs depending on the problem domain. We can conceptualise this with the following:



. (1.1)

In Equation 1.1, the values I_1 through I_n define an input of rank n. The values of H_1 at the subsequent layer (often termed the hidden layer) are defined by I_1 through I_n multiplied by their respective weights and summed together, denoted as black lines in the above diagram. The final value of H_1 is given by applying an arbitrary function (termed an activation function) to the summed quantity, resulting in a non-linear output (where the specified function is non-linear). Similarly, the other hidden layer node values up to and including H_j are given by the same operation, except each node of the set H have their own respective weights mapping I to the output value. The same procedure is then carried out in our two-layer example to yield O_1 through O_k, given by sum-multiplying H_1 through H_j by their respective weights linking them to nodes of O. Again, an activation function is applied yielding the final O value.

Owing to their general flexibility, several variants of classical neural networks exist. One such example is the convolutional neural network, which is specifically designed to leverage spatial patterns in imaging data types.

1.5 Convolutional Neural Networks

Convolutional neural networks were developed in the late 1980s as a generalisable, shift-invariant extension to standard neural network architectures [80]. This means that learned patterns can be present at any coordinate position respective weight filters are multiplied by every image patch, making their rep-

resentations pattern-specific as opposed to position-specific. In contrast, classic neural networks – or feed-forward architectures – are spatially insensitive. These developments increased the generalisability of image-based predictive models with fewer free parameters than fully-connected feed forward architectures. Such approaches began gaining renewed attention with the development of efficient backpropogation strategies to apply weight updates to large networks via gradient descent [81]. Soon after their introduction, their application became widespread in a variety of imaging domains. For example, the authors of [80] achieved impressive testing accuracies of convolutional neural network models on handwritten digit recognition tasks. Other studies demonstrated similarly impressive results in facial recognition [82]. Early studies of these methods applied to biomedical imaging domains found encouraging initial results in the identification of lung nodules in chest radiographs [83]. Further, the shift invariant properties of convolutional models proved attractive for other aspects of computer vision, such as in hand tracking [84].

The application of such technologies to neuroimaging was a natural next step. However, this was limited by lack of access to sufficient sample sizes. This was further compounded by the technological constraints impacting efficient training schema, as the amount of parameters to be trained is often large. The accessibility of standardised neuroimaging collections allowed renewed focus on this topic, and computing infrastructure has become rapidly more sophisticated in recent years [85, 86]. These advances provided the necessary conditions for novel lines of enquiry in the field of brain disorder neuroimaging.

1.6 Deep learning in brain disorder neuroimaging

With large sample sizes, standardised imaging collections, and rapid computing infrastructures, convolutional neural network studies of neuroimaging data became widespread. Notable examples include classification tasks of AD, BD, and schizophrenia [87–89]. Strikingly, these studies demonstrated higher testing predictive performances than other linear approaches. Particularly, studies making use of ADNI data achieved near-perfect classification accuracies in several instances [90–92]. Further examples of convolutional neural network applications to brain disorder phenotypes will be considered in a later chapter. Additionally, the modelling framework proved useful for other diverse tasks, including brain age estimation [93] and tumour segmentation [94]. These varied studies demonstrated the potential of the technology. Further, the features used for predictive purposes may comprise multiple brain regions or be non linear, which may be of interest in the context of endophenotye and biomarker derivation. The interpretability of such approaches is complicated by several factors, including model depth and the fact that model coefficients do not have the same properties as statistical coefficients. Despite this, interpretative

efforts have been developed which allow for visual examination of regions important to model classification. They involve taking the gradient of the output with respect to the input and overlaying this quantity on the input image to act as an "importance heatmap" [95]. Variants of this procedure multiply this quantity by learned model weights [96]. Other approaches, such as the use of counterfactuals, attempt to reverse-engineer input examples with specific properties and measure the change in model output [97]. This allows researchers to understand what features bear the most importance in the model's learned representation. Despite this, the reliability of interpretability methods for convolutional neural networks is contentious, with recent studies critiquing several popular methods [98]. Convolutional neural networks will be dealt with in greater detail in the next chapter.

1.7 Thesis statement and general outline

With these factors in mind, this work aims to derive endophenotypes of debilitating brain disorders using deep learning methods and examine their genetic properties. We also seek to investigate formally the causal relationship between neuroimaging and brain disorders. In Chapter 2, we begin by systematically reviewing the literature on predictive CNNs applied to magnetic resonance imaging data of brain disorders. In Chapter 3, we use insights from our observations to develop an experimental framework centering interpretability in a custom CNN model applied to an AD dataset. We examine the neuroanatomical features associated with AD through our derived endophenotype and carry out a GWAS, characterising its genetic architecture. In Chapter 4, we develop an alternate deep learning approach – a custom autoencoder – to model the same problem in the same AD dataset with several additionally useful unsupervised properties compared to CNNs. We determine the brain imaging features driving discriminative node activity and carry out a GWAS of its activity. We further compare our two GWAS to a GWAS of binary AD status and simulation experiments. Finally, we use Mendelian randomization approaches to build networks of causality between neuroimaging measures and a separate psychiatric phenotype, bipolar disorder. We query our results at the network level to determine regions and collections of regions that are of particular interest in the context of the overall disorder.

Preamble to Chapter 2

This work has been published in Human Brain Mapping and is available via the following link: https://onlinelibrary.wiley.com/doi/10.1002/hbm.26521[99]. It was jointly supervised by Dr. Pilib Ó Broin and Prof. Dara Cannon; the candidate's contribution included reading all papers, writing the manuscript, making the figures, and engaging with the peer review process.

Chapter 2

Predictive Modelling of Brain Disorders with Magnetic Resonance Imaging: A Systematic Review of Modelling Practices, Transparency, and Explanatory Efforts in the use of Convolutional Neural Networks

2.1 Introduction

Brain disorders, which include bipolar disorder, Alzheimer's disease, and schizophrenia, are a collection of debilitating neurological and psychiatric conditions characterised by a variety of features including impaired cognition, altered mood states, psychosis, neurodegeneration, and memory loss [2]. These phenotypes, each with varied clinical presentations, are all associated with neuroanatomical changes, incurring public and personal health burdens through reduced quality of life, social stigma, and increased mortality [2, 100]. As such, these conditions are the focus of intense research across multiple disciplines. There is significant interest in building predictive models designed to differentiate conditions and their subtypes, which could incorporate biological information into current clinical frameworks and yield mechanistic

insights via the biomarkers used [101, 102]. Additionally, biomarker-informed diagnoses could offer the potential of early intervention and management, a concept well understood in general medicine [103]. Magnetic resonance imaging (MRI) provides non-invasive measures of brain structure and the increasing availability of large-scale collections of MRI data has enabled a wealth of predictive modelling studies [104, 105].

Previously, machine learning and classical statistical approaches have been used to highlight differential neuroanatomical patterns across several conditions, including subcortical structure volume reduction in bipolar disorder and Alzheimer's disease [55, 106]. However, incorporating such information into clinical systems is non-trivial, as the dynamics and limitations of a particular biomarker must be addressed prior to use [107, 108]. Additionally, the methods used to identify discriminative features have their own considerations, such as requiring preprocessing tools to derive tabular brain summary information [109, 110]. These tools can produce variable results depending on the parameters chosen, even when applied to the same dataset, highlighting the importance of domain expertise to justify decisions [111]. Additionally, statistical modelling often requires formal specification of expected variable relationships, and generally are unsuited to high-dimensional imaging data structures. Traditional machine learning approaches are also limited by their inability to consider spatial dependencies between groups of pixels, making it necessary to use tabular summary data. With these factors in mind, deep learning algorithms – and particularly those well-suited to imaging – have become a popular methodology. This is because of their ability to consider arbitrarily complex relationships, providing greater model flexibility without specification of expected variable relationships. Convolutional neural networks (CNNs) are deep learning models designed to detect spatial patterns in imaging data and have shown impressive predictive performances in various classification tasks. They have also been widely applied in the field of medical imaging for segmentation and prediction, particularly in the context of aging and psychiatric/neurological disorder diagnosis [95, 112–114].

These recent developments have been enabled by access to large standardised neuroimaging data collections, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) [53] and the UK Biobank [63]. The predictive capabilities of these approaches is promising in the context of potential clinical applications; brain disorder classifications are usually based on life history information and questionnaires, and thus leveraging models making use of neuroanatomical measures could supplement existing diagnostic frameworks. However, there are a few caveats that bear consideration; firstly, deep learning models can suffer from a number of limitations, such as high parameter dimensionality, lack of interpretability, weight stochasticity, lack of uncertainty, and difficulty to train [115–118]. Secondly, clinical decision systems require rigorous validation and reporting frameworks for more interpretable models; the use of opaque

deep learning algorithms make validation and transparency more difficult to achieve [119, 120]. Clinical decision systems that offer no explanation of a classification are less likely to be incorporated into patient care frameworks. These factors combine to make the application of deep learning to clinical settings challenging, particularly where medical imaging is concerned.

As the number of studies applying deep learning to brain disorder prediction using neuroimaging data increases, the opportunity arises to examine factors that may limit their potential for clinical application. In this work, we systematically review 55 papers which report on such approaches. While many of the studies examined have been designed to demonstrate the predictive capabilities, we sought to assess the existing literature with the aim of identifying key principles that can maximise the potential clinical value of future work; these principles are: 1) modelling practices, 2) transparency, and 3) explanatory efforts. Below, we first provide a brief overview of CNNs and their workflow in the context of brain disorder imaging-based models, and subsequently detail our motivation behind these three principles; we then analyse the selected papers in the context of these principles and suggest several recommendations for future studies based on our results.

2.1.1 Convolutional Neural Networks

CNNs are a popular deep learning algorithm for many areas of research, particularly those utilising MRI data [95, 112, 114, 121]. Their structure is designed to account for spatial data patterns; this is accomplished through the use of filters and feature maps. A feature map is derived via *convolutional operations*, which are a matrix multiplication between a weights vector of an arbitrary window size and an input image patch of the same size. Every number in the input window is multiplied by every number in the filter and summed together, becoming a pixel value of a new feature map at the next layer. The convolution of the same filter over every patch of the input image generates the entire output feature map, which is usually the same size as the input image. Multiple feature maps are used in CNN architectures, each with their own filters, which, throughout model training, can detect distinct data patterns such as shapes and/or edges. CNNs build increasingly abstract representations of data through iterative transformation operations until such a time as feature maps are represented in a 1-dimensional list and fed to a fully connected neural network. Similarly, all variables at successive layers are the sum-weighted combination of all previous layer variables. Weight initialisation is often random and training is carried out via backpropagation. More in-depth considerations of neural networks and their training can be found in LeCun et al., 1995 and 2012 [117, 118].

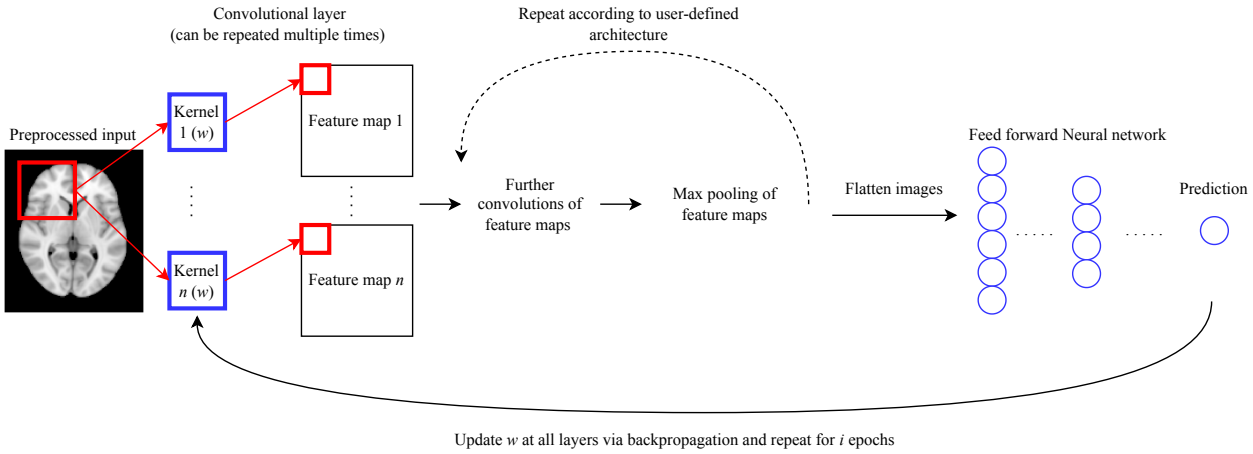


Figure 2.1: General experimental workflow. The preprocessed input image, either in 2 or 3 dimensional format is passed to a CNN model (or ensemble of CNN models) for training and prediction, The weights vector w is updated via backpropagation at each epoch, minimising the error of the loss function.

2.1.2 CNN Implementations

MRI-based predictive modelling of brain conditions with deep learning models generally follows the pipeline presented in Figure 1, or a variant thereof. Preprocessing is usually applied to skull strip, register raw input images, crop, resize, and/or contrast normalise. The preprocessed inputs are then used as training data for a CNN (or an ensemble of CNNs). Owing to the fact that many existing CNN models have been applied to 2D data domains, studies in the medical imaging field can adapt their data to fit existing architectures via transfer learning or train new models in the 3D space, as structural MRI scans are usually 3D [122, 123]. Some studies also train custom architectures on 2D data [124–126]. The output is usually presented as a probability, which is then used to calculate performance metrics such as the area under the receiver operating characteristic curve (AUC) and accuracy.

In the following sections, we define and justify our emphasis on modelling practices, transparency, and explanatory efforts in the context of brain disorder classification using neuroimaging data for potential clinical benefit. We note that these principles bear domain-agnostic importance for predictive studies and overlap with recent recommendations for improving the clinical and biological translational potential of machine learning and deep-learning models [127].

2.1.3 Modelling Practices

Modelling practices here refers to the reliability of the methodology used; procedures demonstrating the reliability of experimental results by maximising reproducibility are increasing the potential clinical value of a study. Generally, studies that can be reproduced and that have attempted to mitigate factors that can influence the reliability of results are more likely to witness clinical integration. As previously mentioned, deep learning models have a number of unique features that can make this task difficult, but several procedures can be observed. We examine the presence of repeat experiments, the data splitting procedure, the reported accuracy, and the data representation strategy to evaluate this principle.

Repeat experiments ensure that the reported performance metrics are trustworthy across multiple random weight initialisations and that the system as a whole can be expected to perform well if retrained. This is pertinent given that CNNs are parameter-dense, making them more prone to overfitting. A useful type of repeat experiment includes k-fold cross validation, whereby data is split into k folds, and $k-1$ folds are used to train the model and the k-th fold serves as the testing set. This procedure is repeated k times, until every fold has served as the testing set, providing an estimate of model performance on multiple data splits.

The reported accuracy is the final performance of the model as estimated from an evaluation strategy, which can include *k*-fold cross-validation, performance estimation on a separate test split within the same population, or estimation on a separate population. The overall capacity of a model to classify a brain disorder with fidelity will ultimately impact its potential for clinical integration.

The data representation strategy is of specific importance for CNN models in this domain, as structural MRI data is 3D, whereby each number is represented by a pixel. Thus, modelling entire volumes can be computationally expensive, and some studies may opt to split data into individual 2D slices. This comes with a set of considerations: firstly, each 2D slice is treated as an individual instance during conventional training procedures, meaning that performance metrics can either be reported per slice or combined to derive patient-level quantities, prompting consideration of voting strategies; secondly, 2D data are more prone to information leakage if train-test splitting is carried out after 2D slice derivation because the model may have been exposed to data from the same patient during training and evaluation. Information leakage can inflate performance estimates by exposing a model to testing information during training steps. Additionally, studies may take multiple 3D patches per patient that could result in similar issues. Together, these issues can contribute to inflated estimates of performance.

2.1.4 Transparency

Transparency refers to how clearly the study's methods are reported, including code and model sharing. This principle bears general importance, particularly for models with clinical potential [128]. Several important advantages to code sharing having been described previously, including facilitating greater understanding of experiments and facilitating reproducibility [129, 130]. There are many hyperparameters associated with deep learning models which can affect performance, making transparent reportage necessary. Descriptions of model architectural choices and training schedules can help to increase potential for clinical translation through increased reproducibility and understanding of studies. Furthermore, model weight sharing can mitigate the computational overhead of model training. Additionally, explicitly stating data sources is important for readers to understand the data demographics that yielded the reported results and model.

2.1.5 Explanatory efforts

Explanatory efforts, otherwise known as interpretability efforts, refer to the attempts made to identify and explain features driving model predictions [131]. Deep learning systems can be difficult to interpret, but efforts can be made to examine image regions that are used during prediction to determine whether that information is relevant. This is particularly important as CNNs are prone to overfitting and can make use of any image feature, in turn making algorithmic biases more likely if not examined [132, 133]. Ensuring CNNs are using relevant information can increase clinical potential and confidence in the system. Models can be interpreted by saliency methods such as gradient-based class activation mapping [95, 96], which rely on deriving the gradient of model output with respect to input and weighting that quantity by the input – the final metric is then overlaid on the input for visualisation. This can indicate what regions are most 'important', but they are not directly comparable to coefficients from classical regression models. Another approach to understanding model behaviour is counterfactuals, which involve measuring the changes in predictive performance of models when they are exposed to inputs with known qualities; for example, noting the change in model output when a patient image with a thicker amygdala is used as the input [97]. We investigate explanatory efforts via the application of methods that produce a saliency map (such as [96] or [95]), which are gradient-based, or the visualisation of internal feature map outputs. These steps serve to increase confidence in models and consequently its potential clinical value.

2.2 Methods

We conducted a systematic literature review according to PRISMA guidelines [134], the details of which are provided below.

2.2.1 Inclusion/exclusion criteria

We limited our search to consider studies making use of traditional CNN architectures exclusively, whereby convolutional layer outputs, or other model outputs, are not used to train separate machine learning models. This is because they are the most common architecture and as such this better enables comparisons across studies. We also focused our attention on studies that use structural MRI data, as functional MRI data structures can often have different modelling requirements, including the use of time series methodologies that make them more difficult to compare.

2.2.2 Search details

We performed a Web of Science (all databases) and Pubmed search with the following keywords:

((((structural) or (T1-weighted)) AND (imaging)) AND ((MRI) OR (T1 MRI)) AND ((CNN) OR

(convolutional neural network) OR (3D-CNN))) AND (psychiatric OR depression OR autism OR bipolar OR Alzheimer's OR neurological) NOT (segmentation)

For Web of Science, 77 results were returned, and 114 results were returned from Pubmed. Titles and abstracts were screened for relevance to the research question, and duplicates across both databases were removed, leaving a total of 74 papers. Nineteen studies were excluded for using functional MRI data and applying hybrid models where CNNs were not the primary modelling method; this resulted in a total of 55 papers remaining for review. The flowchart of this process is presented in Figure 2.

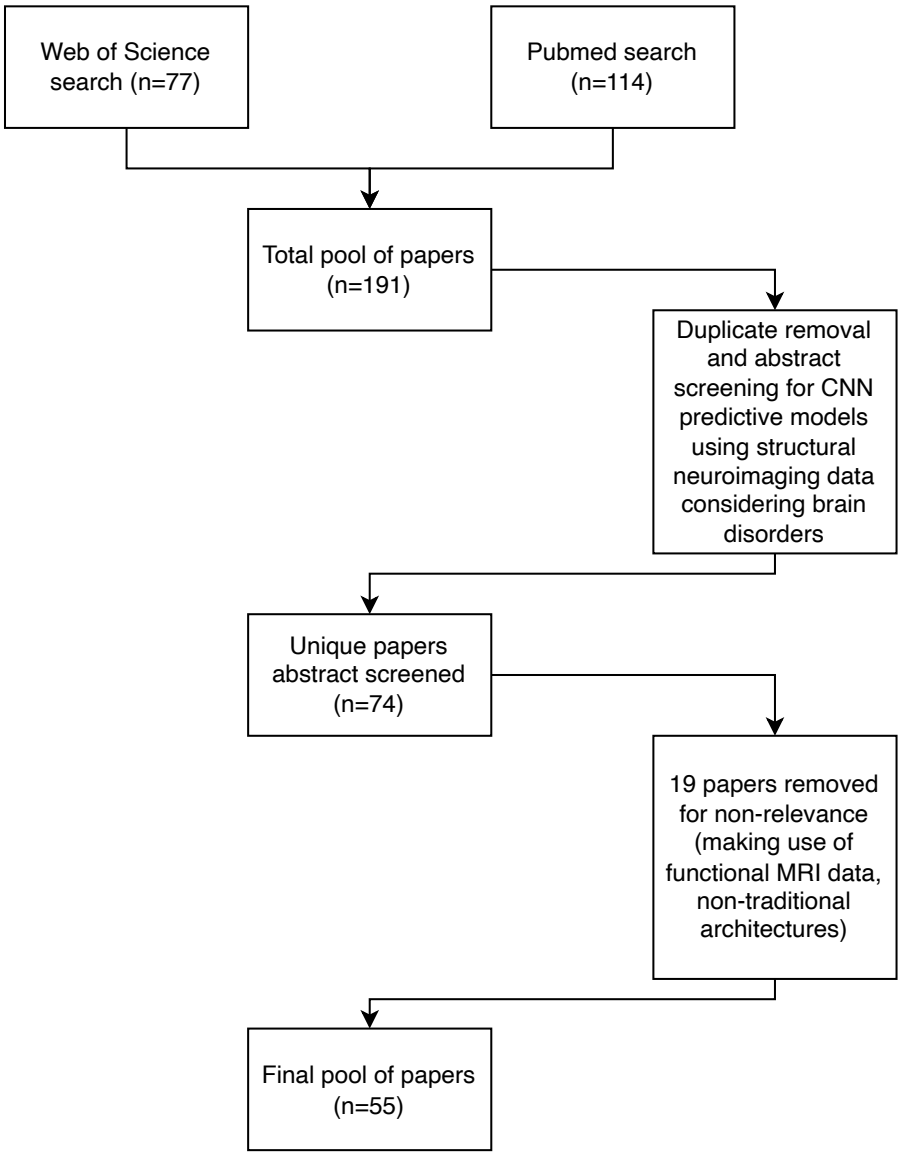


Figure 2.2: Flowchart detailing the paper selection process.

2.2.3 Desired variables

A standardised questionnaire was designed to evaluate the methodological details of the studies considered, including the presence or absence of repeat experiments, the overall data representation strategy, the reported accuracy, the sample size, the data source, whether or not an explanatory method was applied, and whether or not code was made available. For obtaining accuracy, we recorded the highest performances in testing experiments reported by authors as the value for that study, with an aim to capture the accuracy in the main classification test (for example, certain studies made use of 3-way classifiers for Alzheimer's,

cognitive impairment, and controls - we took only the Alzheimer's vs. control performances as our value for that study). We marked sample size as NA where patient-level data numbers were not reported. We fit t-tests to determine whether or not accuracy statistically varied across binary categories and a linear regression to examine the relationship between accuracy and sample size. We also fitted a linear regression of accuracy against all measured variables (except code availability) to examine whether or not different study attributes were predictive of reported accuracy.

2.3 Results

We organise our findings according to our three principles: modelling practices, transparency, and explanatory efforts. The selected papers and their attributes can be found in Table 1, and a numerical summary of the results can be found in Table 2.

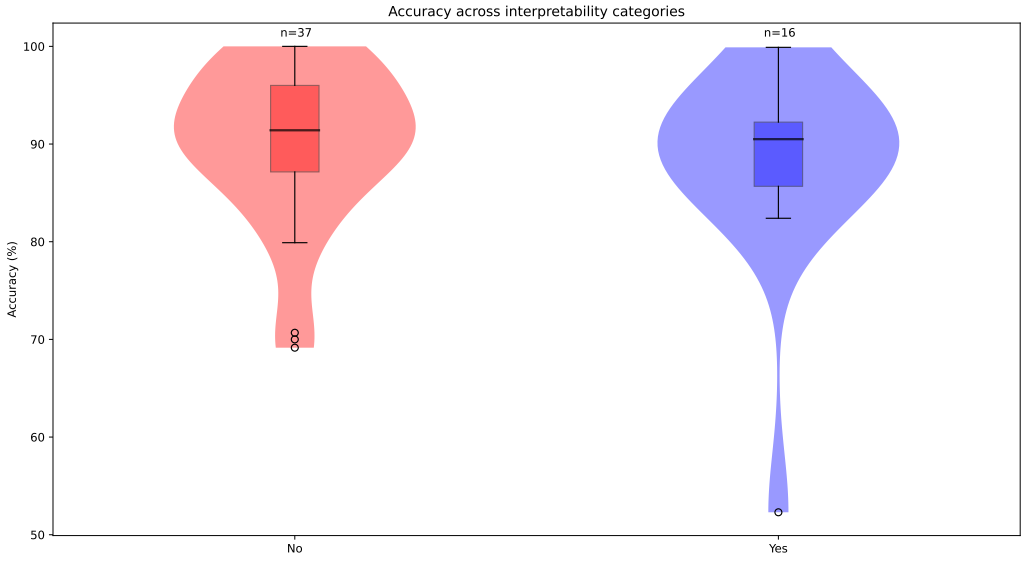
2.3.1 Modelling practices across studies

We found that 24 out of 55 papers represented data in 2D (Table 2). While this is more computationally efficient than 3D, it can make information leakage more likely. This is because 3D-based data representations have one quantity per patient and 2D-based representations will often have multiple quantities per patient, requiring extra care during the data splitting procedure. This is not unique to 2D-based studies as there may be patch-based 3D approaches deriving multiple quantities per patient. Accuracy calculation can be carried out per slice or per patient, introducing issues surrounding optimal voting strategies. Of the 24 studies making use of 2D slices, only one referred to voting methods [135]. Several studies made use of single slices per patients [125, 136, 137]. Even with comprehensive image registration, there is no guarantee that the same biological information is considered per patient with this approach. One paper making use of 2D slices provided code, detailing how individual patient volumes were split into 2D images [138].
We noted that 24 out of 55 studies made use of multiple models for training and prediction, with some papers using the output of one trained CNN as the input to another [139–143]. This may impact generalisation by increasing the chances of overfitting. A number of studies used statistical tests to pre-select informative image patches which can introduce bias by focusing the model on regions which may not be informative in full models [136, 142, 143]. Furthermore, pre-selecting regions based on accuracy metrics in one population may influence generalisation capacity in another. In several studies, one model was trained on the whole dataset the weights from that model were used for transfer learning of another model in the same dataset, leading to potential leakage or overfitting [126, 135, 136, 144, 145]. We note that while

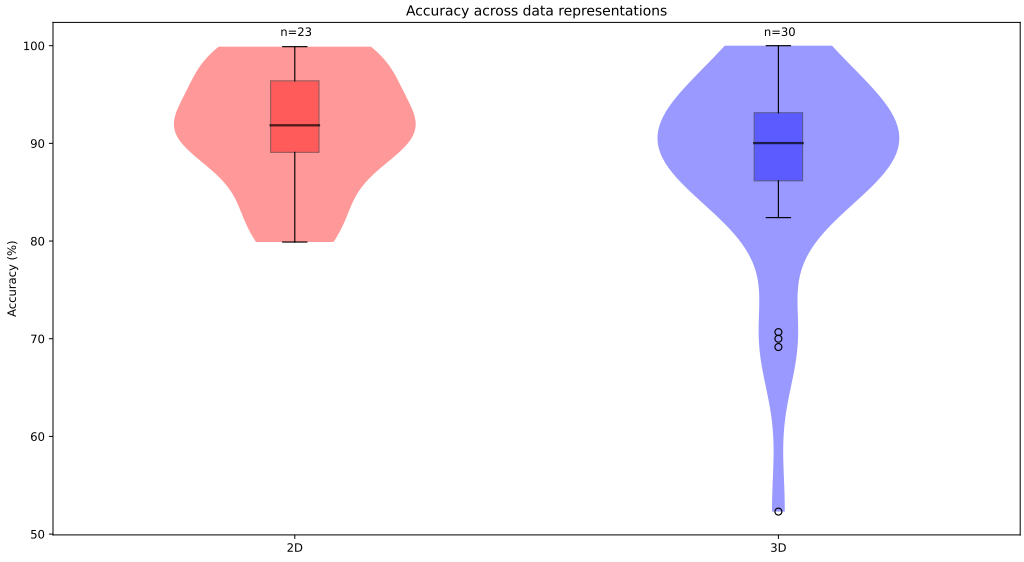
overfitting mitigation strategies can be employed, in cases where the weight training has been informed by access to testing labels, no degree of post-leakage mitigation can remedy these specific effects.
Thirty out of 55 studies employed repeat experiments, ten of which reported only point estimate performance metrics. We also found the mean accuracy across all studies was $89.36 \pm 8.694\%$ ($\mu \pm \sigma$, mean and standard deviation). There was no significant relationship between reported accuracy metrics and any study attributes, although accuracy appeared to be inversely correlated with increased sample size (Figure 2.4). Further, a regression of accuracy against all variables yielded non-significant test statistics for every coefficient. 44 of 55 studies made use of the ADNI dataset during either training or testing. The mean sample size of the 55 considered studies was 828 ± 691.

2.3.2 Transparency, interpretability, and explanatory efforts across studies

We found that 49 out of 55 papers did not provide code or model weights, meaning that the majority of studies relied on textual methods summaries. This implies limited methodological transparency which is an issue considering how modelling choices can impact system performance. Studies reporting code facilitate clear, reproducible experimental practices [138, 144, 146–149]. However, we note that every study made explicit mention of the database their studies came from and textually described a preprocessing pipeline.

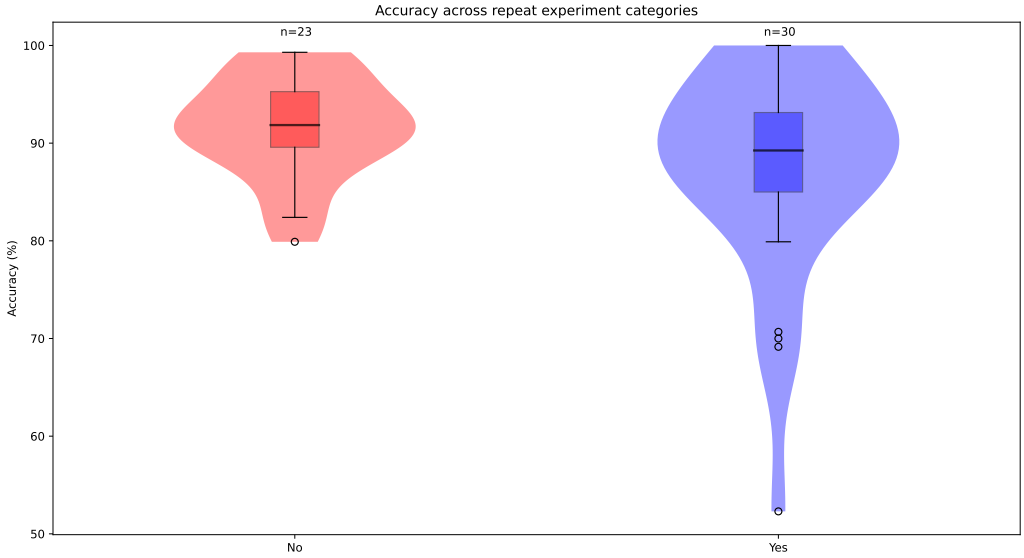


(a)

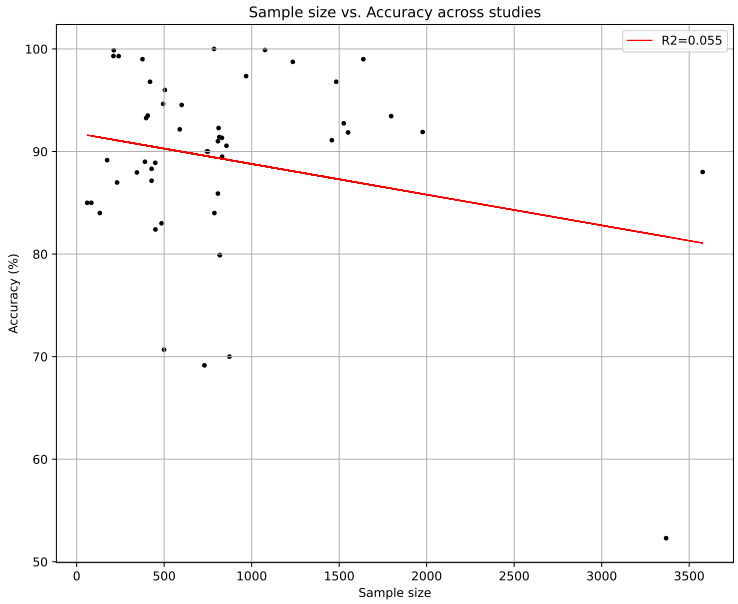


(b)

Figure 2.3: Plots of accuracy variation across binary categories (a=interpretability, b=representation). Studies where accuracy was not reported were excluded.



(a)



(b)

Figure 2.4: Violin plot of accuracy variation across repeat experiment studies (a) and scatter plot of accuracy against sample size per study (b). Studies where accuracy was not reported were excluded. The line of best fit estimated using a linear regression is denoted in red (b).

Table 2.1: Tabular presentation of the studies considered for this systematic literature review.

Authors and citation	Modelling Practices			Transparency	Explanatory Efforts	Database(s)	Sample size
	Data representation	Repeat experiments	Accuracy (%)	Code availability	Saliency		
Zou *et al.* (2017) [114]	3D	Yes, repeated on same data split	69.15	No	No	ADHD-200	730
Taheri Gorji & Kaabouch (2019) [90]	2D	No	94.54	No	No	ADNI	600
Spasov *et al.* (2018) [91]	2D	No	99.0	No	No	ADNI	376
Li & Liu (2019) [150]	3D	Yes, cross-validation	91.0	No	Yes, saliency map	ADNI	807
Liu *et al.* (2020) [92]	3D	Yes, cross-validation	88.9	No	No	ADNI	449
Li *et al.* (2017) [151]	3D	Yes, cross-validation	88.31	No	No	ADNI	428
Folego *et al.* (2020) [144]	3D	Yes, cross-validation	52.3	Yes, *github.com/gfolego/alzheimers*	Yes, saliency map	ADNI, CADD, AIBL, MIRIAD, OASIS	3368
Marzban *et al.* (2020) [152]	3D	No	93.5	No	No	ADNI	406
Hosseini-Asl *et al.* (2018) [121]	3D	Yes, cross-validation	99.31	No	No	ADNI	210
Gunawardena *et al.* (2017) [153]	2D	No	96.0	No	No	ADNI	504
Basaia *et al.* (2019) [154]	3D	No	99.0	No	No	ADNI, In-house	1638
Tufail *et al.* (2020) [155]	2D	Yes, cross-validation	94.64	No	No	OASIS	494
Hu *et al.* (2020) [156]	3D	Yes, cross-validation	70.68	No	No	NUSDAST,IMH	499
Cheng *et al.* (2017) [157]	3D	No	87.15	No	No	ADNI	428
Nanni *et al.* (2020) [158]	3D	No	NA	No	No	ADNI, S509	773
Lin *et al.* (2018) [145]	2D	Yes, cross-validation	79.9	No	No	ADNI	818
Billones *et al.* (2016) [123]	2D	No	91.85	No	No	ADNI	NA
Barbaroux *et al.* (2020) [124]	2D	Yes, cross-validation	92.16	No	No	ADNI	589
Yigit & Işik (2020) [159]	2D	No	83.0	No	No	OASIS, MIRIAD	485
Pan *et al.* (2020) [160]	2D	Yes, cross-validation	84.0	No	No	ADNI	787
Ahmed *et al.* (2020) [135]	2D	No	90.73	No	No	GARD, ADNI	NA
Ortiz-Suárez *et al.* (2017) [161]	2D	Yes, cross-validation	85.0	No	Yes, feature map	OASIS	84
Aderghal *et al.* (2017) [125]	2D	No	91.41	No	No	ADNI	815
Lian *et al.* (2020) [162]	3D	No	91.1	No	Yes, saliency map	ADNI	1458
Li *et al.* (2021) [163]	3D	Yes, cross-validation	85.9	No	Yes, saliency map	ADNI	807
Cui & Liu (2018) [164]	3D	Yes, cross-validation	86.98	No	No	ADNI	231
Aderghal *et al.* (2020) [165]	2D	No	91.86	No	No	ADNI	1551
Böhle *et al.* (2019) [146]	3D	No	87.96	Yes, *github.com/moboehle/Pytorch-LRP*	Yes, saliency map	ADNI	344
Sarraf *et al.* (2019) [138]	2D	Yes, cross-validation	99.9	Yes, *github.com/samansarraf/MCADNNet*	Yes, feature map	ADNI	1076
Liu *et al.* (2018) [166]	3D	Yes, cross-validation	93.26	No	Yes, saliency map	ADNI	397
Zhang *et al.* (2020) [167]	3D	Yes, cross-validation	85.0	No	Yes, feature map	In-house	60
Lee *et al.* (2019) [168]	2D	Yes, cross-validation	98.74	No	No	OASIS, ADNI	1235
Qiu *et al.* (2020) [147]	3D	Yes, repeated on same data split	96.8	Yes, *github.com/vkola-lab/brain2020*	Yes, feature map	ADNI, AIBL, FH, NACC	1483
Spasov *et al.* (2019) [148]	3D	Yes, repeated on different data splits	100.0	Yes, *github.com/simeon-spasov/MCI*	No	ADNI	785
Sun *et al.* (2020) [169]	3D	Yes, cross-validation	84.0	No	No	ADNI	132
Oh *et al.* (2020) [170]	3D	Yes, cross-validation	70.0	No	No	BGS, COBRE, MCICS, NMCH, NUSDAST	873
Lian *et al.* (2020) [139]	3D	No	91.9	No	Yes, feature map	ADNI, AIBL	1977
Cui & Liu (2019) [171]	3D	Yes, cross-validation	91.33	No	Yes, saliency map	ADNI	830
Mendoza-Léon *et al.* (2020) [136]	2D	No	89.16	No	No	OASIS	174
Pelka *et al.* (2020) [126]	2D	Yes, cross-validation	90.0	No	Yes, saliency map	Heinz Nixdorf Recall, ADNI	744
Li & Liu (2018) [140]	3D	Yes, cross-validation	89.5	No	Yes, feature map	ADNI	831
Bae *et al.* (2021) [172]	3D	No	82.4	No	Yes, saliency map	ADNI	450
Cui & Liu (2019) [141]	3D	Yes, cross-validation	92.29	No	No	ADNI	811
Liu *et al.* (2018) [142]	3D	No	90.56	No	No	ADNI, MIRIAD	856
Liu *et al.* (2018) [143]	3D	Yes, cross-validation	92.75	No	No	ADNI, MIRIAD	1526
Al-Khuzaie *et al.* (2021) [173]	2D	No	99.3	No	No	OASIS	240
Zhang *et al.* (2021) [174]	3D	No	97.35	No	No	ADNI	968
Hu *et al.* (2021) [149]	3D	No	93.45	Yes, *github.com/BigBug-NJU/FTD_AD_transfer*	Yes, feature map	ADNI, NIFD	1797
Herzog & Magoulas (2021) [137]	2D	No	90.0	No	No	ADNI	750
Yee *et al.* (2021) [175]	3D	Yes, cross-validation	88.0	No	No	ADNI, AIBL, OASIS, MIRIAD	3577
Mukhtar & Farhan (2020) [176]	2D	No	79.9	No	No	ADNI	818
Bae *et al.* (2020) [177]	2D	Yes, cross-validation	89.0	No	No	ADNI, SNUBH	390
Nigri *et al.* (2020) [178]	2D	No	NA	No	Yes, saliency map	ADNI, AIBL	826
Li *et al.* (2021) [89]	2D	Yes, cross-validation	99.87	No	No	In-house	212
Kiryu *et al.* (2019) [179]	2D	No	96.8	No	No	In-house	419

Question	Answer
How are data represented?	2D(n=24), 3D(n=31)
Is code available?	No(n=49), Yes(n=6)
Mean Accuracy ($\mu \pm STD$):	$89.36 \pm 8.694\%$
Is interpretability considered?	No(n=38), Yes(n=17)
Are there repeat experiments?	No(n=25), Yes(n=30)

Table 2.2: Numeric summary of study attributes from the 55 papers satisfying selection criteria.

Seventeen of 55 studies made efforts to explain models by applying a saliency method [96] or visualising feature maps [95]. Of these 17, 4 papers discussed their interpretation of interpretability outputs in their findings [146, 147, 166, 178]. Five of the 17 papers employing applying interpretability methods provided code, but just three code releases contained information on saliency method implementation [146, 147, 149].

2.4 Discussion

Below, we discuss the findings summarised in Table 2 in and propose several recommendations to maximise the potential clinical value of future studies making use of CNNs to predict brain disorders from structural neuroimaging data collections.

2.4.1 Data representation

A majority of papers made use of 3D data representations, either by deriving multiple patches per patient of arbitrary sizes or modelling entire brain volumes (patch-based/region-of-interest-based). Using entire brain volumes on a subject level can reduce the chances of information leakage while preserving

one volume per patient, but is computationally expensive. Modelling multiple 3D patches per patient can potentially lead to information leakage if patch derivation is carried out before patient-level train-test splitting, and testing accuracies can be represented by different voting strategies. While 3D patch modelling can consider the spatial dependencies across 3 axes of brain data, it is unclear as to the benefits of focusing on smaller 3D regions compared to entire 2D images along one axis. A significant minority of papers made use of 2D data structures (24/57), which are an attractive prospect considering the high computational burden of modelling in 3D and the ability to capture all brain information along one dimension. Workflows making use of multiple 2D images per patient can also be prone to information leakage and may represent testing accuracies by different means, thus requiring a greater level of care. It is difficult to declare one data representation strategy as optimal compared to another, given the limiting factors associated with every approach. Further, we found that there was no statistically significant differences between reported accuracies across data representations (2D = $91.5 \pm 6\%$, 3D=$88 \pm 9\%$). This suggests that if performance inflation has occurred, it does not appear to be enriched for a specific data representation strategy. Nevertheless, researchers should be cognizant of the individual limitations associated with each experimental approach and proactively address issues where possible. For example, information leakage can be mitigated by ensuring slice (or patch) conversion post-data splitting at the patient level, which can be verified by providing well-annotated code. Additionally, where multiple slices or patches have been used per patient, the voting strategy should be explicitly detailed.

Utilising single 2D slices may lead to performance estimation inflation owing to the fact that there is no guarantee the same biological information is being considered per patient at the same slice index. Several studies also made use of model stacking, whereby the input of a model is the output of another trained model. This may impact the model's ability to generalise to different data by increasing the chances of overfitting. This is because the first model in stacking situations has already derived a representation of the data informed by test labels. This is distinct from using traditional unsupervised dimensionality reduction techniques to derive an input for a subsequent predictive model due to bias. Additionally, deep learning systems can be opaque, making it difficult to understand the first deep learning model's data representation and consequently the properties of the input used for the final predictive model.

2.4.2 Repeat experiments

Most studies implemented repeat experiments via cross-validation, which can account for performance estimation variation caused by weight initialisation stochasticity and fold splitting. A related strategy that may be considered is nested cross-fold validation, whereby the training set every iteration undergoes its own cross-fold validation process to determine an optimal set of hyperparameters. This has the advantage

of reducing the chances of information leakage by allowing for hyperparameter optimisation that is not informed by exposure to independent test sets. While studies varied in the amount of data available for training, and consequently the number of folds they considered during cross-validation, evidence of repeat experiments greatly increases the reliability of reported performance metrics. This is of crucial importance for both clinical integration and scientific integrity. As underlined in [180], reproducibility is not guaranteed even when code is provided, making repeat experiments especially important. Twenty-five of the 55 considered papers did not employ repeat experiments, which reduces confidence in reported results. A number of repeat experiment studies reported point estimates, which does not fully convey the range of performance metrics, potentially leading the reader to believe the spread is not large. Code inaccessibility exacerbates this issue, leaving the reader unclear as to the procedure followed. We again found no significant difference between the accuracy metrics reported across repeat experiment procedures (repeat experiment studies=$87.944 \pm 10.43\%$, non-repeat experiment studies=$91.58 \pm 5.025\%$), although this does not minimise the importance of carrying out repeat experiments. It should also be noted that the experimental designs of each study differ, making a fair statistical comparison difficult. We recommend that researchers continue to employ repeat experiments and report their results with means and standard deviations.

2.4.3 Code availability

Most studies did not provide code. As detailed in [181], the principles of fairness, accountability, and transparency studies are of paramount importance for deep learning modelling studies, and code inaccessibility acts as an impediment to these ideals. The construction of deep learning systems requires many algorithmic decisions which can influence performance, introduce bias, and impact reproducibility. Deep learning models optimise an objective function over a set of arguments, meaning that any decisions taken in preprocessing and model construction can affect the capabilities of the system as a whole, and propagate subjective choices throughout ostensibly objective models [133]. For instance, several studies have examined algorithmic biases against underrepresented and/or marginalised groups [182–184]. Aside from domain-specific benefits to code sharing, the larger scientific community has recently shifted towards open science frameworks, with several high-profile journals requiring methodological transparency [129, 185–187]. Therefore, code availability and transparent methodological descriptions are an important aspect of deep learning experiments in this domain independent of potential clinical applications. Within a patient-care context, we underscore the importance of constructing reproducible systems to increase trust, both from a clinician and patient perspective. Studies making code available are proactively embracing these essential principles. We further encourage that minimal Jupyter/Google Colab notebooks, and

other literate programming tools, be explored to enhance understanding and reproducibility [129, 188, 189]. This would also have the useful properties of allowing researchers to examine pipelines and identify potential 'blind spots' that the model authors may have overlooked in their modelling decisions, encouraging accountability [181]. Additionally, model training is often computationally intensive; having access to models trained in similar domains could enable transfer learning approaches and allow researchers to examine different data representation strategies. Therefore, we recommend that authors share model weights and code to increase the potential of clinical translation ind facilitate reproducibility.

2.4.4 Saliency and explanatory efforts

We found many studies did not interrogate their presented models to ensure that relevant information is being used. Where irrelevant information is included, such as skull thickness when examining Alzheimer's disease neurodegeneration, without confirmation that the model is utilising brain information, attempts at patient care integration will have limited success. Even in cases where known irrelevant information can be removed by preprocessing, visual maps can draw attention to previously-unknown irrelevant information. Models that are opaque are less likely to be implemented in patient care settings, making the use of explanatory efforts of crucial importance. As previously stated, algorithmic biases in predictive settings is concerning, and saliency methods can help researchers to identify sources of bias where they occur. Additionally, attempting to understand the image features driving model predictions can help to relate new models to previous findings. These methods may also be used to generate new hypotheses for downstream experiments. Seventeen studies investigated neuroanatomical features driving model predictions via explanatory methods, thus increasing the potential to highlight sources of bias in model training. While there exists variability in application, especially in terms of region understanding, it is recommended that future studies apply explanatory methods to highlight relevant brain regions being used. Potential clinical utility can be further increased by supplementing these methods with code.
However, explanatory methods have a number of considerations that may limit their utility, prompting a discussion around how best to understand opaque models. Most existing methods deriving a saliency map return an 'importance' per pixel, which has no direct link to human-interpretable neuroanatomy. Usually, this represents the degree of change in the output relative to a small perturbation in the input pixel, collapsing a potentially non-linear relationship to single values. While it provides an empirical assessment of captured patterns and is a useful visual aid, it offers little interpretative value compared to coefficients returned by classical statistics. The deep learning field in general is focused on prediction as opposed to inference, meaning that the mechanistic understanding of relationship dynamics is often secondary to test accuracy. This is challenging in the context of discovery and clinical settings. Furthermore, saliency

methods have their own limitations arising from their algorithmic derivation of 'importance', which can affect interpretation [98]. Similarly, counterfactuals, while promising, are difficult to empiricize and require significant computational overhead. Nonetheless, interpretability efforts allow researchers to visually evaluate model attention, which, for clinical translation, can serve to increase confidence and reduce bias, a topic of concern with respect to models applied to society at large [133, 183].

2.4.5 Accuracy metrics, sample sizes, and data sources

We found that studies overall reported high predictive accuracies in their primary modelling questions ($89.36 \pm 8.694\%$). This underscores the potential of deep learning models to aid clinical procedures by leveraging a non-invasive data source. This makes the careful consideration of our outlined principles all the more important, given the capability of deep learning models to accurately classify brain disorders. We note that despite no significant differences in reported accuracies across questionnaire categories, the significance of applying these principles from a qualitative framework remains undiminished, especially when considering the information used by models to achieve reported accuracies; further, studies with high accuracies that have applied repeat experiments and thought carefully about data representation strategies can be considered more trustworthy. This trust can be further enhanced by making code available so results can be reproduced, with the additional benefit of allowing researchers to apply models to their own domains. A significant barrier to full reproducibility in this context is data privacy concerns, which limit the release of a fully reproducible paper.

Further, sample sizes were generally high, although there was large variation from study to study. While there was no significant relationship between sample size and accuracy, it appears that there is a weak negative correlation between the two variables, even in spite of large database crossover between studies, with 44 studies making use of ADNI ($R^2 = 0.055, p = 0.09$). However, the variation in experimental designs makes it difficult to draw conclusions based on this result.

2.4.6 Future perspectives and commentary

This systematic literature review highlights areas of focus across modelling practices, transparency, and explanatory efforts in the context of maximising the potential for clinical utility and reproducibility. These points underscore long-standing differences between deep learning and classical statistics, whereby the former is usually concerned with predictive performance and the latter with making inferential statements. The predictive imperative has led to numerous advances in image processing, with several state-of-the-art approaches developed to address tasks not suited to classical statistics [95, 117]. Neural networks have clear

advantages where describing data-generating parameters are not a concern, and the reported accuracies of the considered research is further evidence of this.
However, as deep learning becomes more readily applied to medical imaging domains, with potential consequences for patients, dichotomies of prediction versus inference should be retired, even where models have clear discriminative potential. Researchers can maximise potential clinical benefit and potentially increase the quality of patient care by embracing the principles of reproducibility, transparency, and explanation for predictive models. This can increase the confidence in such methods and consequently increase the potential for future clinical integration. We summarise our key recommendations in Table 3.

2.5 Limitations

This work reviewed studies from 2 database sources, but is not guaranteed to have evaluated all available relevant research. This study also did consider studies making use of functional neuroimaging data sources, which comprises a large corpus of research. We did not endeavour to comprehensively identify potential information leakage – an in-depth consideration of this concept is explored in [181]. Additionally, while we encourage the use of interpretability methods, we acknowledge the multiple drawbacks which may limit their utility and application. We further note that it is difficult to identify a unified set of optimal experimental parameters across every context – our commentary is designed to draw attention to the limitations arising from specific procedures and encourage researchers to carry out experiments that mitigate these issues as much as possible. We also note that fair comparison of reported accuracies across a myriad of diverse studies is challenging. This is particularly relevant given that there may be class imbalances between studies and different phenotypes considered. Although many studies carried out experiments in the ADNI dataset, there is no guarantee the same patient population or preprocessing efforts were carried out. This work did also not endavour to evaluate the variation in all preprocessing procedures of the selected studies.
Finally, our analysis using survey results is not reflective of overall study quality. Our binary questions are intended to serve as a vehicle to discuss important concepts that all CNN-based predictive studies should be cognizant of.

2.6 Conclusion

In summation, we conducted a systematic literature review of 55 studies carrying out CNN-based predictive modelling of brain disorders using structural brain imaging data and evaluated them in the context

Table 2.3: Key recommendations arising from the results of this systematic literature review, their benefits, and the risks associated with non-adherence.

Key Recommendations	**Benefits**	**Risk(s) mitigated**
Make well-annotated code freely available	- Improve chances of reproducibility – Readers can better understand workflow – Encourage accountability and transparency	- Limit reproducibility efforts – Models remain opaque
Employ repeat experiments	- Improve confidence in model estimation – Mitigate random weight initialisation	- Risk reporting overfitted results – Performance estimation inflation – Diminished confidence in system overall
Employ interpretability methods	- Validate that model is using relevant information – Potential biomarker discovery – Improve confidence in system overall	- Models remain opaque – Diminished confidence in system overall – Unsure what information is being used by models
Consider specifics of data representation	– Lessen chance of information leakage – Ensure optimal voting strategies – Cognizant of strengths and weaknesses of different representations	- Information leakage is more likely to occur – May not consider multiple voting strategies – May not consider optimal representation given experimental context

of their modelling practices, transparency, and interpretability. We set forth recommendations that we believe will increase the future potential clinical value of deep learning systems in this domain. Careful consideration of these concepts can help to inform a clinical framework that can effectively incorporate deep learning into diagnostic and prognostic systems, enhancing our ability to improve patient care.

2.7 Declaration of Competing Interest

All authors report no competing interests.

2.8 Acknowledgements

This work was conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6214.

2.9 Data availability

All studies in this systematic literature review are accessible via PubMed and Web of Science.

Chapter 3

Genetic Association Studies of Convolutional Neural Network-Derived Intermediate Phenotypes of Brain Disorders

.

3.1 Introduction

As detailed in our systematic literature review, the application of convolutional neural networks (CNNs) to neuroimaging data of brain disorders can yield high predictive accuracies in testing datasets. This is of great interest given the lack of reliable biomarkers available for a range of brain disorders, which could prove useful for diagnostic/prognostic applications. However, brain disorder biomarkers are also of great interest in the context of elucidating disorder pathology and potential underlying biological mechanisms. For example, the diagnostic criteria of bipolar disorder and schizophrenia are informed by differences in cognitive systems, but identifying neuroanatomical correlates of those phenotypes has remained difficult [2]. Further, quantifiable biomarkers are used primarily to provide context in the presence of other behavioural or cognitive features in clinical diagnostic frameworks, including for Alzheimer's disease (AD), where neuroanatomical differences are commonly observed in cases relative to controls [59, 77]. Nonetheless, identifying the biological manifestations of distinct psychiatric and neurological disorders may yield insight into the biological means by which behavioural or cognitive features used in clinical

practice may come about. In considering our previous review, the application of CNNs to brain disorders to identify biomarkers is limited by several factors. Firstly, as a general rule, it is difficult to establish causality from observational studies of any kind. This means that a feature associated with a phenotype by way of its contribution to an accurate prediction may not be informative for understanding any systems related to the disorder. Secondly, any confounding variables we can measure cannot be controlled for in the traditional statistical sense in deep learning models; this is because regression models are designed to facilitate inferential statements about the data generating process, whereas deep learning models are usually designed to maximise predictive performance by leveraging relationships of any nature in the data. This means that if we can identify input features deemed important by a predictive neural network, this 'importance' must be cautiously interpreted. Thirdly, neural networks and related models are designed with primarily predictive tasks in mind, making it difficult to identify features important for classification in the first instance. While methods exist to measure the importance of image features to CNNs, they do not have the same interpretation as classical regression coefficients [95].

Despite these issues, the predictive capabilities of CNNs applied to brain disorders incentivises efforts to understand their data representations. In identifying features important to classification accuracy in an interpretable fashion, we have the potential to better understand brain disorder pathology as it relates to differences in structural neuroimaging. Furthermore, if these features can be understood and represented as phenotypes, we can query the genetic basis of a potentially novel biomarker associated with a particular disorder. This is advantageous from both a confirmatory and exploratory perspective, by allowing us to test whether or not the molecular correlates of our biomarkers make sense in the context of the disorder in question.

To this end, we sought to train a CNN to classify Alzheimer's disease using neuroimaging data and carry out a genome wide association study (GWAS) on the features deemed important by the model. The remainder of this chapter describes the neuroanatomical features and genetic architecture of Alzheimer's disease, interpretability in CNN models, and subsequently our methods, results, and discussion.

3.1.1 Alzheimer's disease

Alzheimer's disease (AD) is a debilitating progressive neurodegenerative condition characterised by impaired cognitive function and loss of memory [190]. It is especially prevalent in elderly populations [191]. Due to increased global life expectancy, its incidence, and subsequent disease burden, has increased significantly over the past 3 decades [192]. As such, it has been the focus of research efforts since its description [193, 194]. Its etiology is multi-faceted, with a range of social, environmental, and genetic risk factors contributing to its progression [195]. As medical imaging became widespread, it was observed that brain

region atrophy was frequently observed in individuals with AD [196, 197]. As previously mentioned, access to neuroimaging consortia such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) and large sample sizes for GWAS yielded several biological correlates of AD [35, 53, 60, 61]. Early genetic linkage studies implicated mutations in the $\epsilon 4$ allele of the apolipoprotein-E (ApoE) gene as strong genetic predictors of late onset AD, a result verified by later GWAS [35, 198]. The ApoE4 isoform results in impeded clearance of amyloid-β, thus contributing to its increased toxic accumulation in neural tissues, leading to subsequent neurodegeneration [199]. Its characterisation spawned several theories on AD molecular pathology. Accumulation of neurofibrillary tangles and β-amyloid plaques in the brain are thought to be potential causal factors, leading to impaired neurotransmitter function and potential neurotoxicity [200–202]. Additionally, atrophy in several brain regions – such as temporal lobe structures thought to be responsible for memory formation – has been repeatedly described in AD participants [203, 204]. Its clinical diagnosis is still primarily reliant on behavioural features despite the identification of several neuroimaging correlates [190]. As previously noted, neuroanatomical endophenotypes of AD include medial temporal lobe and hippocampal atrophy [71].

Recent predictive studies of AD using neuroimaging data have demonstrated high predictive accuracies [181]. For example, our systematic literature review found that the mean predictive accuracy of 57 studies classifying brain disorders was $89.53 \pm 8.69\%$, with the majority of studies using AD or cognitive impairment as outcomes. This suggests that CNNs have the potential to classify AD with high accuracy. This also demonstrates the potential of using such models to describe novel endophenotypes of AD. However, the use of deep learning approaches for endophenotype derivation in AD has been scarcely explored. In [205], the authors train a classifier to identify transcriptomic signals associated with AD that may serve as candidate endophenotypes. No studies have attempted a similar principle using neuroimaging data. This may be due to interpretability issues in CNN models.

3.1.2 Interpretability of CNNs

CNNs are often high-dimensional models creating abstract representations of input data [117]. Since their introduction, researchers have been concerned with means by which to understand their internal representations. This could be accomplished by examining the makeup of feature maps while models did not have millions of parameters [206]. More recently, models have increased exponentially in size, making this task more difficult [95]. As such, researchers considered alternate strategies, including class activation maps [207]. These methods offered immediate visual explanatory insight by overlaying 'importance-ranked' sum-weighted feature maps over input images. While useful, its implementation required that a specific layer be included before prediction. This lead to the introduction of a generalisable variant

termed gradient-based class activation mapping (Grad-CAM)[96]. Such methods could be used with pre-trained models and required only obtaining the gradient of output with respect to the input. Further, these approaches were related to other gradient-based methods that overlayed gradients directly onto input images [95]. These approaches were utilised for identifying features important for several classification tasks, including tuberculosis screening from X-ray data [208] and skin lesions [209]. Further, Grad-CAM has been used in CNN models of multiple sclerosis [210], lung cancer [211], and coronavirus [212] to identify relevant variables. Recently, the use of counterfactuals has also been explored for brain imaging classifiers [213].

Given the predictive potential of CNNs applied to neuroimaging collections and the recent advances in interpretability methods, it would be of interest to attempt to derive understandable endophenotypes of AD using brain structural information.

3.2 Methods

3.2.1 Phenotype representation

An attractive potential approach to endophenotype derivation from imaging data is Grad-CAM owing to its widespread use and properties. However, in order to use endophenotypes in GWAS, they must be converted to empirical phenotypes as opposed to image data types. This poses issues as the output of Grad-CAM is a heatmap. The use of bounding boxes may be considered, whereby a summary of the pixel values serve as the phenotype per participant. However, it would be difficult to ensure that the same brain information is considered per individual owing to subtle differences in image registration. Additionally, the final step of Grad-CAM usually involves resizing the output, which may cause information loss when the final convolutional layer is small.

Alternatively, we can consider a variant of the class activation mapping approach to output a phenotype capturing the internal model representation. This can be accomplished through a global average pooling layer following the final convolutional operation, the outputs of which are sum-weighted to give an output prediction. This can be expressed as:

$$\hat{y} = \alpha \left(\sum_{k=1}^{k} w_k \bar{A}_k \right), \tag{3.1}$$

where $\bar{A}_k$ represents the average activation of feature map A indexed by k at the final convolutional layer, w is its associated weights vector, α is a sigmoidal transformation that bounds the output between zero

and one, and $\hat{y}$ is the output prediction. Equation 3.1 is very similar to equation **??**, and possesses some of the same interpretable properties. For instance, w_k represents the contribution of the average of feature map A_k to the final prediction, affording us the opportunity to rank the most significant feature map of the set k based on its value in w. This is useful for two reasons: firstly, it allows a single number that is easily computed to be used as the phenotype in a GWAS; secondly, we can calculate the gradient of this quantity with respect to the input pixel values to identify brain regions contributing to its value. This is similar to standard class activation mapping in architecture, whereby the global average pooling output acts as the final phenotype. We can also correlate this quantity with volumetric brain information variation to better understand its properties.

3.2.2 Data demographics

We trained a custom CNN model on skull-stripped and registered MRI data from 423 participants from the ADNI consortium, comprising 196 AD participants and 227 control individuals. In order to reduce the amount of noise in the classification space, we only considered participants with an AD diagnosis or a cognitively normal (CN) phenotype during training, excluding an intermediate subset of participants with mild cognitive impairment. This is because our primary modelling objective is to train a classifier that can distinguish clinically confirmed AD from CN participants in a binary fashion – greater sample sizes may be required to build an efficient three-class model.

Figure 3.1 shows the demographic breakdown of the training and testing split, which contained 284 and 139 participants respectively. Linear models of age regressed against partition status ($P = 0.967$) and sex ($P = 0.521$) revealed no significant associations, nor did logistic regressions of sex against partition status ($P = 0.779$) and age ($P = 0.519$), suggesting no statistically significant differences between demographic characteristics in the training and test sets. In order to relate our results back to human-interpretable concepts, we also obtained Freesurfer-processed tabular structural neuroimaging data from the 423 aforementioned participants, representing 249 neuroanatomical measurements per participant. This was obtained from the *aparc* and *aseg* output files provided on the ADNI website resulting from the *recon-all* command applied using FreeSurfer 4.3.

3.2.3 Data splitting

Owing to the large input sample size, we endeavoured to reduce computational overhead while still preserving important information per participant. We can preserve the majority of information per participant by leveraging the fact that there are usually many zeros in MRI volumes, and as such, a large amount of

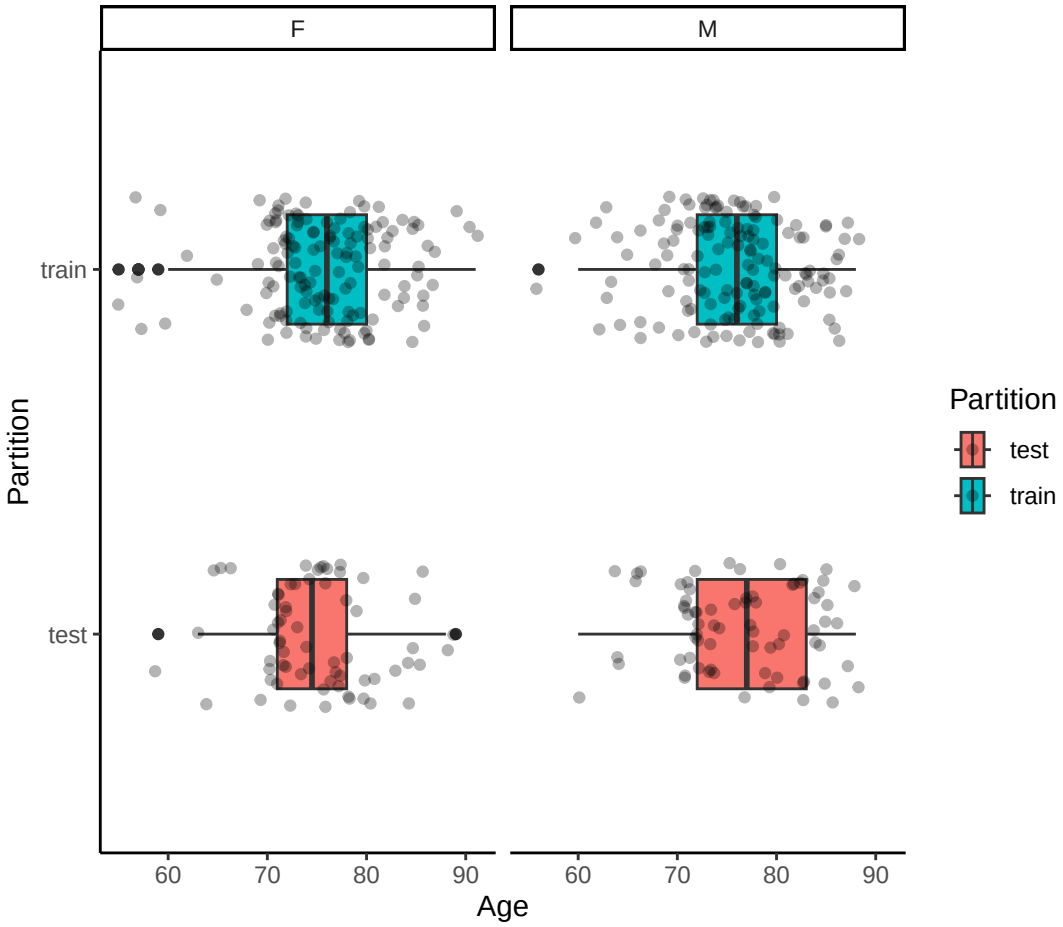


Figure 3.1: Demographic information for 423 participants based on their training/test partition, sex, and age. We found no statistically significant differences between age or sex across partitions using linear and logistic regression models respectively.

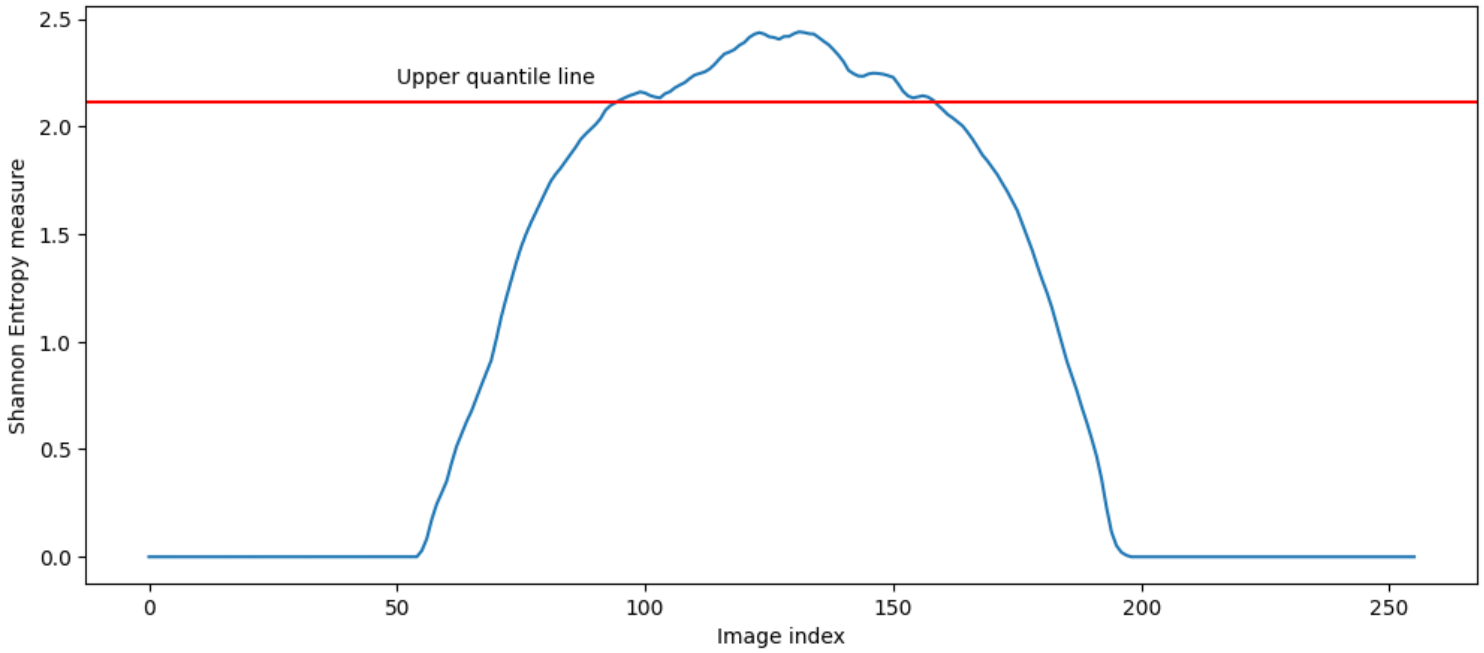


Figure 3.2: The distribution of S values on the y-axis for every slice of a sample participant (x-axis). Here, the Shannon entropy value S is displayed on the y-axis and the slice index is displayed on the x-axis. The upper quantile of the S distribution is denoted in red, and is highest towards the middle indices of the participant's image, indicating that there is more information in bits contained in the corresponding slices.

the input volumes can be discarded. We can formalise this by measuring the Shannon entropy (S) of every image slice along a given axis, denoted by:

$$S = -\sum_{x} (p(x) log_2 p(x)) \,, \tag{3.2}$$

where $p(x)$ represents the proportion of pixels of that value as a fraction of all pixels (ie., how many pixels out of a 100×100 image take on the value i) for every pixel value indexed by x. We can then retain slices above an individual-specific information threshold, resulting in a set number of informative image slices per participant (Figure 3.2).

We found that this methodology results in the same number of slices with S values above the upper quantile per participant, resulting in 64 slices per participant representing the most informative parts of the 3D volume along the given axis. This ensures that the same information is being considered per participant across the 64 slices.

3.2.4 Volume splitting

We split our MRI volumes of size 256^3 into participant-level collections of 2D slices of size $64 \times 256 \times 256$ pixels according to our S distribution. This operation was performed independently in training and test sets after training-test set partition to prevent information leakage. We further split the training set into separate validation sets at the subject level to monitor performance, whereby 15% of the training data subjects were held out for independent testing during training. Finally, we concatenated every slice into a dataset of size $241 \times (64 \times 256 \times 256)$ for the training set, $42 \times (64 \times 256 \times 256)$ for the validation set, and $140 \times (64 \times 256 \times 256)$ for the independent testing set respectively. We also implemented a custom image augmentation algorithm which added 3000 extra images to the training set that underwent a random amount of rotation, resulting in a $\approx 19\%$ increase in training dataset size. Image augmentation techniques seek to increase training dataset size by re-purposing original images using rotation or cropping procedures. This has the advantage of potentially increasing generalisability performance owing to the fact that augmented images are essentially noisy analogues of real inputs [214].

3.2.5 Model construction and evaluation

We constructed a model according to the schematic in Figure 3.3. Our final two layers promote a sparse data representation, with a global average pooling layer fed directly to a prediction layer. We implemented a rectified linear unit transformation as the activation function at every convolutional layer and a sigmoid transformation at the last layer. We used the adaptive moment estimation optimiser (*Adam*) with a learning rate of $3e{-}4$ during training [215]. This optimisation algorithm applies an adaptive learning rate based on the severity of previous weight updates, demonstrating stable convergence solutions in a range of CNN training tasks [215].

We measured the accuracy and area under the receiver operating characteristic curve (AUC) in the test population at the slice level and averaged the predictions per participant.

3.2.6 Phenotype extraction

After training and evaluation, we ranked the final weights vector to find the feature map with the greatest effect on the output. Because the activation function used throughout the network is a rectified linear unit ($max(0, x)$), every entry was positive, making our ranking task trivial. We extracted this output feature map average for every slice in both train and test datasets and averaged the score per participant. To boost our sample size, we applied the trained model to an additional set of mild cognitive impairment participants to increase the power for a GWAS, yielding 744 participant feature map scores. Mild cognitive

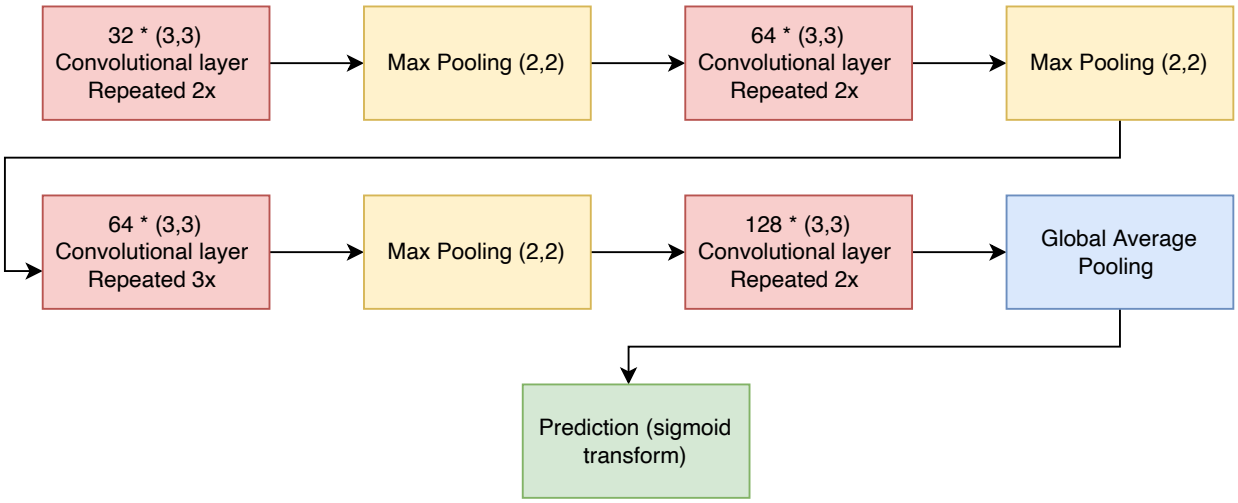


Figure 3.3: Schematic of our CNN model architecture. Here, we denote the individual operations in boxes; where we describe repeating procedures, the output of the first pass of the operation is used as the input to the second. For example, in our first layer, the output of the 32 feature maps populated during the first convolutional operation is passed through subsequent convolutional operations that populate another 32 feature maps. In brackets, we denote the window size of the associated operation – for instance, (3,3) in a convolutional layer describes a convolutional filter with a dimensionality of 3×3; (2,2) in a max pooling layer indicates that the maximum number in a 2×2 window is returned as the output of the operation.

impairment is an intermediate state between a cognitive control phenotype and AD. We excluded these samples from our predictive model training to ensure that the signal considered for our CNN was based entirely on clinically defined AD; however, we included them in our GWAS step, expecting that their feature map output scores would fall between control and AD participants.

3.2.7 GWAS

We imputed genetic information with *beagle* using the 1000 Genomes project haplotype maps as a reference [216, 217]. We constructed a genetic relationship matrix using a set of 556,516 HAPMAP3 variants pruned with the following parameters: r^2 greater than 0.9 in a 100 variant sliding window of size 1000 and minor allele frequency > 0.01. We calculated principal components using linkage disequilibrium-pruned variants with $r^2 > 0.05$ in a sliding window of 1000 bp width and 50 bp step size. We validated that ancestry effects were captured by principal components (Supplementary materials). We then used *fastGWA* to fit a linear mixed model to the imputed variants, covarying for age, site, and sex as fixed effects, and the top 10 ancestry principal components as random effects [218]. We further fit a separate GWAS following the same procedure restricted to individuals of white ethnicity only. Further, we carried out a power analysis to understand what genetic effects we were capable of detecting. Aside from imputation, we followed the pipeline described in [219] to preprocess genotype data and carry out the GWAS.

3.2.8 Interpretation and visualisation

We applied the full suite of *FUMA* tools to our results to investigate the tissue enrichment profile of our results and single nucleotide polymorphisms (SNPs) in linkage disequilibrium with significant variants [220]. We further defined a gradient map of score values using the approach presented in **??**, whereby the expected change in score value given a change in input pixel value was overlaid on its respective input image. We prioritised the image slices with the largest feature map score per participant for visualisation to identify global trends driving score activity. Further, we calculated the top 40 principal components for any participants that had tabular data available (533/744) to capture as many axes of variation as possible. We then regressed 40 principal components of structural neuroimaging features for a subset of the participants (533) against our CNN score and held age and sex as fixed covariates. Specifically, we fit the following equation:

$$\bar{A}_k = \beta_0 PC_i + \beta_1 age + \beta_2 sex + \epsilon \tag{3.3}$$

whereby $\bar{A}_k$ is the averaged feature map with the largest contribution to the final layer, PC_i is a principal component indexed by i, β_{age} is the effect of age, β_{sex} is the effect of sex, and ϵ is random noise. We

seek to estimate β_0, which represents the magnitude change expected in $\bar{A}_k$ given a unit increase in PC_i. We accounted for multiple hypothesis testing of many principal components by setting a Bonferroni-corrected p-value threshold of $0.05 \div 40$. The resultant significant components were visualised and their eigenvectors were examined to identify the structural variables contributing most to the axes of direction correlated with our activation score. This is motivated by a desire to relate the variation in our output score to human-interpretable neuroanatomical features.

Finally, we regressed our CNN-derived score against every structural variable to obtain estimates of mediated coefficient effects. These quantities represent the expected change in CNN-derived feature map score given a unit increase in the measured variable holding constant every other variable. We added age and sex covariates as terms in our regression model.

3.3 Results

3.3.1 CNN performance and phenotype extraction

We obtained participant-level accuracies of $82\%, 80\%$, and 79% on unseen test data across 3 repeat experiments for our proposed model. Figure 3.4 shows the ROC curve of the highest performing model and its AUC (0.84). We extracted feature map average scores from the feature map 47 of this model, which had the highest weight connection at the final layer. We aggregated global average pooling outputs for this feature map per participant and plotted them according to our three phenotype categories. Figure 3.5 shows the standardised feature map 47 score, which appears to be stratify participants continuously as cognitive impairment increases.

3.3.2 GWAS results

Our GWAS of feature map 47 scores yielded 5 genomic risk loci on chromosome 2 (Table 3.1) and 7 significant independent lead SNPs on the same chromosome ($P < 5e-8$). The overall estimated heritability was $9.8e-16$ using the method presented in [221] given by estimating the phenotypic variance explained by additive genetic effects using a genetic relationship matrix. Fourteen genes were mapped based on the associated genomic risk loci (Table 3.2), and Figure 3.6 displays the Manhattan plot for the associated GWAS results, with the gene closest to the most significant SNP (*GPR75-ASB3*) annotated. One genomic risk locus (rs2542584, $\beta = -1.9$) is an intronic variant in high LD with 160 other SNPs ($r^2 \geq 0.6$, Table 3.1) and is mapped to eight of our fourteen mapped genes (*GPR75-ASB3, ASB3, CHAC2, ERLEC1, GPR75, PSME4, ACYP2*, and *SPTBN1*). Its raw CADD score, a measure of a variants likely

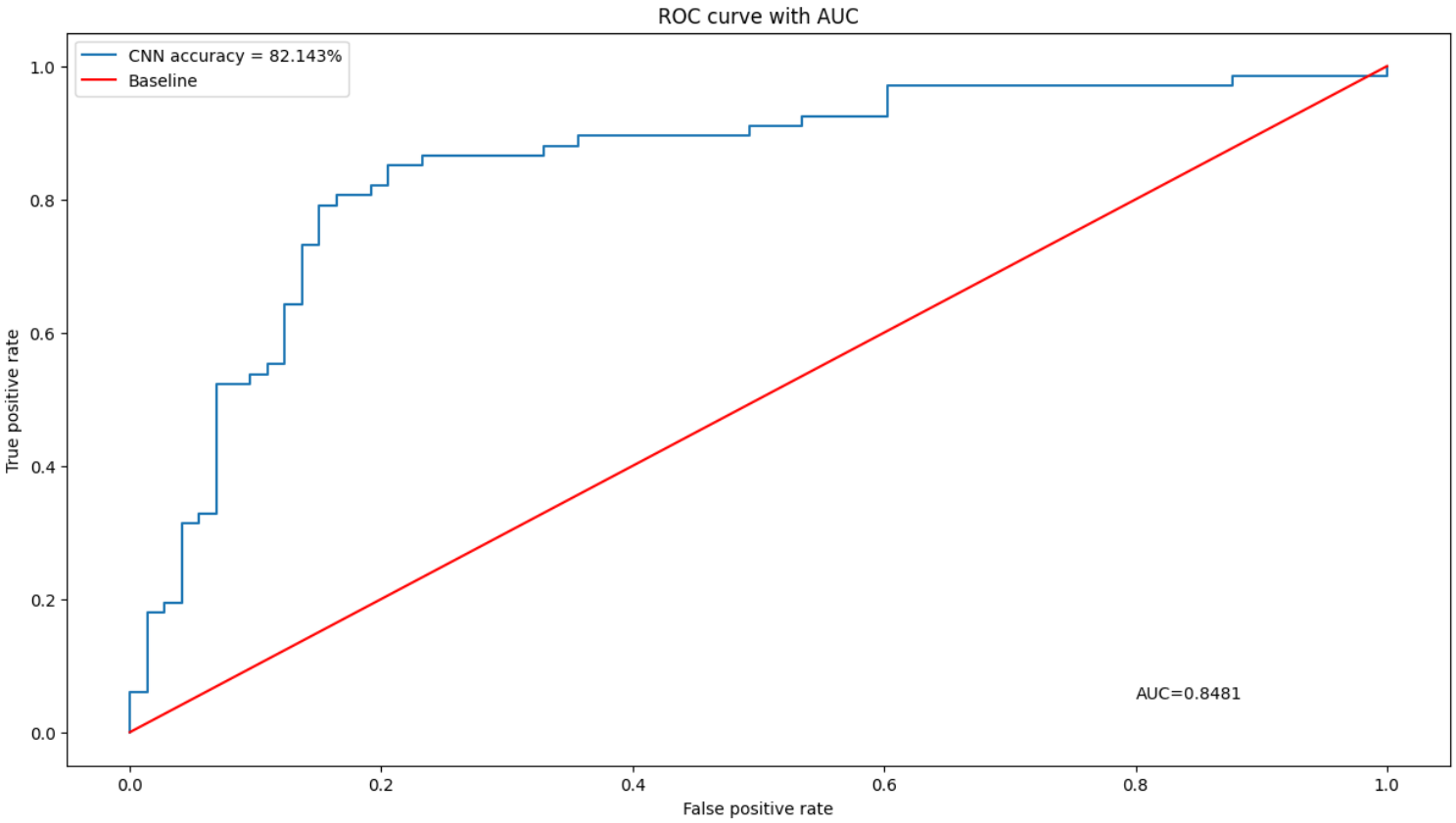


Figure 3.4: ROC curve and AUC for the highest performing model, with the x-axis displaying the false positive rate and the y-axis displaying the true positive rate. The red line denotes the relationship between these quantities in a random-guess scenario.

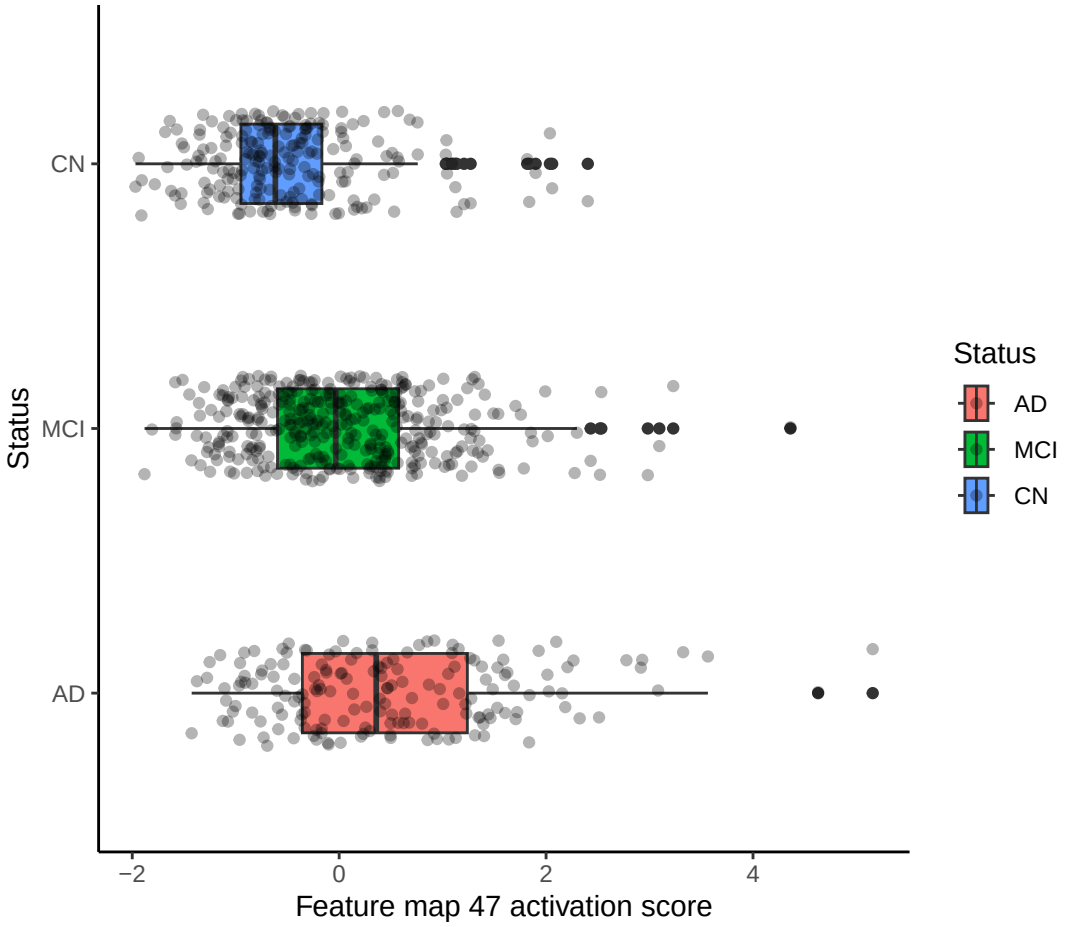


Figure 3.5: Standardised participant-level feature map 47 average scores across phenotypic designations. The feature map 47 score per participant is visualised on the x-axis and the y-axis describe group categories. The pairwise p-values between groups were not computed as the feature map in question was chosen using a weight ranking system, meaning that the statistical differences between groupings are not necessarily relevant to subsequent experiments.

deleteriousness, is 5.94, with a normalised value of 0.32 (given by dividing every raw value by the maximum from annotated SNPs). Another intronic genomic risk locus (rs1318672, $\beta = 1.54$) was in LD with 157 other SNPs, including an exonic nonsysnonymous single nucleotide polymorphism not tagged in our dataset (rs3762513, $r^2 = 0.65$, Figure 3.7). However, the exonic variant's CADD score is lower (0.173) than that of our intronic genomic risk locus (raw = 5.512, scaled = 0.29). Further, it was mapped to three genes of our fourteen given by *MAGMA* (*CCDC104*,*SMEK2*, and *PNPT1*). We found that a variant in LD with the intronic rs10194108 variant ($\beta = 1.5$) from our genomic risk loci list located in the *SPTBN1* gene had a high predicted CADD score despite being intronic (rs78496188, raw = 17.23, scaled = 0.92). Further, rs10194108 was also mapped to the *EML6* gene, which codes for a microtubulin transport protein. Both rs376336838 and rs141237032 ($\beta = 1.52$ and 1.49 respectively) were in LD with 2 and 1 SNP(s) respectively. Both SNPs were intronic, with rs141237032 mapping to a non-coding intronic region. Hence, only rs376336838 was mapped to protein-forming genes (*STON1* and *STON1-GTF2A1L*).

A collection of mapped genes were shown to have literature support for involvement with neurodegeneration phenotypes, proteasome function, and cellular homeostasis in the context of AD and other psychiatric conditions [222–224]. The majority of tagged SNPs (including those not present in the dataset but in LD ($r^2 \geq 0.6$) with independent significant SNPs) were estimated to be enriched for intronic, intergenic, and non-coding RNA intronic functional categories, with all categories reaching significance in a Fisher's exact test as estimated from *ANNOVAR* ($P = 1.9e{-}61, 1.6e{-}36, 1.7e{-}14$ respectively). This test also included linked SNPs.

Figure 3.8 suggested the nervous system was one of the primary areas of tissue-specific expression amongst the 14 mapped genes provided by FUMA according to average log_2 genotype-tissue expression (GTEx) results. However, this result was not significant in a hypergeometric test of enrichment ($P > 0.05$).

Figure 3.9 shows a SNP-wise mean test of p-values, using an aggregated p-value per gene as a test statistic as per MAGMA methodology [225]. Several genes on chromosome 2 reached significance, including *CCDC88A, CCDC104,* and *SMEK2*. Additionally, *ITGA1* and *TBC1D22A* were also found to be significant on chromosomes 5 and 22 respectively. *CCDC104* and *SMEK2* were also part of the set of mapped genes, with both genes containing variants in LD ($r^2 > 0.6$) with with CADD scores of ≈ 18.

3.3.3 Principal component correlation

We found that 5 out of the 40 principal components (PCs) of tabular neuroanatomical variation measures had regression terms with significant p-values after Bonferroni correction (Figure 3.10). We focused on the PC with the highest R^2 value, PC2, and examined its top 10 loadings, representing their contribution

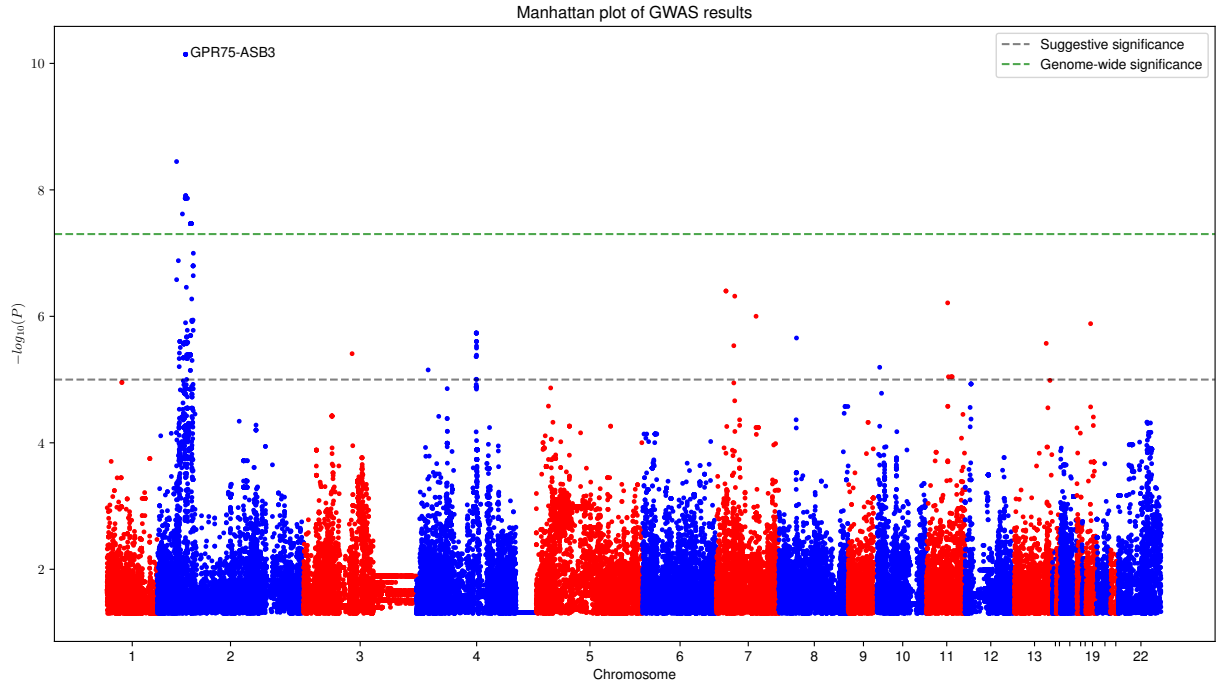


Figure 3.6: Manhattan plot of GWAS results. Suggestive significance threshold lines were drawn at $-log_{10}(1e-5)$, and genome-wide significance threshold lines were drawn at $-log10(5e-8)$. The closest mapped gene to the top independent SNP, *GPR75-ASB3*, is annotated.

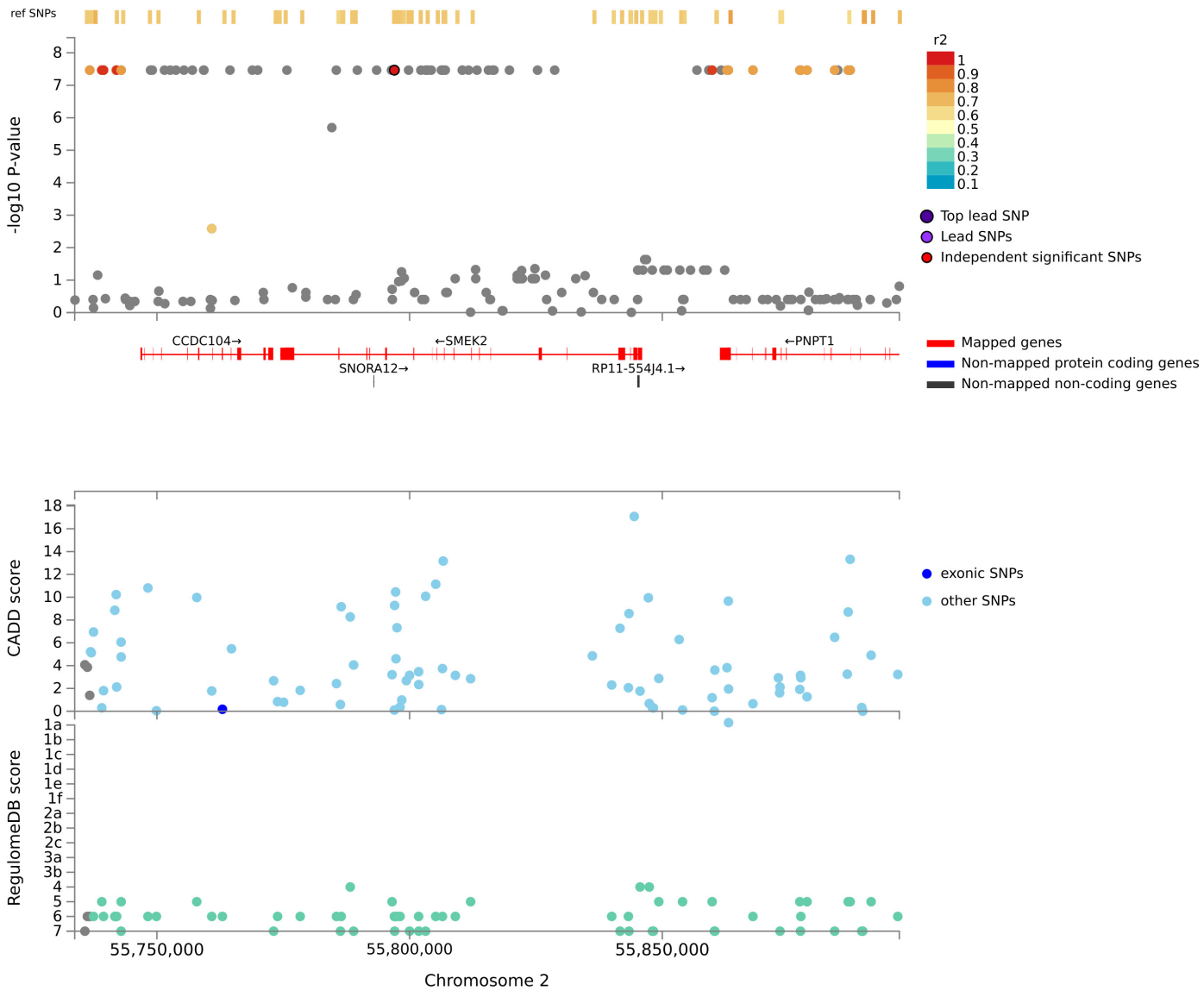


Figure 3.7: Regional association plot of gene *CCDC104* and its tagged variants both in and outside of the dataset. CADD score is a measurement of the predicted deleteriousness of a given variant, with higher scores indicating greater chance of deleterious effects based on computational and experimental evidence. The RegulomeDB score indicates the likelihood of a difference in binding activity of a gene given the variant in question according to functional assay experimental evidence, with larger scores indicating less evidence of binding disruption. The dark blue point on the CADD plot denotes rs3762513, a SNP in LD with rs13027853 ($r^2 = 0.65$, dark red point on upper plot), an exonic non-synonymous variant in the CCDC104 gene, with low CADD (0.173) and RegulomeDB (6) scores.

Table 3.1: Genomic risk loci identified by *FUMA*. Genomic risk loci are defined as aggregations of SNPs in LD ($r^2 \geq 0.6$) with independent significant SNPs identified by GWAS results. nSNPs denotes the number of SNPs in LD ($r^2 \geq 0.6$) with the given SNP that may or may not be in the GWAS dataset; nGWASSNPS denotes the number of SNPs in LD ($r^2 \geq 0.6$) with the given SNP that are present in the GWAS dataset; nIndSigSNPs denotes the number of independent significant SNPs ($r^2 < 0.6$) in the locus; nLeadSNPs denotes the number of lead SNPs ($r^2 < 0.1$) in the locus.

rsID	chr	pos	p	start	end	nSNPs	nGWASSNPs	nIndSigSNPs	nLeadSNPs
rs376336838	2	48803637	3.5e-09	48803637	48808594	2	1	1	1
rs141237032	2	52381261	2.4e-08	52381261	52381261	1	1	1	1
rs2542584	2	54009844	7.2e-11	53654561	54386344	160	29	3	3
rs10194108	2	54964021	1.3e-08	54736382	54969379	90	7	1	1
rs1318672	2	55768883	3.4e-08	55723643	55923831	157	64	3	1

Table 3.2: Mapped genes given by MAGMA as part of FUMA.

symbol	chr	start	end	strand	entrezID	minGwasP	IndSigSNPs
STON1	2	48756522	48826025	+ve	11037	3.5e-09	rs376336838
STON1-GTF2A1L	2	48757064	49003654	+ve	286749	3.5e-09	rs376336838
GPR75-ASB3	2	53759810	54087170	-ve	51130	7.1e-11	rs530198155;rs2542584
ASB3	2	53897430	54087297	-ve	100302652	7.1e-11	rs2542584
CHAC2	2	53994929	54002333	+ve	494143	7.1e-11	rs2542584
ERLEC1	2	54014181	54045956	+ve	27248	7.1e-11	rs2542584
GPR75	2	54080050	54087126	-ve	10936	NaN	rs2542584
PSME4	2	54091204	54197977	-ve	23198	NaN	rs2542584
ACYP2	2	54197975	54532437	+ve	98	4.9e-02	rs2542584
SPTBN1	2	54683422	54896812	+ve	6711	NaN	rs10194108
EML6	2	54950636	55199157	+ve	400954	1.3e-08	rs10194108
CCDC104	2	55746740	55773015	+ve	112942	3.4e-08	rs13027853;rs1318672
SMEK2	2	55774428	55846015	-ve	57223	3.4e-08	rs1318672;rs13027853;rs6737995
PNPT1	2	55861400	55921045	-ve	87178	3.4e-08	rs13027853;rs1318672;rs6737995

to the final PC2 value which correlated significantly with our CNN activation score (Table 3.3). The surface areas of both hemispheres had the highest loadings for this PC, and the total variance explained by this component was 37.2%. Other regions in the top 10 eigenvectors included the lateral ventricles, the choiroid plexus, and the precentral gyrus.

3.3.4 Gradient-based visualisation

We plotted the absolute gradient of the participant-average feature map 47 score to visually examine region importance according to our model (Figure 3.11). In this plot, the colorbar represents the expected output change (feature map activation average score) based on a perturbation in the input value of the pixel in question. We specifically focused on brain slices from AD participants with the highest model predictions - visual examination of these regions shows that areas of the parietal lobes and lateral ventricles appear to

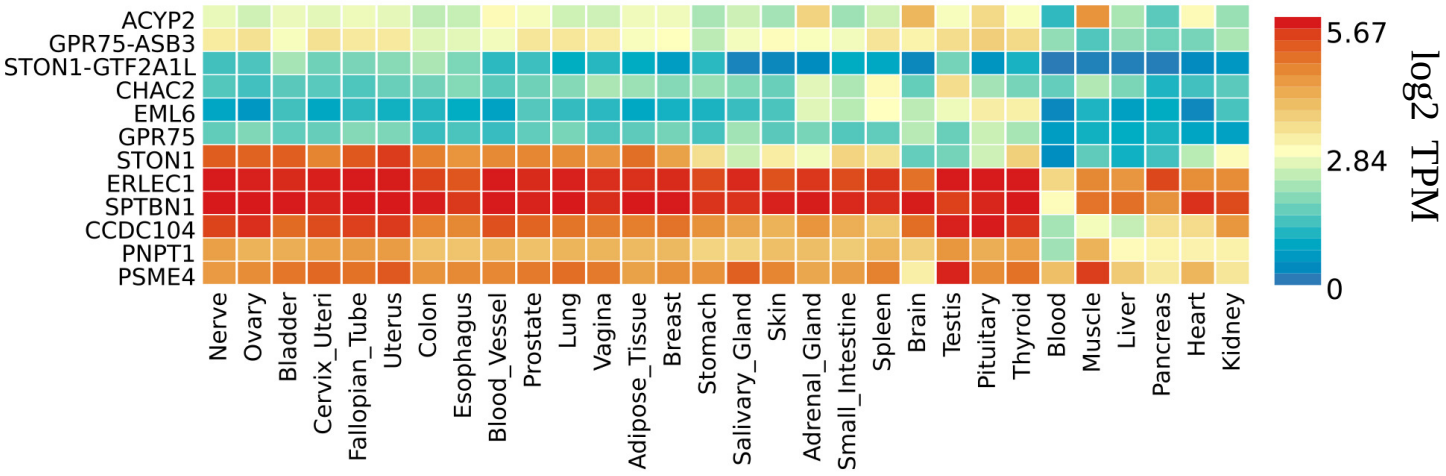


Figure 3.8: Heatmap of GtEX average log_2 transcripts per million (TPM) values across tissues for the 14 genes mapped from 5 genomic risk loci. Tissues are displayed on the x-axis and the associated genes are displayed on the y-axis, with colors corresponding to the magnitude of the average log_2 expression of that gene in the corresponding tissue.

have the greatest influence on feature map activation score. Additionally, we observe that gradient values are particularly high around the edges of brain tissue, suggesting that our score is sensitive to changes in boundary areas. This may suggest that overall brain size is an encoded feature owing to the fact that pixel value changes in these regions imply volumetric differences. However, the direction of effect is unclear as the gradient calculation does not presuppose a specific direction of perturbation, making it difficult to interpret these gradient maps as evidence for atrophy or volumetric increase in any region.

3.3.5 Regression of CNN score against structural variables

We found that 11 terms were significant in our regression of CNN score against brain structural variables holding constant age and sex. In Figure 3.12, we plot 2 variables with the highest R^2 values from the set of coefficients with test statistic p-values less than 0.05 (volume of left lateral ventricle and thickness of right inferior parietal, $R^2 = 0.1$ respectively). According to our scatter plots, high FMAP47 score is associated with cortical thinning in the inferior parietal and increased volume of the left lateral ventricle.

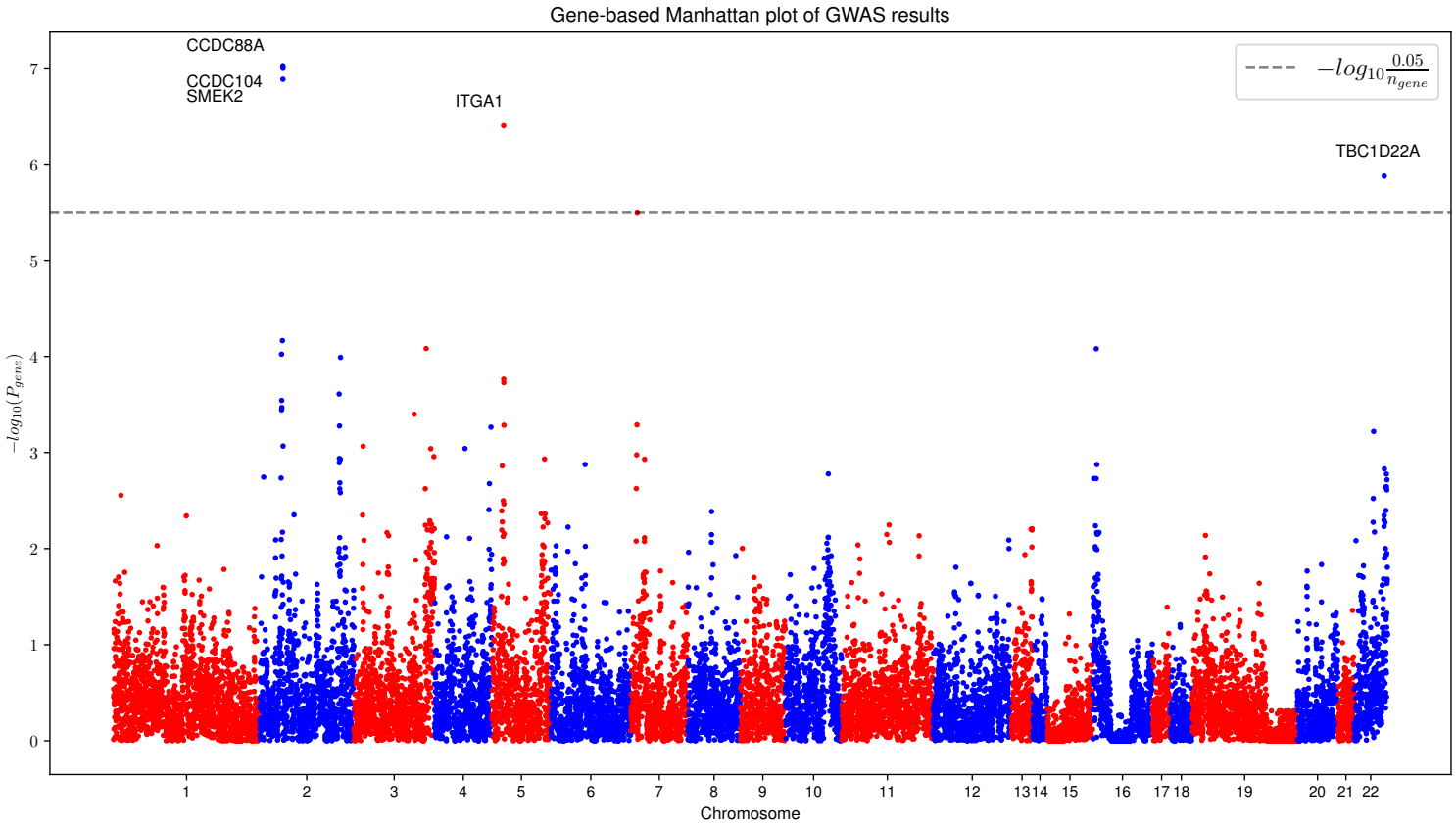


Figure 3.9: Manhattan plot of gene-based results from a SNP-wise mean model provided by MAGMA. Significance threshold is set as 0.05 divided by the number of protein coding genes with mappings from the summary statistics (15833).

Table 3.3: Table of top 10 highest region eigenvectors for PC2.

Region	Eigenvector Value
Surface Area of Right Hemisphere	0.094
Surface Area of Left Hemisphere	0.091
Volume of Left Inferior Lateral Ventricle	0.084
Volume of Right Choroid Plexus	0.083
Surface Area of Right Precentral	0.082
Volume of Third Ventricle	0.081
Volume of Right Lateral Ventricle	0.081
Volume of Left Choroid Plexus	0.080
Volume of Right Inferior Lateral Ventricle	0.080
Surface Area of Right Postcentral	0.080

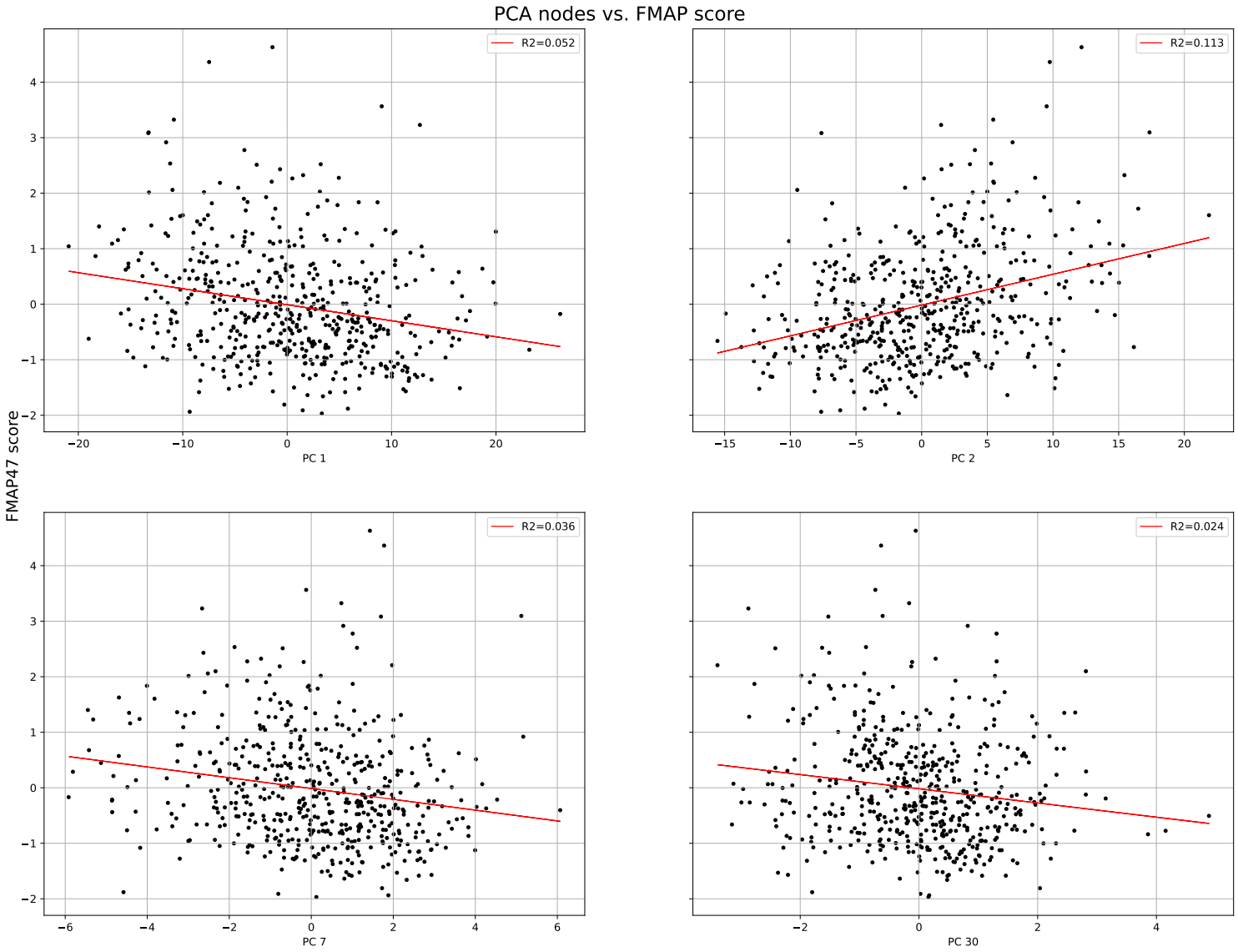


Figure 3.10: Principal component nodes surviving Bonferroni correction significantly correlated with our feature map activation score. PC2 (first row, second column) had the highest R^2 value (0.11) and explained 37.2% of the variation in the dataset.

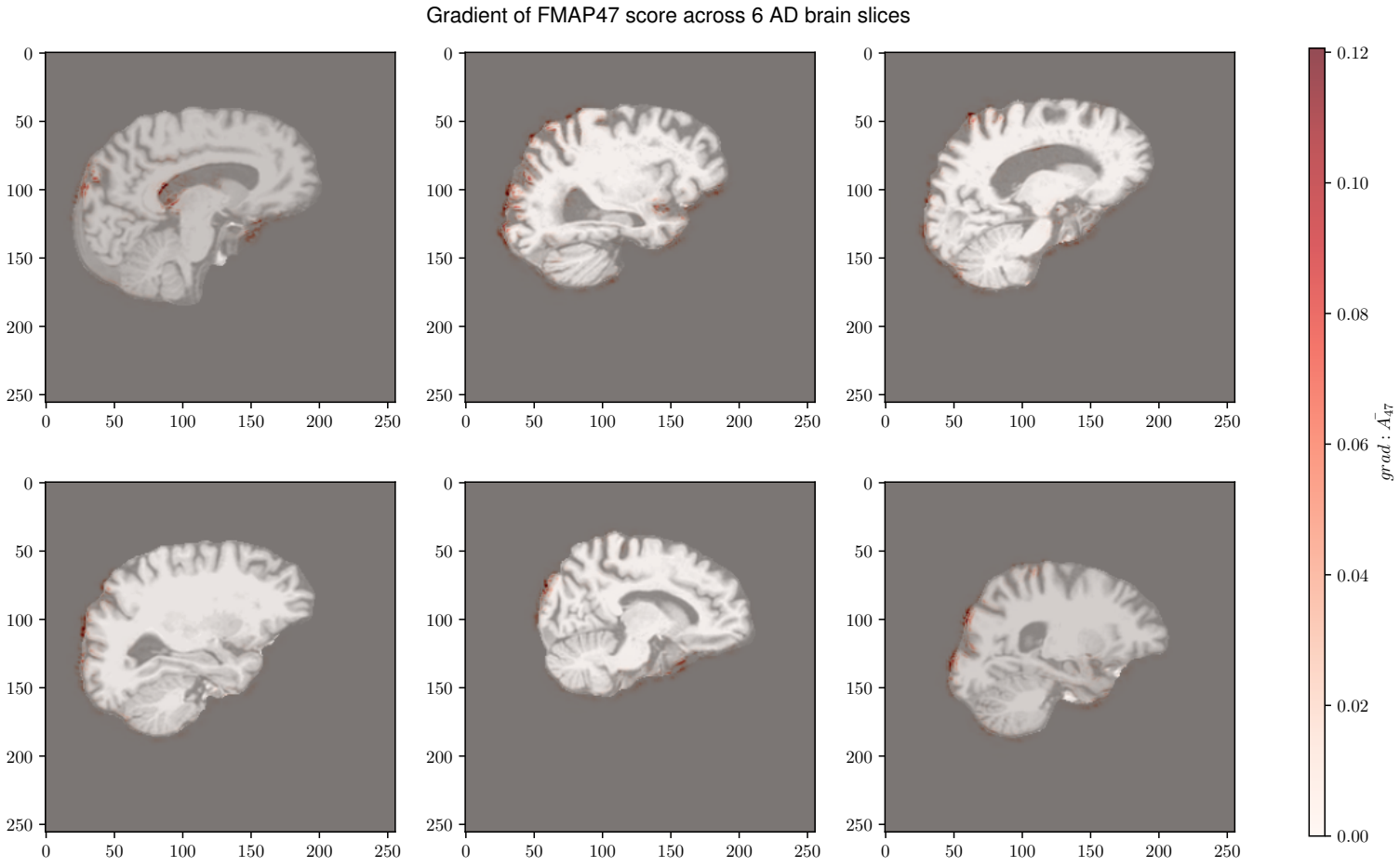


Figure 3.11: Maps of the absolute gradient of feature map 47 scores in 6 brain slices from 6 different participants with AD. The slices with the highest returned prediction for AD were chosen for visualisation. Here, the absolute gradient of the feature map 47 output score for that slice given a change in the input denotes the magnitude of the color intensity. Therefore, more intense color values correspond to greater feature map 47 score given pixel changes in those respective regions.

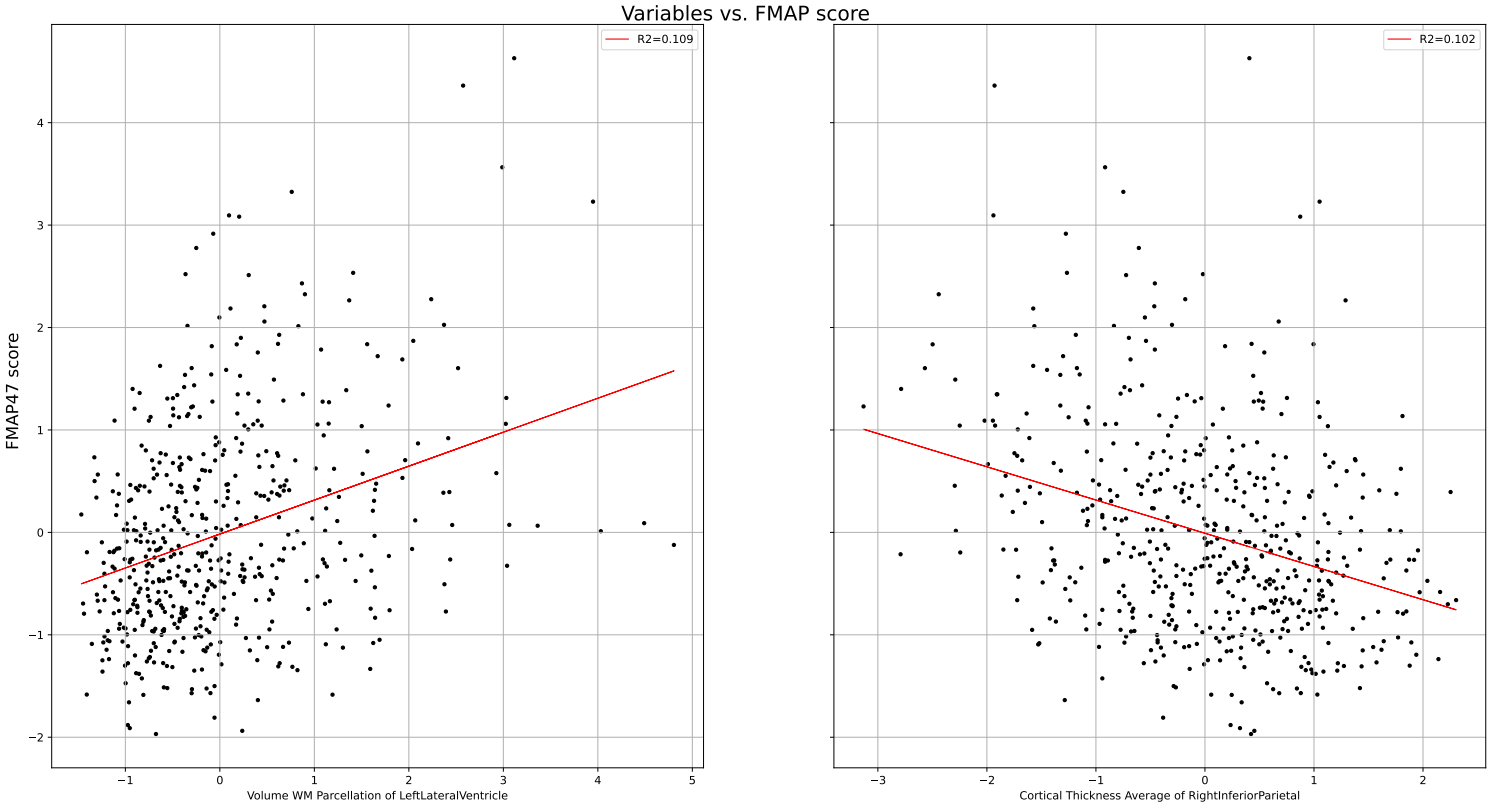


Figure 3.12: Scatter plots of terms significantly associated with FMAP47 score activity holding constant every other variable plus age and sex (volume of left lateral ventricle and thickness of right inferior parietal, left to right).

3.3.6 Power analysis

Our power analysis indicated that with a sample size of 744 and a genome wide significance threshold of $5e-8$, we are at 80% power to detect variants of effect size 0.318 when the minor allele frequency of those variants is 0.5 (Figure A.3). The effect size refers to the variance explained by the SNP in question. To maintain 80% power at lower minor allele frequencies, the effect size of the variant must be larger. For example, our study is powered to detect 80% of variants at minor allele frequencies of 0.1 provided they have true effect sizes of greater than 0.53, which translates to a large amount of variance explained for a single variant.

3.3.7 Ancestrally restricted GWAS

We found that results did not differ substantially when 53 samples were removed for an ancestrally restricted GWAS of Caucasian participants. Instead of five genomic risk loci, 3 were reported, with 5 lead SNPs returned instead of 7. There were no differences in functional annotation results, with the same genes reported in *MAGMA* tests and the same annotation enrichment categories (intergenic, intronic, and ncRNA intronic). The removal of 53 samples meant that p-value magnitude decreased across all variants (Figure A.5). Principal components 1 and 2 appeared to capture ancestry effects (Figure A.1).

3.4 Discussion

3.4.1 Model construction, evaluation, and performance

The principles underlying Chapter 1 influenced our architectural choices and general experimental practices. We created a parsimonious model with just 397 thousand trainable parameters. In comparison, many popular CNN image modelling architectures (such as VGG-16), have upwards of 100 million trainable parameters [95]. While the use of a global average pooling layer was primarily intended to facilitate interpretability at the penultimate layer, it also has the added advantage of enforcing sparsity onto the model's representation of the modelling space through averaging the entire signal from a feature map into a single number. This acts in tandem with our choice of activation function (the rectified linear unit) which can propagate zeros throughout the network (the function returns a 0 where values are negative). Such modelling choices that emphasize sparsity may reduce prediction bias in different data domains.

Furthermore, our model achieved competitive predictive accuracies compared to other models trained on ADNI data, especially considering the relatively small amount of trainable parameters in our architec-

ture. The final accuracy of our best-performing model on a hold-out test set was 82%, whereas the mean testing accuracy reported by studies making use of the ADNI dataset was 90%. While further model tuning may yield gains in predictive performance, it is worth noting that further performance evaluation on separate datasets would first be required to measure generalisation performance. While model weights are not directly informed by testing labels, the vast majority of deep learning applications are architecturally informed by testing set performance – therefore, conditioning model hyperparameter choices on testing performances in one population may limit the applicability of the model at large to other populations. It is of interest to consider the relationship between testing accuracy and genetic signal strength – is a model with increased testing accuracy more likely to yield a higher number of significant genetic variants? We will expound upon this point at a later point of the discussion.

3.4.2 GWAS results

Our heritability estimate for the feature map with the largest contribution to AD classification was low and statistically insignificant. Despite this, Figures 3.6 and 3.9 indicate several genome-wide significant results of biological interest. For example, several genes mapped to our lead SNPs have been investigated in AD-related molecular phenotypes. *PSME4* encodes a subunit of the proteasome complex which is responsible for maintenance of cellular homeostasis through degradation of aberrant proteins. Additionally, one of the defining characteristics of AD is the deposition of extracellular plaques, which may be facilitated by impaired proteasomal function. In [222], it was reported that proteasome subunit gene expression in a population of AD mice was increased relative to age-matched controls. While the fold change of *PSME4* was consistently greater in AD cortex gene expression profiles compared to controls, the authors reported no significant test statistics. However, *PSME4* contains 31 SNPs in linkage disequilibrium ($r^2 > 0.6$) with lead SNP rs2542584, the majority of which are intronic. One such tagged SNP, rs805309, with the highest CADD score ($CADD > 8, r^2 = 0.76$) is located in the first intron of the *PSME4*. How altered expression of *PSME4* and its effect on proteasomal activity effects AD pathology is currently unknown. *PSME4* was also found to be significantly differentially expressed in Parkinson's disease (negative log fold change relative to controls) and AD (positive log fold change relative to controls), suggesting that the proteaosme in general is a biological system of interest with respect to neurodegeneration [226, 227].

Additionally, *SPTBN1* is a member of the spectrin forming complex, an often-disrupted system in neurodegenerative diseases [223]. According to Figure 3.8, this gene appears to exhibit high expression in brain and nerve tissues. However, its observed log fold change according to gene tissue expression signatures is non-significant in both cases. Whole-exome protein-truncating variant analyses of children with autism spectrum disorder and attention deficit hyperactivity disorder in a Danish population showed

that *SPTBN1* variants had a significant odds ratio in a logistic regression of cases vs. controls, but this association did not survive genome-wide Bonferroni significance testing ($OR = 14.9, p = 9.90 \times 10^{-4}$) [228]. Interestingly, the tagged SNP with the highest CADD score, rs78496188 (CADD=17.23) is located in the second intron of the gene (which has 2 supported transcripts). Additionally, the variant in question falls within a region identified as a likely transcription start site by the ENCODE consortium [229]. However, its functional consequence remains difficult to interpret - there is no evidence of any splicing effect caused by this SNP, or indeed any understanding of the effect of splicing variants on the expression of *SPTBN1* in the context of spectrin formation. According to GTEx results, this gene is expressed the most in nervous tissues in units of transcripts per million (Figure 3.8). Given its role in determining cell shape in brain tissues through spectrin formation, this finding indicates that our phenotype may be associated with a gene known to be associated with neurdegenerative and psychiatric phenotypes via associated SNPs [223, 228].

Another gene of interest in our results, *ERLEC1*, has been previously implicated in the molecular pathology of AD in mouse brains. In [224], the authors suggest that membralin is a co-precipatory protein of *ERLEC1*, which forms part of the endoplasmic reticulum associated degradation (ERAD) pathway, responsible for cellular homeostasis. Further, the authors simulate amyloid-β like pathology in mouse AD brains by downregulating membralin, indicating that *ERLEC1* and the ERAD system at large may be involved with the molecular pathology of AD. Furthermore, analyses of human RNA-Seq experimental datasets in [230] classified *ERLEC1* as a hub gene amongst networks of AD-differential gene sets. While these findings appear to increase the biological plausibility of our results, it is unknown how the specifics of the ERAD system function in human brain cells, and moreover, how the system interacts with AD molecular pathology.

Two genes reaching significance in MAGMA's gene-based test, *CCDC104* and *SMEK2*, had variants in LD with rs8080, which has a CADD score of 18.03 (Figure 3.9). The variant is located in the 17th exon of the *SMEK2* gene, also known as *PPP4R3B*. Curiously, there is no literature support indicating these genes and the associated high-CADD variant are involved with AD, although GTEx results suggest *SMEK2* has high expression in brain and testicular tissues. In contrast, [231] found *ITGA1* was significantly differentially expressed in APOE-$\epsilon 4$ carriers relative to controls based on blood samples. However, GTEx results show that levels of gene expression are low in every region of the brain relative to other tissues, suggesting limited evidence for plausible molecular association. Further, *TBC1D22A*, a gene showing specifically high expression in lymphocytes, has not been reported by previous studies of AD. Overall, this suggests there is limited prior evidence that results found to be significant in the gene-based test are associated with AD in any meaningful way.

However, our power analysis suggests that we are powered to detect variants of effect sizes of at least 0.32 when the minor allele frequency is 0.5. This effect size is large and by definition the minor allele frequency cannot exceed 0.5, making common variants of these qualities implausible. As the minor allele frequency decreases, the effect size must increase in order for our study to maintain 80% power. It is unlikely that any variants have effect sizes of this magnitude. This means that if there are any variants of small effect associated with our endophenotype, we are underpowered to detect them. It is also worth noting that while the exclusion of 53 samples did not cause major differences in outcome results, the overall magnitude of p-values decreased. This suggests that our study is sensitive to loss of samples.

It is of general interest that the only major peak in our Manhattan plot is on chromosome 2. Additionally, it is a broad peak, which may be caused by artefacts in the genotype data. This is because it is unlikely that multiple independent significant variants of effect sizes large enough to be detected by our study are concentrated in the same region. Further, an examination of the Manhattan plots from the online PheWeb browser provided by [232] shows that imaging derived phenotypes (IDPs) associated with our CNN score do not exhibit broad peaks or LD peaks on chromosome 2. Additionally, an examination of Figure 1 from [35] shows an absence of broad peaks. The tail end of chromosome 3 also contains many variants with similar p-values, a structure not present in any previous studies. This suggests that there is limited variability in this genomic region, which may be further evidence of artefacts in the genotype data. Consequently, this casts doubt on the overall validity of our GWAS results. More stringent quality control procedures in combination with larger sample sizes may be necessary to increase confidence in the results of future studies.

3.4.3 Feature map score correlation with structural neuroimaging measures

Our combination of gradient-based methods, PC-component regression, and standard linear models aided our interpretative efforts by allowing us to query the visual and empirical basis of our feature map score. Interestingly, gradient results indicated changes in collections of pixels on the edge of brain tissue had the greatest impact on feature map score magnitude. While it is difficult to identify the exact direction of effect in all cases, gradient values located near the edges of brain tissue are likely related to brain size differences, because changes in either direction will influence this quantity. This conclusion appears to be supported by our complementary PC regression results; the top-weighted loadings of our most-correlated PC were related to overall surface area. Additionally, we visually observed lateral ventricle importance for our CNN score which another top loading of the same PC. Results from a larger regression model including all structural neuroimaging measures also indicated that lateral ventricle enlargement and brain size were significant terms. In AD, pathology is often marked by overall brain size decrease, which can be

accompanied by lateral ventricle enlargement [233]. Indeed, enlarged ventricular tissue has been hypothesised to be positively correlated with dementia severity since the 1980's, with this finding replicated in the ADNI dataset across cognitive impairment categorisations [204, 234]. Interestingly, gradient maps of important pixel collections appear to highlight parietal regions of the brain, which is further corroborated by our regression results presented in Figure 3.12. Cortical thinning of the inferior parietal lobule has been noted in AD participants, meaning that our visual and empirical results are consistent with previous observations [235].

Together, this indicates our CNN score represents AD-related neuranatomical information, displaying plausible concordance with well-established structural brain results. In comparing our results to genetic loci associated with brain features from the UK Biobank given by [232], we found that none of our lead SNPs were significantly associated with any imaging-derived phenotypes. This may suggest that general neuroanatomical variation has a distinct genetic architecture to AD-related neuroanatomical variation. Based on genetic results alone, this may suggest that the phenotype extracted from our trained CNN is not capturing relevant information, but our empirical and visual results indicate that AD-specific neuroanatomical variation is encoded in our phenotype.

3.4.4 Global trends, future perspectives, and limitations

Our methods attempt to unify several interdisciplinary fields in a coherent fashion – predictive modelling, statistical inference, deep learning, and genomics. Our approach seeks to exploit the potential of CNN models to leverage non-linear data patterns and understand their dynamics. We have provided a general framework by which a usually-opaque predictive model can be trained and queried using a combination of visual and empirical elements to understand the features inherent to its representation of the data domain problem. Our interest in investigating the genetic correlates of psychiatry-related neuroimaging features meant that interpretation was of primary concern. In order to meaningfully conduct a GWAS of a quantity, we must understand its properties. However, without leveraging the field-specific phenomenon of neuroimaging tabular measures to supplement our findings, it is difficult to envision how gradient-based approaches alone could facilitate the same level of interpretative power. Indeed, although the focus on interpretation is primarily borne from an interest in the genetics of this domain, our results speak to the general crisis of interpretability in deep learning. Our understanding of CNN-derived quantities is fundamentally limited when only considering visual components of any kind, which is the primary output of most popular 'saliency' methods in use today [97]. With this in mind, our results highlight the need for straightforward and interpretable quantities derived from deep learning models that are compatible with GWAS.

Furthermore, while our work provides GWAS results of an interpretable deep learning phenotype, we must be cognizant of the potential bias inherent to our approach. Effectively, we train a predictive model on a binary classification problem and run a GWAS of its learned features. This may result in inflated importance of features conditioned on the representation of a model that has been exposed to labels. This bias may arise from the specific training population in question or any number of other factors influencing model output parameters. The current experimental framework relies on computing phenotypes on the training population that informed the model, which may be considered a specific form of statistical 'double dipping' [236, 237]. Quantifying the extent of this potential bias is difficult. Mitigating its impact would require the use of one separate training dataset to be used for model construction and evaluation, and a separate dataset to compute phenotypes and carry out a GWAS. The sample sizes in the ADNI consortium do not allow for this experimental design. Additionally, the derived phenotype is still difficult to interpret, requiring regression of tabular terms and gradient visualisations to understand.

In a larger sense, it is of interest to consider our expected results in ideal experimental conditions. We may consider that a condition is more likely to causally influence neuroimaging measure variation than *vice versa*. This is because the genetic basis of brain disorders is partially determined from birth, meaning that neuroanatomical changes arise as a result of the phenotype. This may also be plausible as brain structural changes are often plastic compared to the lifetime status of brain disorder classification. From this perspective, observed AD-related neuroanatomical variation may likely only genetically determined insofar as AD is genetically determined and causes the observed changes. However, the opposite may also be true, with neuroanatomical variation influencing cognitive systems and resulting in brain disorder phenotypes. This will be discussed later in the thesis.

3.5 Conclusion

Overall, we report several plausible genetic findings from a GWAS of a deep learning derived phenotype associated with AD. We further provide a framework for CNN model construction and evaluation that centers interpretation and empirical understanding as its guiding principles. Our work serves as a useful proof-of-concept investigation into a subject of great interdisciplinary interest, combining predictive modelling, deep learning, and molecular biology to further understand the biological correlates of a debilitating brain disorder.

Chapter 4

Shallow De-noising Autoencoders to Examine the Genetic Architecture of Sub-Phenotypes of Alzheimer's Disease

4.1 Introduction

The derivation of meaningful and interpretable biological correlates of brain disorders is of great clinical and research importance. Many research efforts to derive these markers have focused on the analysis of large neuroimaging collections using a range of statistical approaches. Most recently, large neuroimaging consortia and high-performance computing infrastructures have facilitated the application of deep learning models to this problem domain, yielding high predictive performances in Alzheimer's disease and other conditions [156, 181]. Accordingly, the brain imaging features of these models have been studied to ascertain their potential as biomarkers. However, interpreting the internal representations of deep learning models is notoriously difficult, with methods often developed only after model deployment. This immediately limits the range of methods that can be applied to understand the dynamics of neural network representations. This is further complicated in the context of endophenotype derivation, as we require an interpretable output quantity that can be used in a genome wide association study (GWAS).

As described in Chapter 3, we developed a custom convolutional neural network (CNN) to derive an interpretable quantity capturing neuroanatomical features of Alzheimer's disease. Furthermore, we found seven significant genome-wide significant loci in a GWAS of our candidate endophenotype. While

promising, our GWAS includes scores derived from members of the training population which was used to train the model, which may introduce bias. As previously mentioned, the CNN-derived quantity is also still difficult to interpret. It is therefore of interest to derive a deep-learning sub-phenotype that does not suffer from these limitations. Namely, we seek an approach that can capture non-linear dependencies between neuroimaging variables related to brain disorders that is not reliant on prior knowledge of condition labels. Additionally, we seek a quantity with interpretative potential of the same degree as our CNN experiments or better. In considering these desired properties, we can examine the use of shallow unsupervised approaches, specifically, the use of de-noising autoencoders [238]. Here, shallow refers to the depth of the model in terms of layer numbers. This can enable straightforward interpretation of the model's low dimensionality internal representation post-training. Further, de-noising autoencoders have useful properties which ensure interpretative potential is not gained at the expense of modelling flexibility, several of which we discuss in the next section.

4.1.1 De-noising Autoencoders

De-noising autoencoders (DA) are deep learning models that map a 'noisy' input to an output of the same dimensionality through an arbitrary number of differentiable layers. The noisy input is usually defined by the addition of random noise to create a corrupted input example and the model is trained to minimise the reconstruction loss between input and output, meaning that the goal is to output a quantity as close to the input as possible. The introduction of corrupted examples is motivated by a desire to ensure that the latent representation learned by the model is robust to variation, an idea explored in [238]. The purpose of the differentiable 'hidden' layers – usually of a lower dimensionality than the input – is twofold; firstly, they prevent an exact one-to-one mapping of input to output, and secondly, the middle layers encode an interpretation based on their weight vectors, the values of which are informed by an objective function that minimises the difference between input and output through a theoretical information bottleneck. As such, a latent space embedding is encoded through a dimensionality reduction procedure, whereby inputs are mapped to the next layer via a series of sum-weighted transformations. After an arbitrary number of such operations, the same procedure is applied to yield an output of the same dimensionality as the input, and the encoded latent space is decoded. This means that the latent space which encodes the input information is embedded with information about the intrinsic properties of the modelling space. Further details on the exact algorithmic details of autoencoders and their variants can be found in [239].

The application of autoencoders can therefore produce an interpretable latent space. An example of this can be found in [238], where the authors apply a shallow one-layer autoencoder to a set of gene expression profiles of the bacteria *psuedomonas aeruginosa* obtained from GEO, detailing bulk RNA-

seq experiments carried out under various experimental conditions. By conditioning the latent space on experimental labels, they identified nodes differentially associated with conditions of interest. Further, because the model is shallow, they examined the variables which contribute the most to the activity of nodes significantly associated with a particular test label by examining the weights vector of that node. They identified several biologically relevant gene sets significantly associated with different experimental conditions. For example, they identify a strain-specific node, with genes differentially expressed across strains featuring as the highest-weight inputs for the node in question, and an anaerobic response node that contains several members of the anaerobic stress response as important determinants of node activity. A similar approach was taken in [240] where the authors use denoising autoencoders to identify gene sets from RNA-seq datasets. Additionally, autoencoders have been used in singe cell RNA-seq experiments for transcript counting [241] and batch correction [242].

4.1.2 Theoretical conception – autoencoders applied to neuroimaging

Autoencoders offer several useful properties for our desired output. In practice, the only use requirement is an input numeric matrix of instances and features, whereby the nodes of a latent space of arbitrary dimensionality are defined as follows:

$$h_k = \sigma\left(\sum_{i=1}^{n} w_i x_i + b\right). \tag{4.1}$$

Here, h_k is the node value at hidden node k (where hidden nodes are indexed 1 through l, with l equal to the total number of hidden nodes), x are the feature variables indexed by i from 1 through n, w is the weights vector with the same dimensionality as x, b is a constant bias term (analogous to the intercept in a standard regression), and σ is an arbitrary activation function. The function used for σ is user-defined, which may include a sigmoidal transformation ($\frac{1}{1+e^{-x}}$). Here, Equation 4.1 describes the derivation of a node value as the sum-weighted transformation of input variables. Because we construct a model with just one layer, the values of the weight vector w have a direct interpretation. As input to the model, we provide a tabular representation of multiple neuroimaging features per participant. This can be obtained by taking the *Freesurfer* outputs denoting the neuroanatomical features of an input individual [109].

As demonstrated in our CNN work, we utilised these tabular measures of brain structure to create a matrix of participants and their associated features, which can also serve as an input for applying the approach detailed in the work carried out by the aforementioned authors [238].

4.1.3 General chapter motivation and phenotype selection

We construct a shallow de-noising autoencoder trained to reconstruct tabular representations of structural brain variation and identify nodes with outputs significantly associated with condition labels. This quantity can be considered a candidate endophenotype, as it will represent the non-linear combination of structural brain features associated with a condition of interest. We also perform a GWAS on the resultant quantity. Given our prior work on Alzheimer's disease (AD), we can build upon these previous findings by considering the same phenotype. While we expect that our method will capture non-linear dependencies, it is reasonable to assume that consistently observed linear relationships can also be recapitulated. Furthermore, the common genetic architecture of AD has been previously described by well-powered GWAS experiments, meaning we can compare the genetic properties of our endophenotype to that of the overall condition [35]. Additionally, we simulate simplified experimental conditions under which we expect our methods to work by generating a phenotype composed of multiple independent signals, each with respective genetic architectures. These simulation experiments allow us to contextualise our main GWAS results and are described below.

4.2 Methods

4.2.1 Simulation experiments

Our goal is to derive endophenotypes using a deep learning model and examine the genetics of those endophenotypes. However, the conditions that facilitate a successful application of this experimental framework are unknown; that is to say we seek a positive control phenotype that can be defined as a mixture of multiple phenotypes which can be decomposed and queried at the genetic level through an autoencoder. This is complicated by the lack of suitable real-world outcomes satisfying these properties. Therefore, we simulated an arbitrary number of uncorrelated phenotypes and associated genetic data of known correlation with respective phenotypes and combined the phenotypes to create features which could be used as inputs to an autoencoder model to test our workflow.

We can simulate phenotypes using the *MASS* package in R to sample from a multivariate normal distribution using a pre-imposed correlation structure. A multivariate normal random variable can be defined by the following function:

$$x \sim \mathcal{N}(\mu, \Sigma), \tag{4.2}$$

Where $\mathcal{N}$ is the normal probability density function, μ is a vector of means, and Σ is a covariance matrix. We seek to simulate a set of uncorrelated phenotypes in the first instance to construct phenotypes that are mixtures of distinct variables. In order to impose a specific correlation structure on the generated variables, we can convert a synthetic correlation matrix to a covariance matrix for the above function. Recalling the definition of Pearson's correlation coefficient, we can see its value is given by the following:

$$R = \frac{cov(X_i, X_j)}{\sqrt{var(X_i)var(X_j)}}, \tag{4.3}$$

Where X is a matrix of variables indexed by i and j in a 2 variable example. Therefore, the covariance between two samples can be rewritten as the following:

$$cov(X_i, X_j) = R\sqrt{var(X_i)var(X_j)}. \tag{4.4}$$

If we abstract this to a general matrix solution, we can think of the relationship between the covariance matrix and correlation matrix as:

$$cov(X) = DRD, \tag{4.5}$$

where D is the diagnonal matrix of variable variance entries. The pre-multiplication between D and the correlation matrix R scales the correlation matrix by the columns of the variance matrix and the post-multiplication scales the result by the rows of the variance matrix. This can be used to yield a covariance matrix with an arbitrary correlation structure, which can be used as the Σ input for Equation 4.2.

Therefore, we simulate 5 uncorrelated phenotypes by generating a correlation matrix following the specified form:

$$R = \begin{bmatrix} 1 & 0.05 & 0.05 & 0.05 & 0.05 \\ 0.05 & 1 & 0.05 & 0.05 & 0.05 \\ 0.05 & 0.05 & 1 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.05 & 1 & 0.05 \\ 0.05 & 0.05 & 0.05 & 0.05 & 1 \end{bmatrix}$$

We can then derive Σ by generating the diagonal of an arbitrary variance matrix. This yields the pairs of independent variables illustrated in Figure 4.1, where each phenotype has 1000 observations.

We can then simulate single nucleotide polymorphism (SNP) values of fixed correlation by a similar procedure. This data does not need to be categorical as in the case of actual SNP data – it must only be correlated with its corresponding simulated phenotype. We can then generate a vector following a

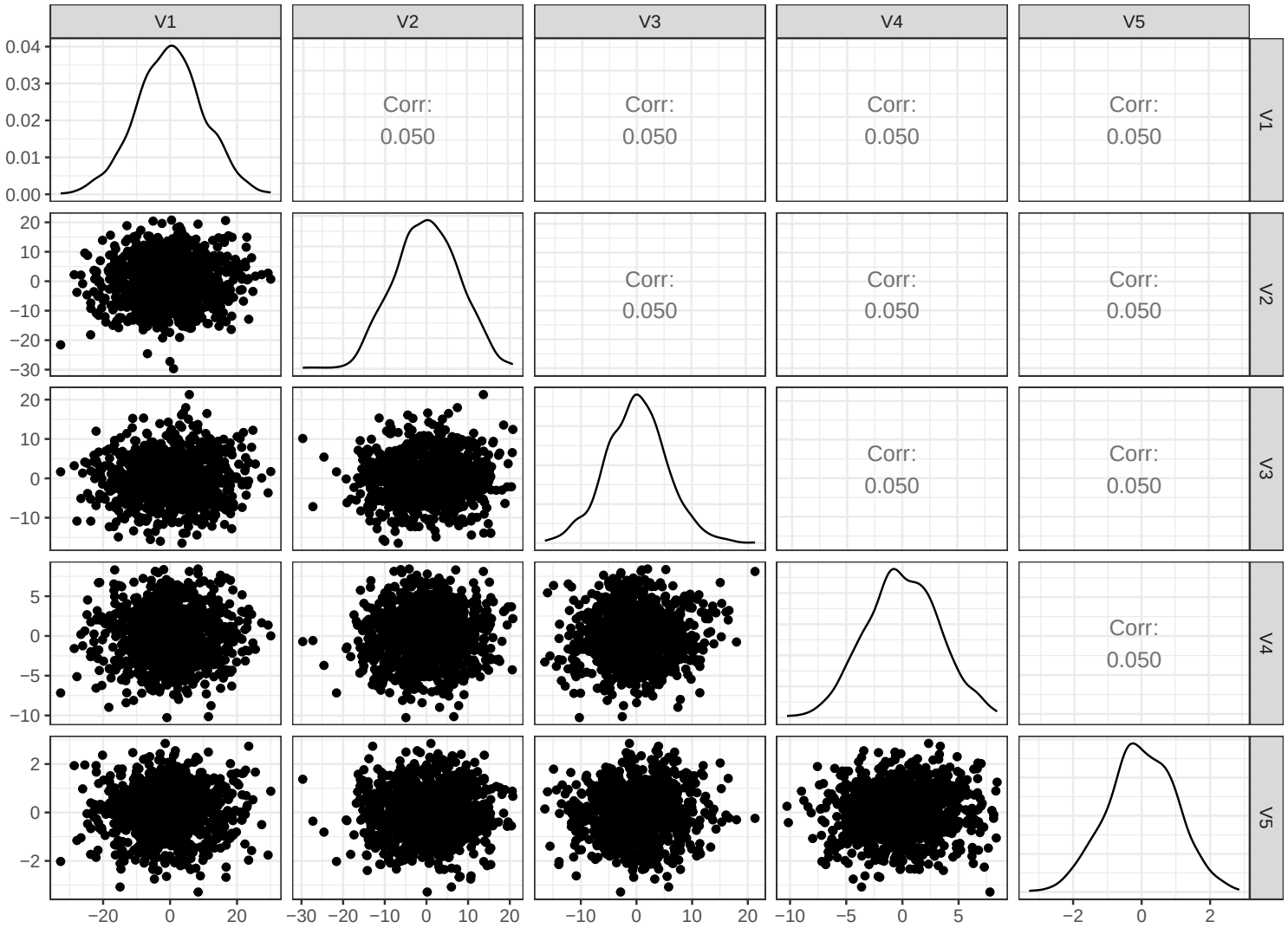


Figure 4.1: Correlation/scatter plots for simulated phenotypes. Here, every phenotype is plotted against the other on the x and y axes respectively; the upper right portion of the panel describes the correlation value between respective phenotypes, identified by the top x-axis and right y-axis of each plot. The density estimates of each individual phenotype are also provided along the diagnonals.

specified distribution and impart a specified correlation structure given our generated phenotypes. This can be achieved by following procedures in [243].

We generate 100 SNPs for each of our 5 phenotypes. Twenty-five SNPs per phenotype are given a combined R^2 value of 0.3 (analogous to a narrow-sense heritability estimate), with 75 SNPs given R^2 values of $1e-6$. The 25 SNPs per phenotype with larger additive R^2 values are considered as 'causal'. For our simulation experiments, we created 10 features per individual given by $\binom{5}{2}$ summed combinations. These features are passed through an autoencoder with five latent nodes and a rectified linear unit activation function. The choice of five latent nodes was motivated by a desire to ensure that the ten input features would pass through an information bottleneck. The latent layer is then decoded to 10 features in an output vector. We trained the model for 100 epochs with the adaptive moment estimator optimizer algorithm using the mean squared loss between input and output as the objective cost function for backpropagation. The resultant latent node activations per individual are then correlated with the underlying phenotypes which form the mixture features that the autoencoder is trained on. A GWAS was carried out on all 500 SNPs by regressing node values against SNP values. Because all data was simulated, no covariates must be accounted for. We rank the most significant SNPs by adjusted p-value and check if the causal SNPs for the corresponding correlated phenotype are significantly enriched in a hypergeometric test (Fisher's one-tailed exact test). Additionally, we took every latent node-phenotype pair and noted the number of causal SNPs from that phenotype present in the top 25 SNPs of the latent node association test results ranked by adjusted p-value. We regressed absolute correlation value against this quantity, hypothesising that pairs with stronger correlation values would consequently have more causal SNPs of that phenotype present in the top 25 SNPs of the respective latent node.

4.2.2 Binary AD GWAS

We also carried out a GWAS of AD status across all individuals with genetic data available in the ADNI dataset to compare to endophenotype GWAS results. AD status was coded as 1 and every other phenotype value was coded as 0, including mild cognitive impairment phenotypes. We then used fastGWA's logistic parameter to carry out logistic regressions of SNPs against AD status, controlling for sex, genotyping centre, age, and the first 10 principal components (as a proxy for ancestry). We used FUMA to investigate the functional enrichment of our results.

4.2.3 MRI brain-derived data selection and preprocessing

We use *Freesurfer*-derived tabular summary information of 576 participants from the Alzheimer's Disease Neuroimaging Initiative [53]. We firstly removed any variables relating to standard deviation measures and retained all neuroimaging measures with information on thickness, volume, or area. We standardised every numeric variable to have mean zero and unit variance. This left us with 249 neuroanatomical variables per participant.

4.2.4 Autoencoder construction and training

We define a shallow autoencoder architecture as presented in Figure 4.2. We set l, the number of hidden nodes, equal to 60 and repeated training twice to assess the reproducibility of genetic results. The choice of setting $l = 60$ was again motivated by a desire to ensure that the 249 variables per participant were mapped to a lower dimensionality space; however, the choice of 60 exactly was arbitrary. We let σ equal the rectified linear unit function.

The model was trained according to a de-noising procedure, whereby the input data has a certain percentage of patient values 'corrupted' by the addition of noise drawn from a random normal distribution ($\mathcal{N}(0, 5e{-}3)$). We trained the model for 8000 epochs and determined the input per epoch by the following:

$$X' = X_j + X_i, \tag{4.6}$$

whereby X' is the input data, X_j is the set of uncorrupted data, and X_i is the set of corrupted data, the total fraction of which with respect to the entire dataset is determined by a float factor, ω. The set of data X_i to be corrupted is given by uniformly sampling data indices at a proportion defined by ω. The data not contained in X_i is X_j. We vary ω gradually, reducing its magnitude every 250 epochs by a constant factor λ (0.85). Therefore,

$$\omega = \begin{cases} \lambda\omega, & \text{if} \quad \mod\ (e, 250) = 0 \\ \omega & \text{otherwise,} \end{cases} \tag{4.7}$$

where e denotes the epoch number. This means that every epoch, a certain proportion of input data is randomly corrupted by the addition of a noise vector; further, every 250 epochs, the proportion of noisy data in the entire set of input training data changes. This schema was applied with the intention of ensuring that the final model latent space representation is robust to variable amounts of input data noise and is related to the idea of learning rate decay [244]. Additionally, this training schema promotes iterative learning of core data patterns owing to large amounts of noise initially followed by refined tuning of

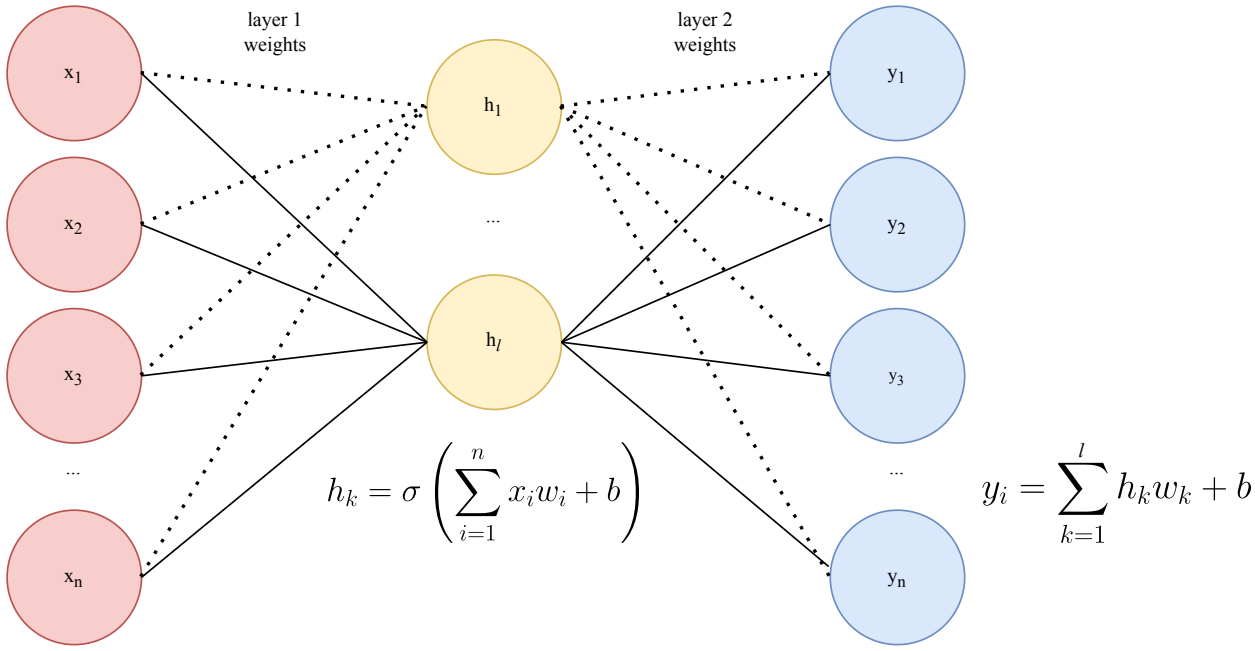


Figure 4.2: Structure of proposed autoencoder. In our application, the index of the hidden node is represented by k and σ is the rectified linear unit function. x is the noisy input. h_k is the value of node k. y_i is the reconstructed output determined by the sum-weighted combination of all hidden node values (h_1 through h_l).

different data patterns imposed by less noise addition. We used the adaptive moment estimator optimizer algorithm with default parameters to train our network [215].

4.2.5 Latent node extraction

After model training, we extracted the latent layer activations for every patient, yielding a matrix of size 576×60. We carried out a 2-sample independent t-test of activation scores in AD participants vs. controls for each of the latent nodes. The resultant p-values were then corrected for multiple testing using the Benjanini-Hochberg method. The most significant node was extracted and any variables with weights $\geq 2\sigma$ outside of the mean activation value were called as 'high-weight' regions and plotted. We further plotted the activity of the most significant node across diagnostic categories to examine whether or not the activity of node values appeared to stratify participants.

4.2.6 GWAS experiments

We followed the GWAS procedure as described in Chapter 3. We carried out a power analysis to explore how well-powered our study was using the given sample size. We also carried out an ancestrally-constrained GWAS of white ethnicity individuals, bring our sample size from 533 to 500.

4.2.7 Correlation experiments

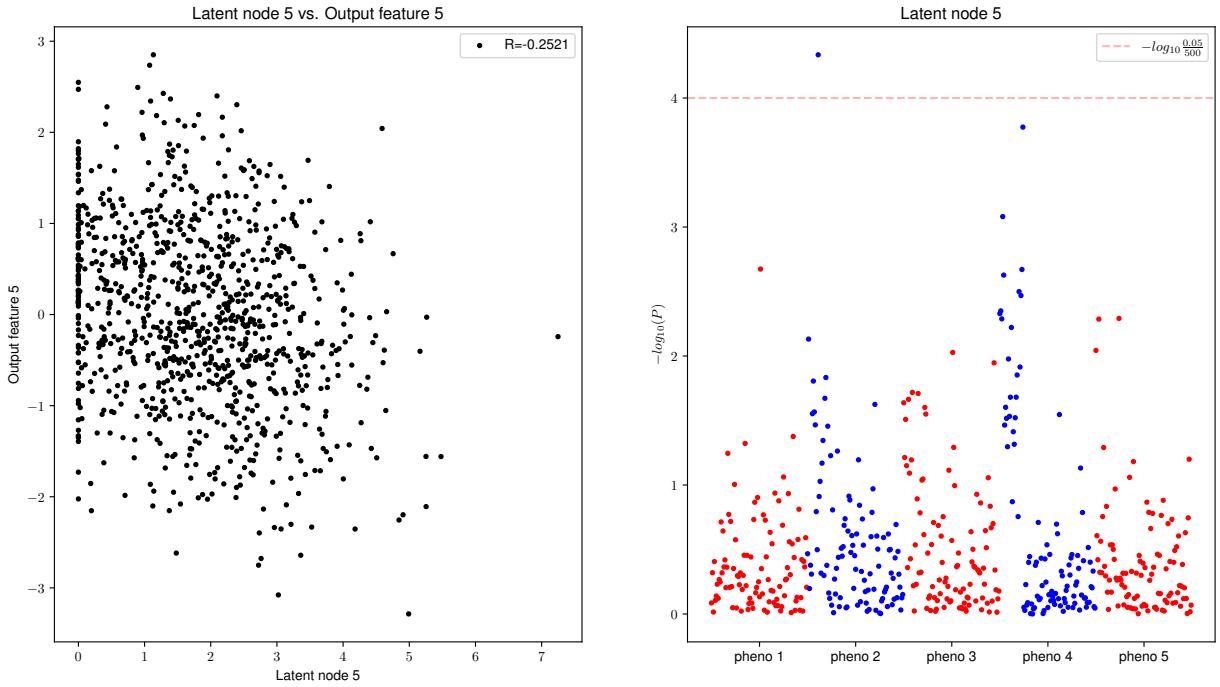
We subset the nominally significant ($P < 0.05$) genetic results of our control AD GWAS, our autoencoder latent node GWAS, and the CNN component GWAS from Chapter 3 and examined the intersections between the three studies. We examined the set membership of the resultant intersections and visualised the β value coefficient correlations in shared sets of nominally significant SNPs across studies.

4.3 Results

4.3.1 Simulation results

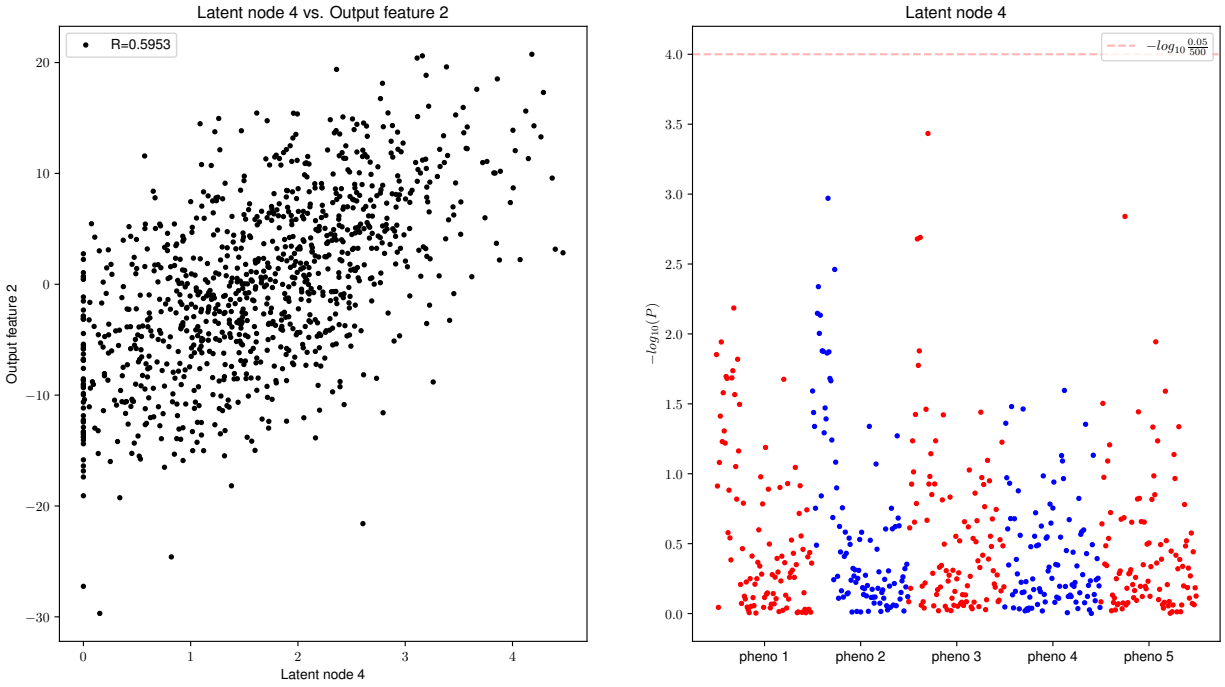
We found that autoencoder latent nodes with high absolute Pearson correlation values with underlying phenotypes had more causal SNPs of that phenotype present in the top 25 p-value ranked SNPs resulting from association tests (Figure 4.3). For example, latent node 3 and phenotype 1 were correlated at -0.79, and there were 18 causal phenotype 1 SNPs present in the top 25 adjusted p-value ranked SNPs of latent node 3 (Figure 4.3c). This finding was found to be significant in Fisher's one-tailed test ($P = 2.67e{-}11$). Further, we observed that correlation strength appeared to be predictive of the number of 'true' causal SNPs present in the top 25 ranked SNPs of respective latent nodes. This can be seen in Figure 4.3b, where latent node 4 and phenotype 2 have a lower magnitude correlation value ($\rho = 0.59$) and thus less phenotype 2 causal SNPs are present in the top 25 ranked SNPs of latent node 4 (11). However, this result was still significant in a Fisher's one-tailed test ($P = 3.11e{-}6$). Nonetheless, where correlation values are lower, as in Figure 4.3a, the observed number of strongly associated phenotype SNPs in the respective latent node top 25 is lesser and furthermore non-significant ($n = 3, P = 0.16$, latent node 5 vs. phenotype 5). Our regression of this relationship is illustrated in Figure 4.4. Here, the absolute correlation between every latent node and real phenotype pair is plotted against the number of strongly associated SNPs of that phenotype present in the latent node's top 25 list. We observe that a standard deviation unit increase in absolute correlation value is associated with an average 0.9 standard deviation unit increase in the number of causal SNPs of that phenotype present in the latent node's top 25 list.

Latent node 5 vs. Output feature 5

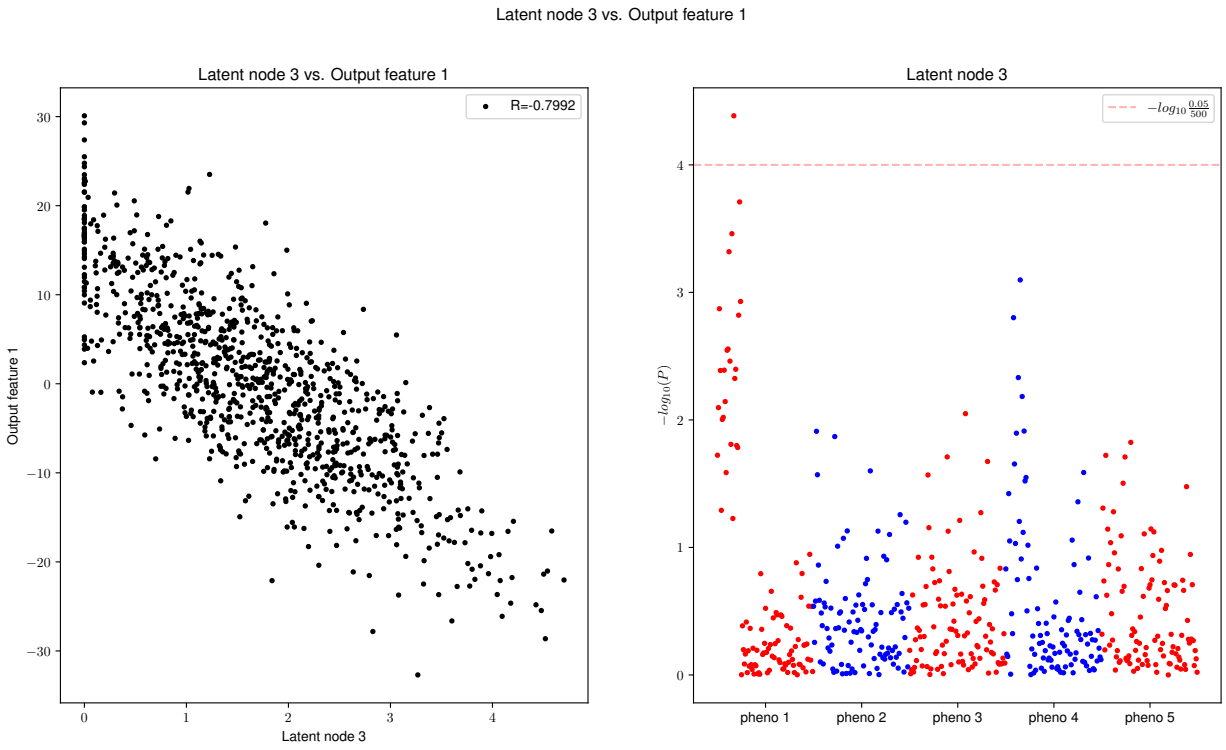


(a)

Latent node 4 vs. Output feature 2



(b)



(c)

Figure 4.3: Positive simulation experiments with scatterplots of phenotypes and latent nodes on the left panel and Manhattan plots of the latent node on the right. (a): Latent node 5 vs. output phenotype 5, where 3 causal phenotype 5 SNPs were present in the top 25 most significant results. (b): Latent node 4 vs. output phenotype 2, where 11 causal phenotype 2 SNPs were present in the top 25 most significant results. (c): Latent node 3 vs. output phenotype 1, where 18 causal phenotype 1 SNPs were present in the top 25 most significant results.

4.3.2 Control AD GWAS

Interestingly, we detected no genome wide significant SNPs in a GWAS of AD status as a binary phenotype (Figure 4.5). Further, no SNPs reached suggestive genome wide significance ($P = 1e - 5$). However, a *MAGMA*-based gene p-value test yielded several significant genes, including *LRRTM4, CTNNA2, SORBS2* and *TENM3*. Further *MAGMA* Genotype-Tissue Expression (GTEx) enrichment analysis revealed that *CTNNA2* has been observed to be significantly upregulated in brain tissues relative to an all-tissue background. Further, we obersved that GTEx enrichment analysis implicated nervous system and brain tissues as the most significantly enriched tissues in a 2-sided differential expression test using

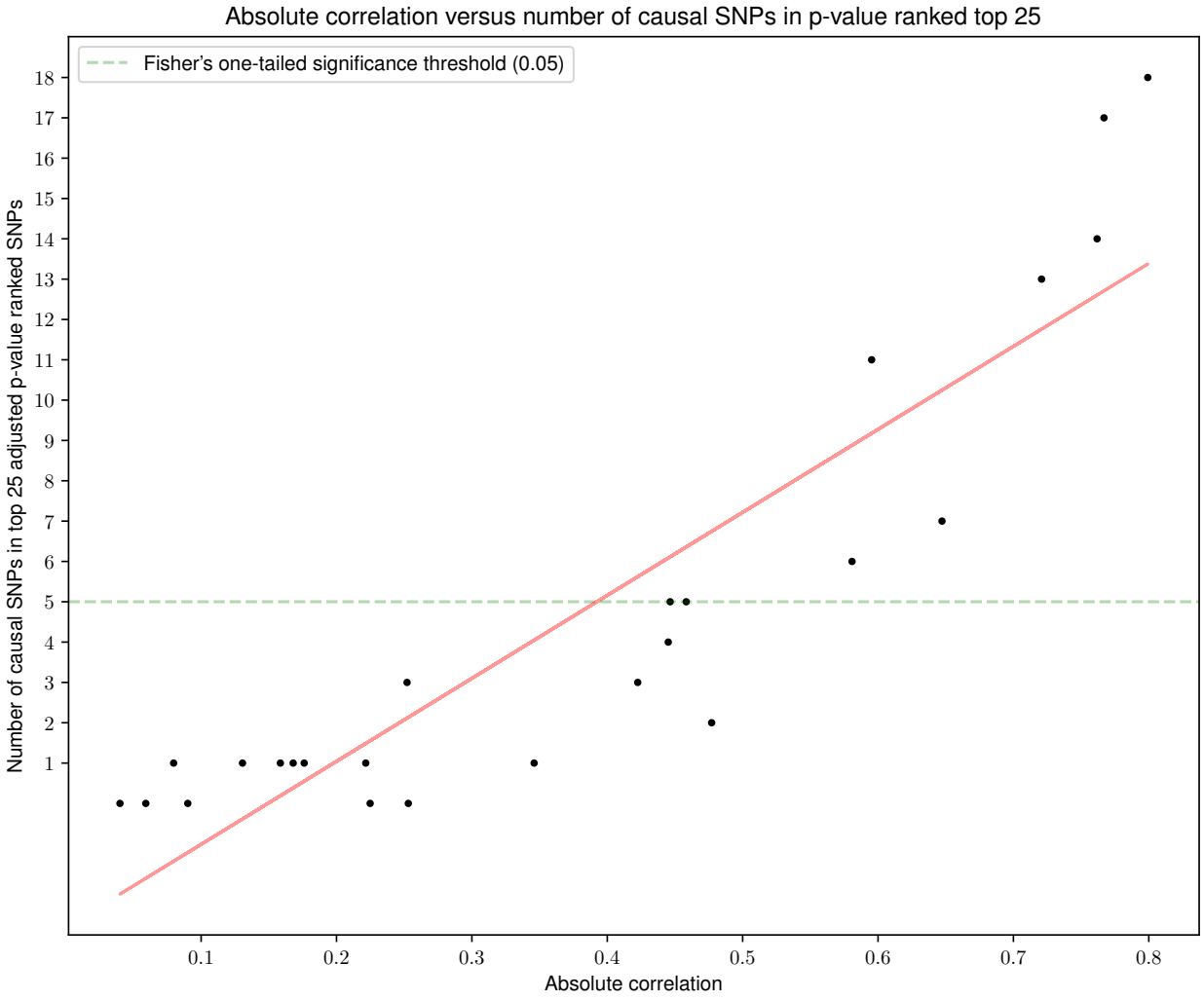


Figure 4.4: Scatterplot of absolute correlation values between pairs of latent nodes and underlying phenotypes against the number of causal SNPs of that underlying phenotype present in the top 25 ranked SNPs of that latent node value.

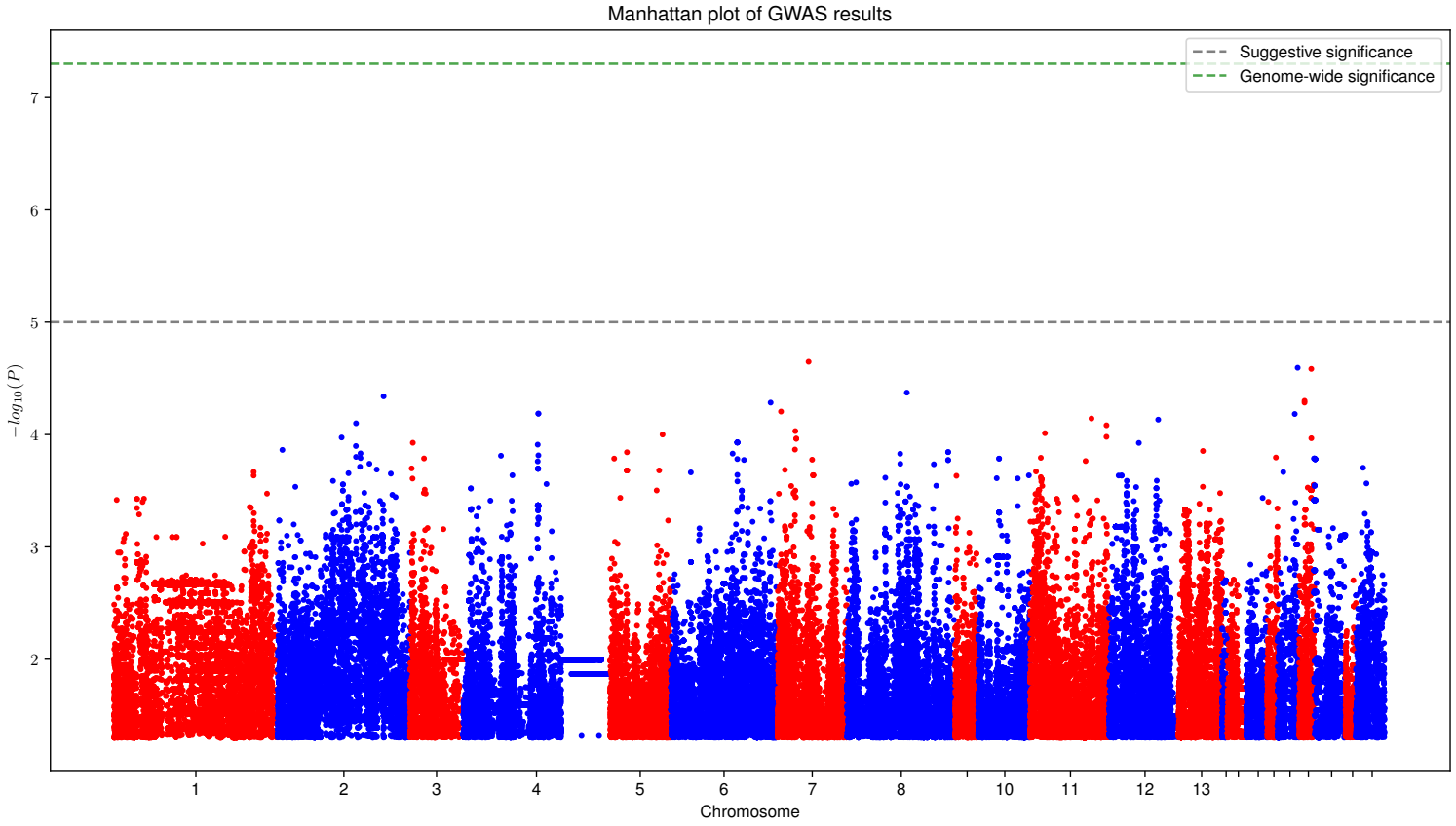


Figure 4.5: GWAS results of AD status in cases and controls. Genome-wide and suggestive significance lines are drawn at $5e-8$ and $1e-5$ respectively.

the aforementioned genes relative to an all-gene backrgound. However, none of these enrichment results were significant after Bonferroni correction.

4.3.3 Autoencoder results and node extraction

We trained our 60-node model to a reconstructive loss of $6.21e{-}9$. Further, we observed that the concordance between values predicted by the model and ground truth values was strong, exhibiting an overall ρ (Pearson correlation coefficient) of 0.92 (Figure 4.6). We identified node 51 as the latent node with the most significant association with AD after Benjanini-Hochberg testing correction of independent t-tests for every node ($adj.P = 2.47e{-}7$). We visualised the mean activity of every node in AD and non-AD participants in Figure 4.7, and observed that nodes with higher or lower values were generally specific to either condition, meaning that they often did not exhibit high values (or low values) in both AD and non-AD participants. Further, we plotted the values of node 51 across diagnostic categories and its activity appeared to be differential with respect to AD status (Figure 4.8). We determined the regions with the most influence on the activity of our discriminative node by examining regions with absolute first-layer weights outside 2 standard deviations of the mean. We then plotted the raw weights of these regions connecting the input to node 51 score in Figure 4.9. We observed patterns of widespread medial temporal lobe atrophy, with lower surface area and/or volume in three medial temporal lobe regions observed as a high-weigh

We observed that increase right thalamic volume, left posterior cingulate volume, left posterior Cingulate surface area, cerebrospinal fluid volume, and fourth venctricle volume were associated with increased node 51 score. The remaining 'high-weight' regions were found to have negative weights, suggesting that lower patient volumetric measures in those regions will increase the value of node 51 activity. This is because standardised inputs with negative values denote lower volumes, and their multiplication with negative weights results in a larger positive output.

4.3.4 GWAS Results for Node 51

We found 3 genome-wide significant SNPs at $P < 5e - 8$ associated with AD in chromosomes 6, 9, and 11 (rs6918430, rs10814283, and rs37621860l, Figure 4.10). The nearest genes to these SNPs are *RP11-239H6.2*, *RP11-509J21.1*, and *RP11-707M1.1* respectively, 3 long non-coding RNAs. These SNPs have not been previously associated with AD or any other neuroimaging phenotype. *MAGMA* GTEx tissue expression enrichment analysis indicated that the brain cerebellum and cerebellar hemisphere tissues were significantly ($P < 0.05$) enriched for genes implicated by our GWAS based on overall p-value distribution.

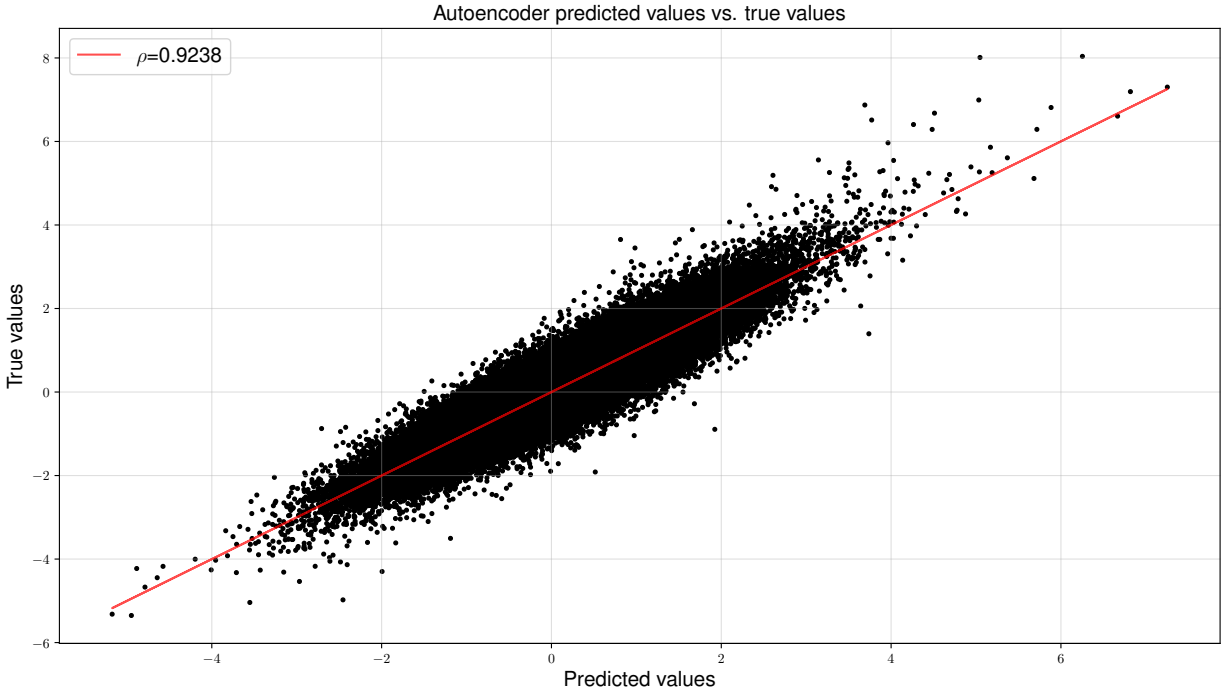


Figure 4.6: Scatterplot of predicted vs. true values from our trained model. Here, the true values are displayed on the y-axis and predicted values from our autoencoder are displayed on the x-axis. The correlation and line of best fit between true and predicted values is displayed in the legend and in red respectively.

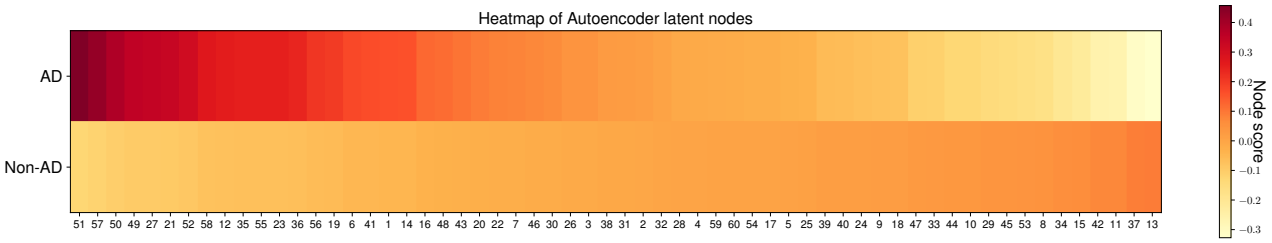


Figure 4.7: Heatmap of autoencoder latent nodes colored by standardised node value. The color intensity represents the value of the node denoted by index in the x-axis; nodes are sorted from high to low based on mean value in AD participants.

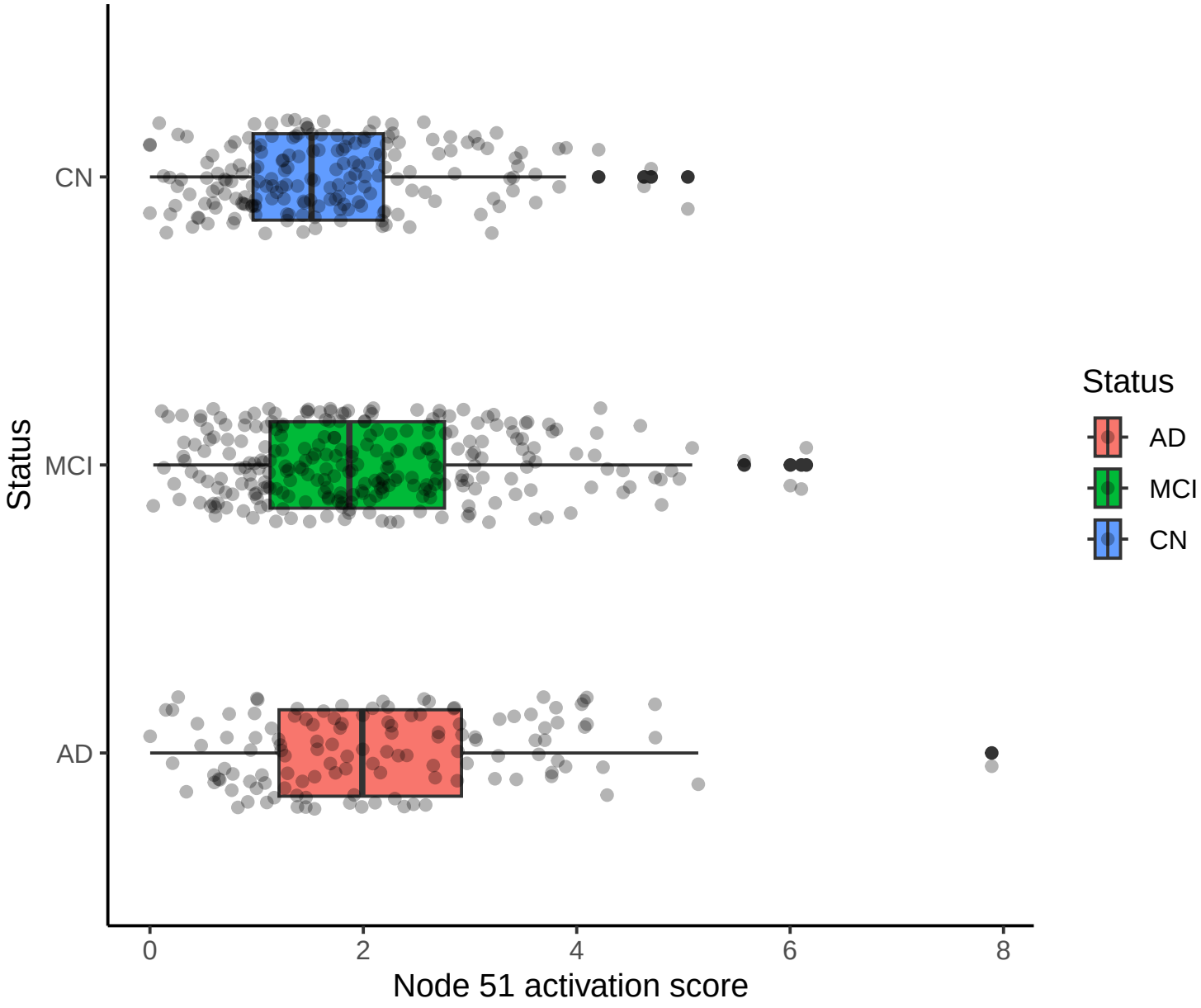


Figure 4.8: Node 51 activity visualised across relevant diagnostic categories. Node 51 was the most significant AD-associated node after multiple testing correction. Pairwise p-values are not displayed as the difference between value in AD and CN participants was already used to select node 51 ($adj.P = 2.47e-07$).

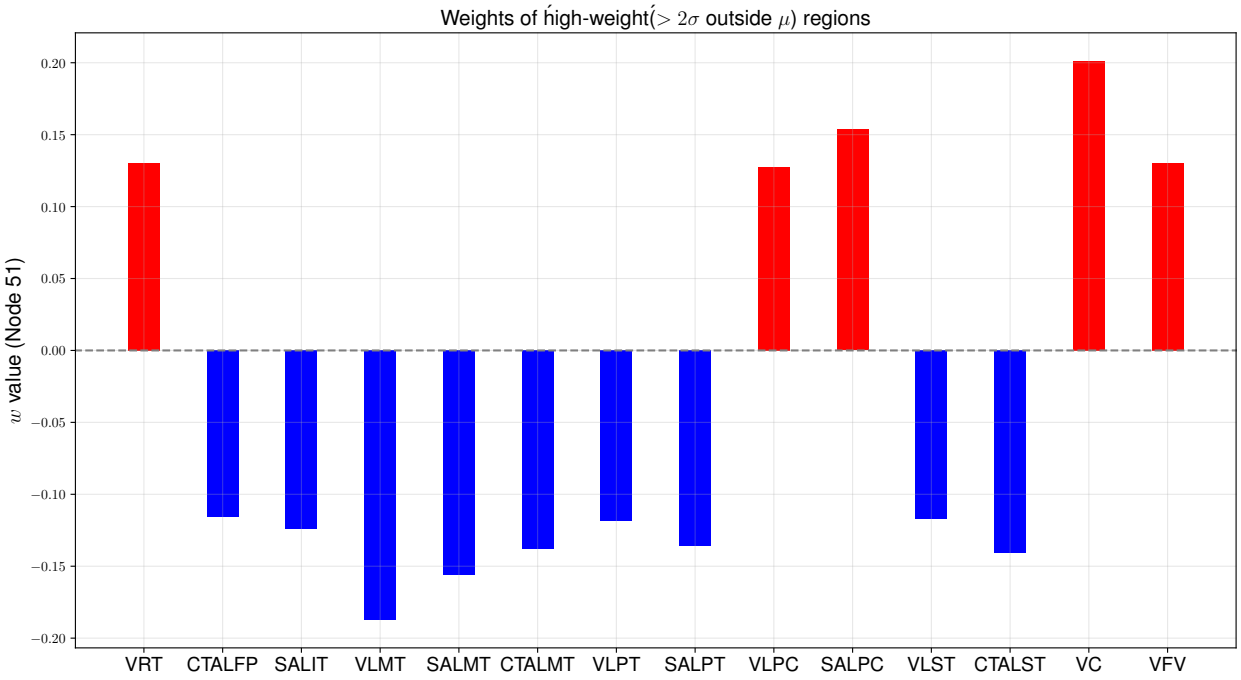


Figure 4.9: Barplot of high weight regions in node 51 (with weights outside of 2 standard deviations of the mean). The x-axis is coded in abbreviations, which correspond to the following - VRT: Volume of Right Thalamus; CTALFP: Cortical Thickness Average of Left Frontal Pole; SALIT: Surface Area of Left Inferior Temporal; VLMT: Volume of Left Middle Temporal; SALMT: Surface Area of Left Middle Temporal; CTALMT: Cortical Thickness Average of Left Middle Temporal; VLPT: Volume of Left Pars Triangularis; SALPT: Surface Area of Left Pars Triangularis; VLPC: Volume of Left Posterior Cingulate; SALPC: Surface Area of Left Posterior Cingulate; VLST: Volume of Left Superior Temporal; CTALST: Cortical Thickness Average of Left Superior Temporal; VC: Volume of Cerebrospinal fluid; VFV: Volume of Fourth Ventricle.

While these results were nominally significant, they did not survive Bonferroni correction across all tissue enrichment tests. Similarly, the most significant ($P = 0.03$) general tissue type implicated by *MAGMA* gene p-value tissue enrichment analysis was the brain, but this did not survive Bonferroni correction.

Our gene-based *MAGMA* analysis yielded several genes deemed significant at our threshold level, defined by $0.05/n_{gene}$ (Table 4.1, Figure 4.11). Further, we found that a *MAGMA* differential expression analysis using the genes in Table 4.1 showed that the brain substantia nigra tissue was the most significantly enriched tissue in a two-sided enrichment test based on our 12 input genes ($P = 0.007$). However, this result did not survive Bonferroni correction. We noted that *EIF2B5* exhibited higher log_2 expression values in this region than other genes, and that *EIF2B5* also exhibited high expression values in other brain tissues, such as the cerebellum, cerebellar hemisphere, and two temporal lobe structures – the amygdala and hippocampus (Figure 4.12).

In our ancestrally-constrained GWAS of 500 samples, we found 15 independent significant SNPs (Figure A.6). This was in spite of accounting for principal components that captured ancestry effects (Figure A.2). They were mapped to seven genes, *MROH8, SUV420H1, KIAA1549L, CHKA, GMDS, CRIM1,* and *MBOAT1*. *SUV420H1* encodes a methyl transferase protein and was identified as an associated gene of AD by a recent whole exome sequencing study of individuals of European ancestry [245]. Further, *MBOAT1* was found to be a target of a microRNA in a network analysis of AD gene expression data [246]. A *MAGMA* gene property test for tissue specificity showed that the brain was the most significant tissue, surviving Bonferroni correction.

We found that with 533 samples we have 80% power to detect variants of a minor allele frequency of 0.1 if their effect sizes are 0.6. When the minor allele frequency of variants is 0.5, we can achieve 80% power if the effect size is at least 0.37 (Figure A.4).

4.3.5 Overlap across GWAS

We found that 2004 SNPs at $P < 0.05$ were shared across the 3 GWAS performed over Chapters 2 and 3 (Figure 4.13). We also found that pairwise Pearson correlations of SNP β values were high, with ρ=0.86, 0.75, and 0.71 for the pairs in Figures 4.14, 4.15, and 4.16 respectively. The largest total set of SNPs detected at $P < 0.05$ were identified in the binary AD status GWAS (n=120,619), as well as the largest unique set of SNPs (n=100,539, Figure 4.13).

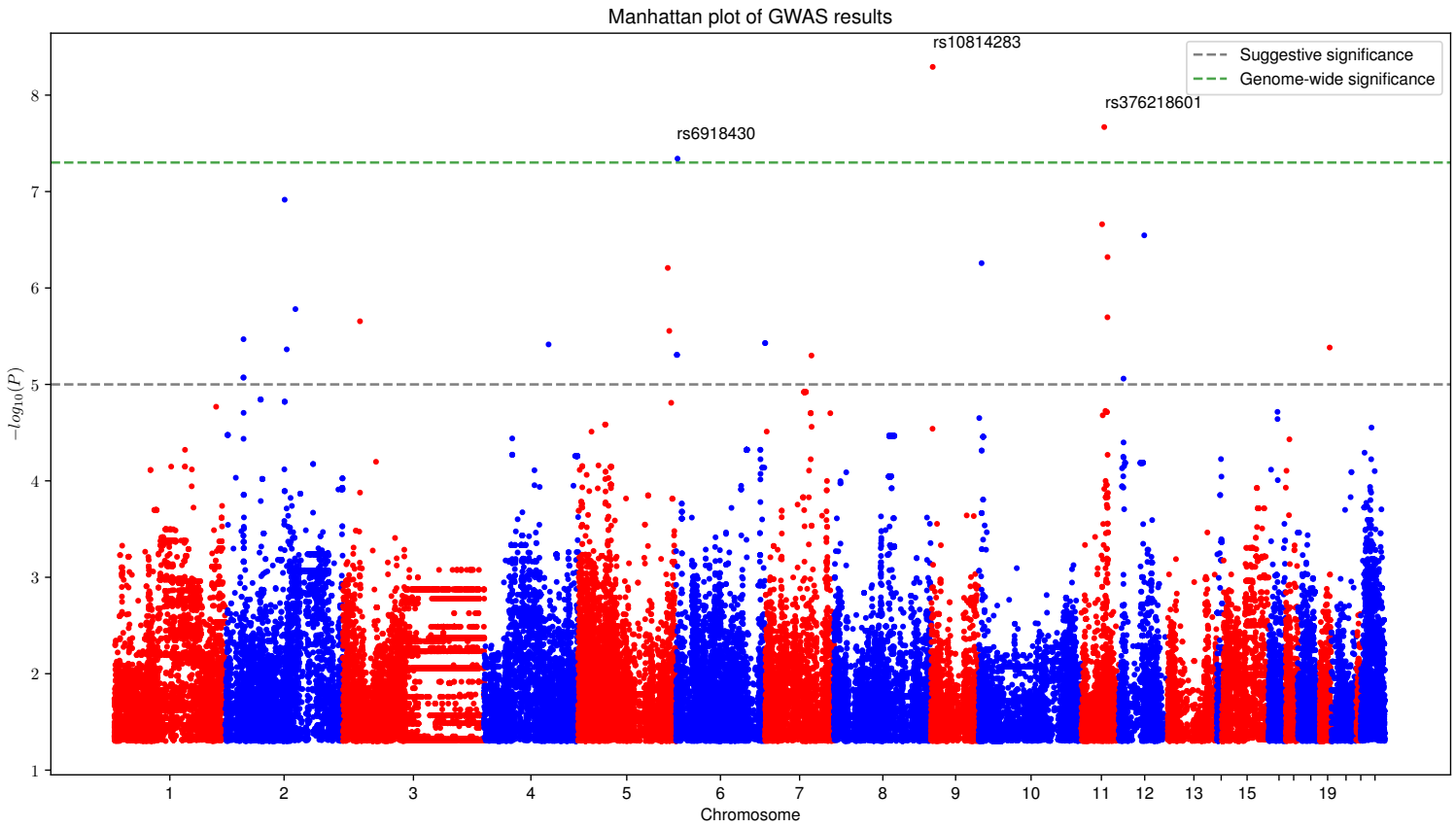


Figure 4.10: GWAS results of autoencoder node 51, which was the most significant node associated with AD after multiple correction testing.

Table 4.1: Genes deemed significant at $P < 0.05/n_{gene}$based on a *MAGMA* test of SNP p-value distribution. NSNPs denotes the number of SNPs present in the corresponding gene.

SYMBOL	P	NSNPS	CHR	START	STOP
ERBB4	1.80e-08	972	2	212240446	213403565
SPAG16	2.22e-07	923	2	214149113	215275225
VWC2L	5.64e-07	162	2	215275789	215443683
ABCA12	2.03e-06	193	2	215796266	216003151
MREG	2.87e-06	52	2	216809213	216898819
PECR	2.98e-06	53	2	216861052	216947678
IGFBP2	2.32e-06	6	2	217497551	217529159
DIRC3	3.78e-09	414	2	218148742	218621316
EIF2B5	1.33e-10	358	3	183852826	184402546
LPP	2.28e-06	353	3	187871072	188608460
TP63	8.19e-08	187	3	189349205	189615068
FGF12	9.66e-08	332	3	191857184	192485553

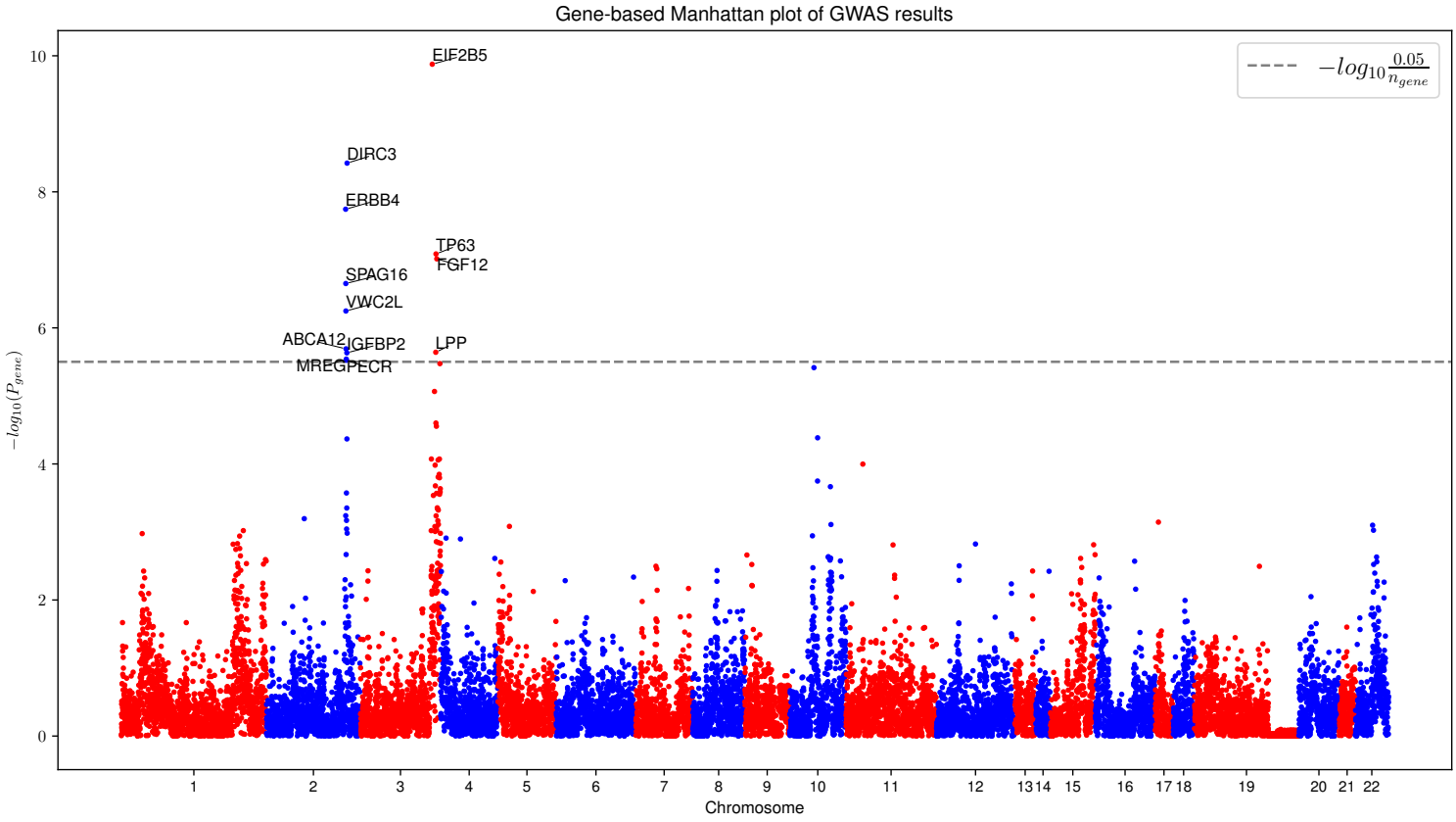


Figure 4.11: *MAGMA* gene-based test results of autoencoder node 51, with genes reaching significance ($0.05/n_{genes}$) annotated.

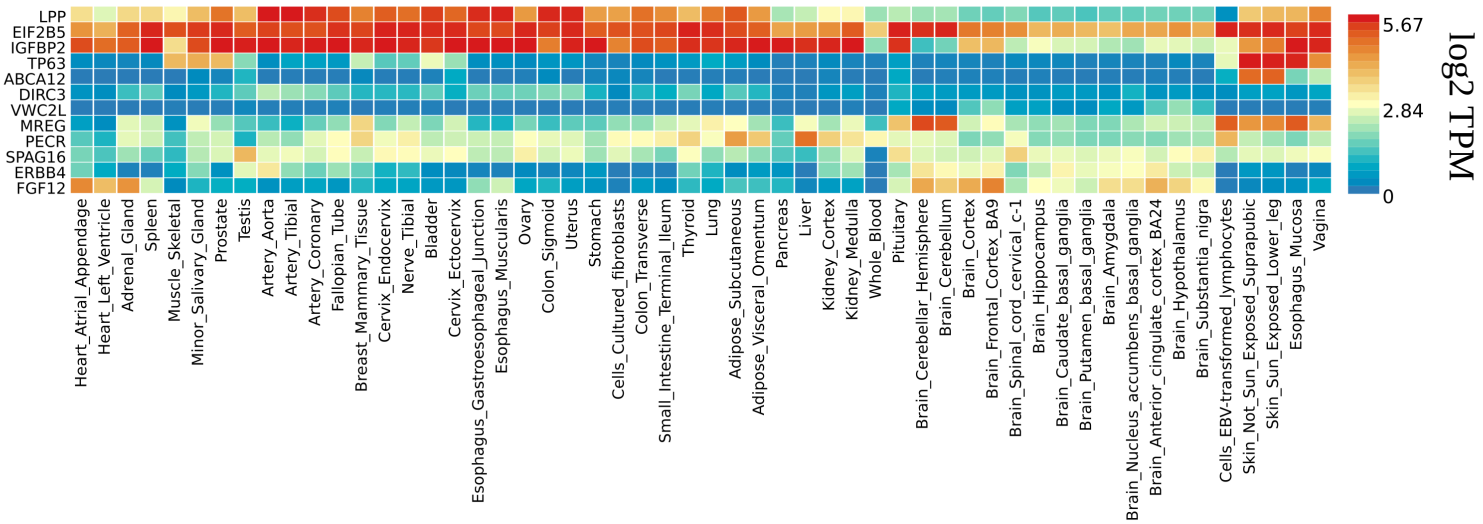


Figure 4.12: GTEx heatmap of gene expression variation across tissues for genes called as significant by a *MAGMA* gene-based p-value test. Values are represented in units of $log_2 TPM$

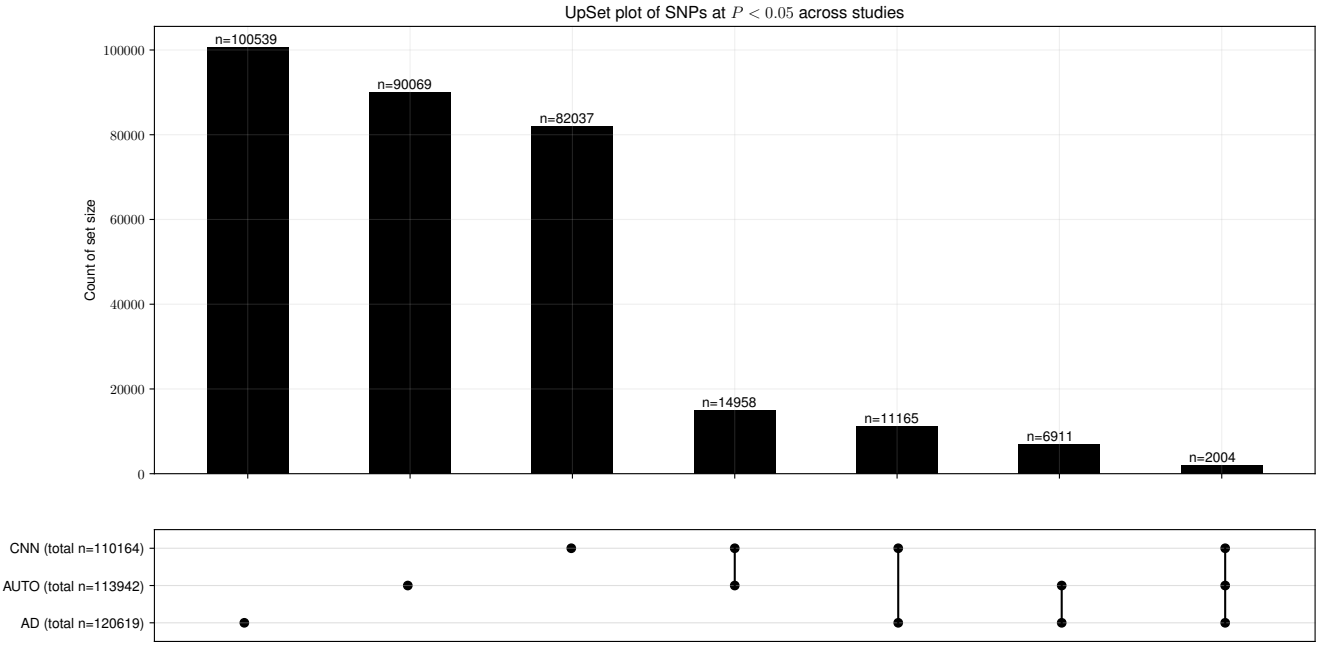


Figure 4.13: UpSet plot of SNPs found to be nominally significant ($P < 0.05$) across the 3 GWAS carried out over Chapters 2 and 3 - CNN denotes feature map 47 from our trained CNN, AUTO is node 51 from our trained autoencoder, and AD denotes our GWAS of binary AD status.

4.4 Discussion

4.4.1 Simulation experiments: decomposing simulated composite phenotypes

Our simulation experiment results indicate that our theoretical framework is empirically verifiable under certain assumptions. Broadly, we found that composite phenotypes of multiple uncorrelated phenotypes can be decomposed using our autoencoder architecture. Consequently, variables strongly correlated with the true phenotypes can also be recovered by regressing the corresponding nodes against those variables. Our regression from Figure 4.4 demonstrates the strong positive relationship between node-phenotype correlation strength and the number of highly correlated variables (in our case, simulated SNPs) present in the top 25 most significant associations. These results show that our proposed methodology can recover data patterns correlated with variables that make up a composite phenotype. Therefore, we can hypothesise that applying an autoencoder to a set of variables associated with a larger phenotype and finding a latent node of sufficient correlation strength with that phenotype has the potential to also highlight variables associated with the larger phenotype.

Interestingly, we also noted that our regression from Figure 4.4 suggests that phenotype correlation with node activity is predictive of causal SNP inclusion in our top 25 lists across all latent node-phenotype correlation pairs; this suggests that where a node is strongly correlated with multiple phenotypes, causal SNPs from those phenotypes will feature at statistically significant levels in the top 25 ranked SNPs of that node. We further note that the 25 strongly associated SNPs of each phenotype only explain 30 percent of the variation in our phenotypes ($R^2 \approx 0.012$ per SNP), meaning that the correct SNPs are still present as significant associations even when their correlation with the latent node is low.

However, it is worth noting that the full set of assumptions which provide the baseline for our simulation experiments may be unrealistic. For instance, no linkage disequilibrium is assumed, and confounding covariates are not simulated. Further, more complex combinations of phenotypes may impact our results. In practice, the number of SNPs in a real-world GWAS is much higher. Our experiments suggest that when correlation is high between latent nodes and their inputs, other variables correlated with those inputs can be identified using an autoencoder. However, we do not know the properties of the larger phenotypes in our example, and therefore cannot measure the correlation of latent nodes directly against them. Our power analysis also suggests that true variants must have larger effect sizes than our simulations account for.

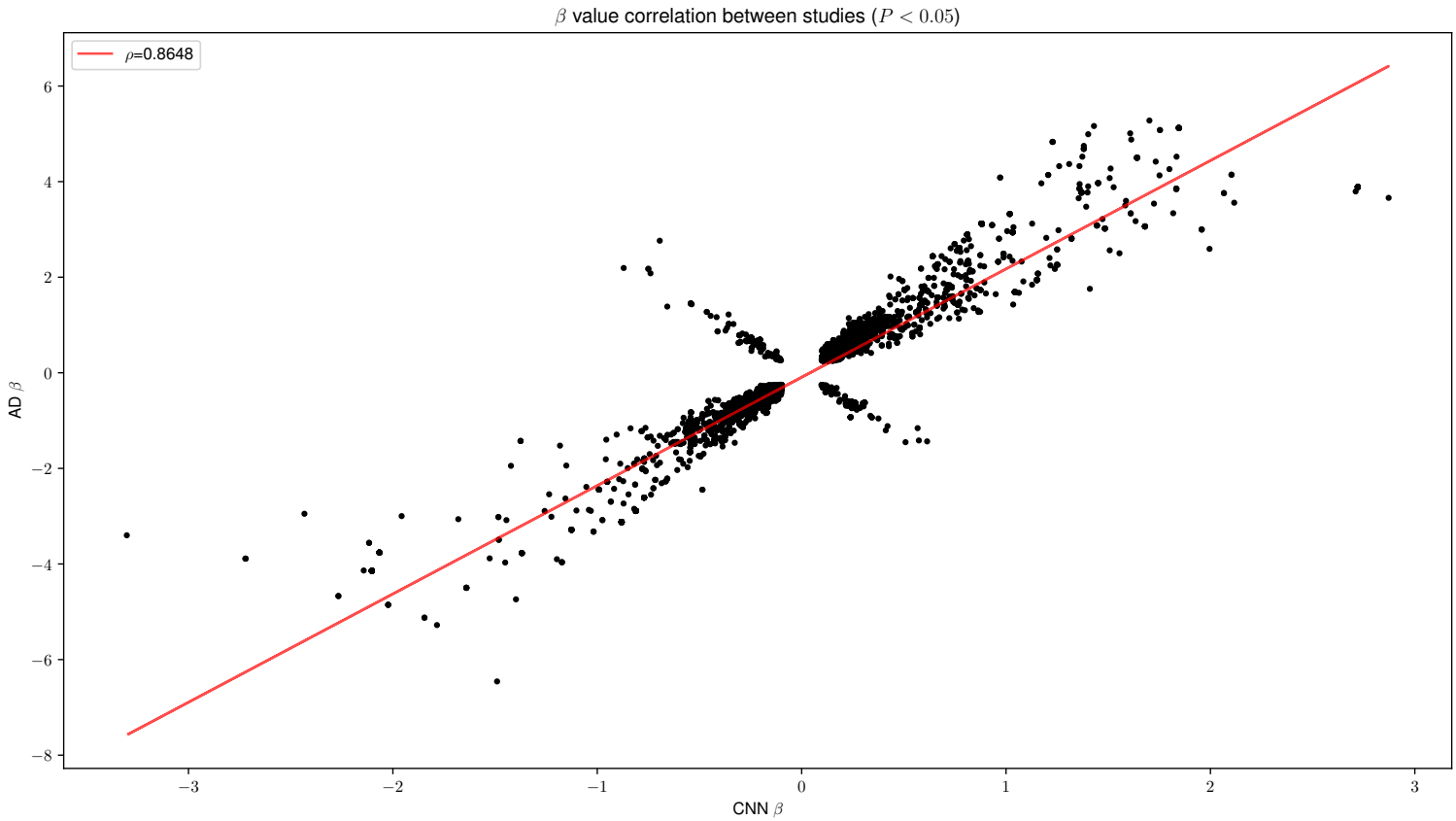


Figure 4.14: Correlation (ρ) between nominally significant SNP β's from our CNN GWAS and the nominally significant SNP β's from our binary AD status GWAS. In the upper left and bottom right corners of the plot, we can see certain SNPs present in both studies at the specified significance level that had differing signs of effect. The line of best fit is denoted in red, with CNN SNP β values on the x-axis and AD SNP β values on the y-axis.

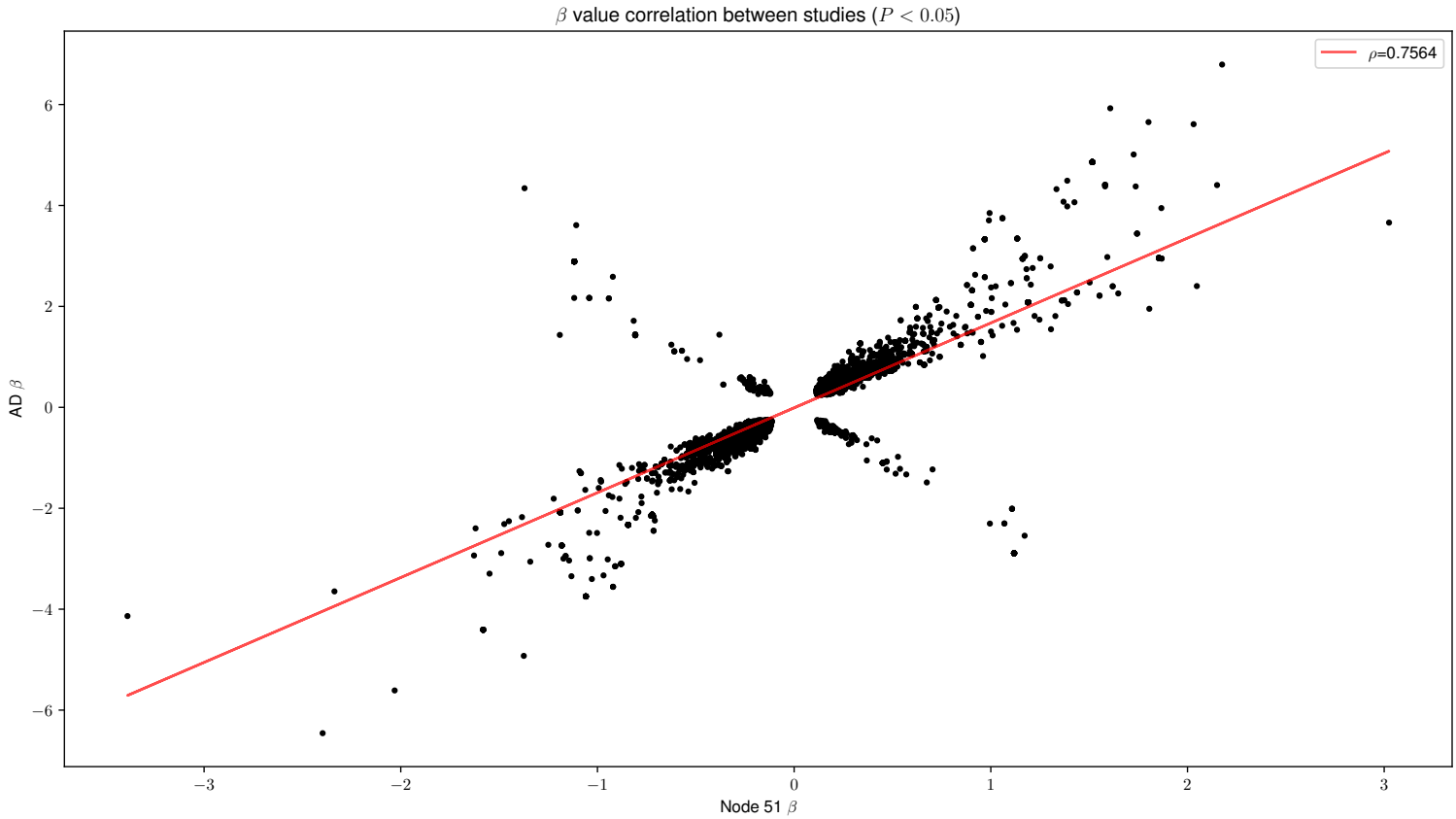


Figure 4.15: Correlation (ρ) between nominally significant SNP β's from our autoencoder node 51 GWAS and the nominally significant SNP β's from our binary AD status GWAS. Similarly, in the upper left and bottom right corners of the plot, we can see SNPs present in both studies with differing directions of effect. The line of best fit is denoted in red, with node 51 SNP β values on the x-axis and AD SNP β values on the y-axis.

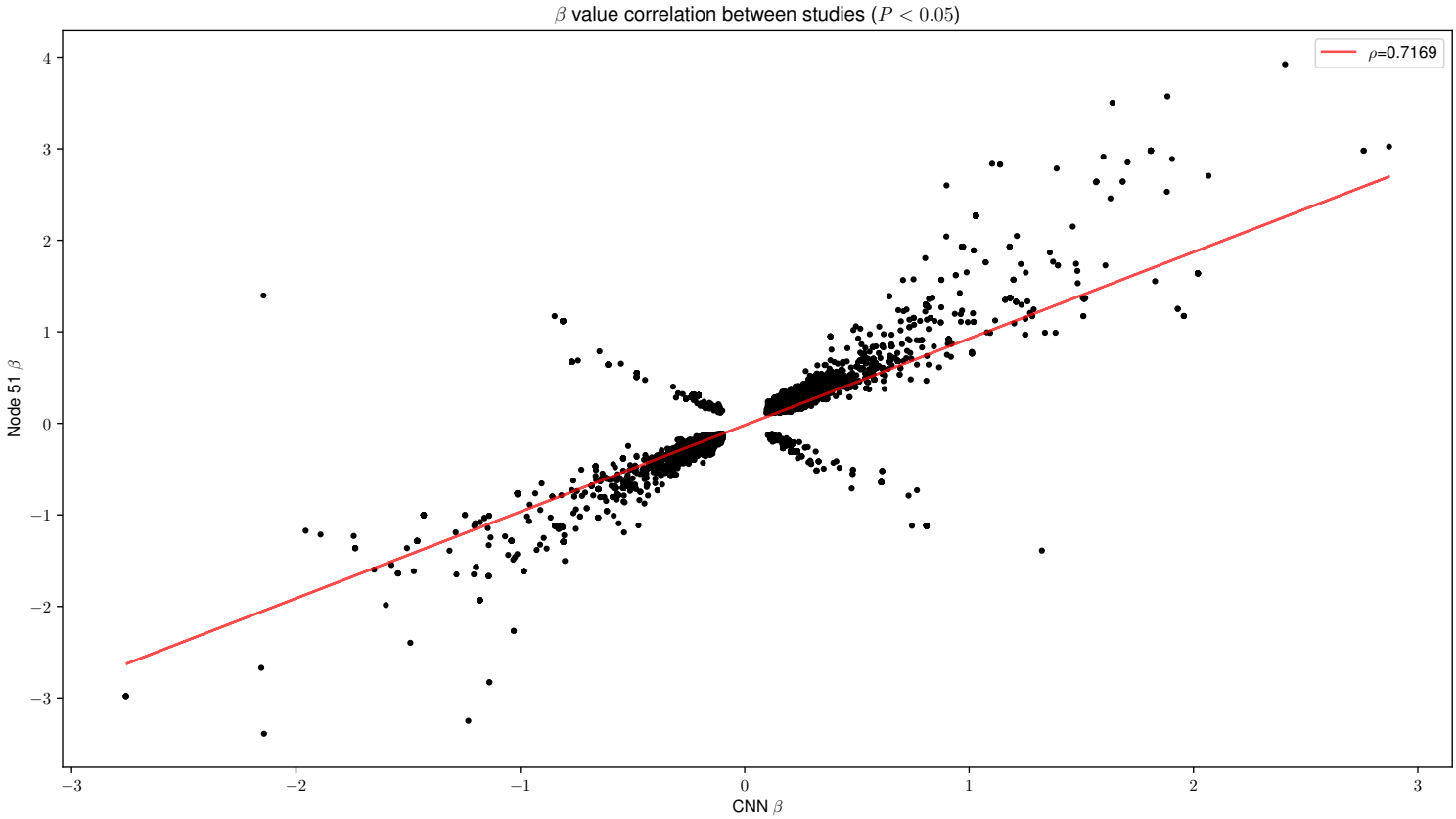


Figure 4.16: Correlation (ρ) between nominally significant SNP β's from our CNN GWAS and the nominally significant SNP β's from our autoencoder node 51 GWAS. The line of best fit is denoted in red, with CNN SNP β values on the x-axis and node 51 SNP β values on the y-axis.

4.4.2 Binary AD GWAS

We report no genome-wide significant SNPs in our GWAS of binary AD status. This is striking considering the GWAS results of our CNN approach and our discriminative autoencoder node, which both found significant associations with limited sample sizes. Additionally, while the gene-based *MAGMA* results yielded several significant genes, only one has published evidence of AD association - the *CTNNA2* locus [247]. Furthermore, this study was also carried out in the ADNI population, meaning the finding may be database-specific.

It is interesting to note that despite our inability to detect significant SNPs, the binary AD GWAS identified the most SNPs at nominal uncorrected significance with $P < 0.05$ relative to other GWAS undertaken (Figure 4.13). This may be due to the fact that the binary AD phenotype is a composite of multiple phenotypic constructs, and as such the signal is not specific to any one of those constructs. Instead of characterising the amount of specifically neuroanatomical phenotypic variation associated with additive genetic effects, we are instead capturing the genetic features associated with variation at a binary phenotype that may be comprised of multiple distinct phenotypes, which may dilute the signal.

4.4.3 High-weight autoencoder regions are biologically relevant

We find that structural measures of the inferior, middle, and superior temporal lobe gyri structures have significant negative weights in node 51, meaning lower values result in higher output scores. This implies that greater region atrophy contributes to higher values in a node discriminative for AD. Temporal lobe atrophy has been observed in participants with AD in a number of previous studies in different databases [248, 249] and longitudinal studies within the same cohort [250]. Additionally, atrophy in medial temporal lobe structures was found to be explain a percentage of variation in clinical dementia rating score across 2 cohorts of AD participants [203]. The body of literature consistent with our findings suggests that our approach can recapitulate a reproducibly observed biomarker of AD through examining the high-weight regions of a discriminative autoencoder node. Further, we observed that every temporal lobe structural measure related to the left hemisphere, which is consistent with previous observations of predominantly left hemispheric temporal lobe atrophy [251]. Observations of asymmetrical volumetric abnormalities have been previously noted in AD, lending further credence to our results capturing relevant pathological features of AD neuroanatomy [163].

We also described increased volume of the fourth ventricle and cerebrospinal fluid as high-weight regions of node 51. Enlarged ventricles (and its proxy variable, increased volume of cerebrospinal fluid present in the ventricles) are a common observation in AD participants relative to other neuroanatomical

structures [233]. It is of interest to note that node 51 score therefore comprises multiple measures of temporal lobe atrophy and ventricle enlargement, two mechanistically relevant AD features with differing directions of effect. Overall, we observed that the majority of high-weight regions had negative weights, meaning that larger negative values, corresponding to volumetric reductions on the standardised scale, resulted in a larger node 51 score (Figure 4.9). Interestingly, we noted that there is broad overlap amongst the regions implicated, and that they typically have the same direction of effect - for instance, the weight values for the surface area and volume of the left posterior cingulate are 0.15 and 0.12 respectively, and their correlation with each other is 0.82. Generally, where multiple measures exist for specific regions, they are often highly correlated; for example, while surface area and volume are not perfectly correlated (and one can not be used to infer the other), they often capture similar information about the dimensions of a region.

Interestingly, while posterior cingulate functional activity has been previously implicated as predictive of AD progression, increased volume and surface area of the left posterior cingulate does not appear to have been previously observed as a feature of AD [252]. Recalling our observation of strong left hemispheric region respresentation in our high-weight lists, a recent review investigated the evidence for potential 'left-brain vulnerability' in AD, citing neuroimaging studies that observe increased levels of cortical thinning in the left hemispheres of cases relative to controls (which may be a function of accelerated brain ageing) [253, 254]. This makes the finding of size increases in a left hemispheric structure difficult to understand; however, our previous discussion of left temporal lobe atrophy is entirely consistent with the current understanding of AD neuropathology.

Despite the fact that multiple descriptors relating to the same region are often highly correlated, as in the case of the left middle temporal, the autoencoder's representation will account for each variable's contribution non-redundantly. This is to say that multiple correlated variables with respectively high weights in a particular node will still contribute to the overall output individually in the summing operation regardless of the correlation structure of the input. Where such situations occur, we can infer that the regions in question, such as the left middle temporal gyrus or left pars triangularis structures, may be especially important to the model's representation of AD patient brain anatomy.

Further, we note that there is limited concordance with CNN structural results aside from an overall finding of left hemispheric size reduction. In particular, the results of our CNN approach implicate increased volume of the third ventricle as important as opposed to the fourth, but the correlation between these variables is modest ($\rho = 0.59$).

4.4.4 GWAS results: the role of long non-coding RNAs in AD and other genetic findings

We detected 3 genome-wide significant loci in our GWAS of discriminative node 51 (Figure 4.10). All three SNPs are located in non-coding regions, with FUMA mapping their respective closest transcripts to long non-coding RNAs. Long non-coding RNAs are regulatory molecules that are not translated into proteins and can contribute to the regulatory landscape of transcript expression across the genome. They can also act as precursors to regulatory micro RNAs, which similarly can affect gene expression. This result is interesting in the context of an emerging body of literature investigating circulating long non-coding RNA's as biomarkers of AD [255]. The meta-analysis by [255] inferred from 7 studies that AD status can be predicted with modest predictive performance based on the expression of several plasma-circulating long non-coding RNAs. Additionally, several long non-coding RNAs have been shown to contribute to AD-related molecular pathways; for example, elevated expression of the antisense transcript of *BACE* leads to amyloid-β aggregation through post-transcriptional gene regulation of *BACE* [256]. The authors of [257] posit several mechanisms by which such RNAs can regulate gene expression pathways within the context of AD, including contribution to micro RNA gene regulation networks and generally promoting amyloid-β aggregation. However, there is limited literature support for any associations with AD amongst the respective transcripts highlighted by our work. For instance, amplifications at the *RP11-239H6.2* locus located at 6p22.3 (approximately 3 megabases from the beginning of the major histocompatibility complex) have been posited as a potential biomarker of bladder cancer, correlating well with late stage cases [258]. Further, SNPs in high LD with this variant ($R^2 > 0.8$) are variants located in the *E2F3* gene, which is known to influence cellular proliferation [259]. Similarly, *RP11-509J21.1* has uncertain function and no previous associations with AD (or any other neurological/neurodegenerative disorders). However, variants in the *RP11-707M1.1* locus have been recently identified as genome-wide significant in a GWAS of attention-deficit hyperactivity disorder [260]. Despite this, the study in question did not identify the variant for *RP11-707M1.1* as significant, and further, our results indicate that our lead variant for this transcript was not in LD with any variants at an R^2 value greater than 0.6. Overall, it is well established that long non-coding RNAs have important roles in regulatory networks of multiple molecular pathways – however, how exactly they may contribute to neuroanatomical variation in the context of AD is unknown. Given that previous work appears to indicate that amyloid-β aggregation is affected in part by long non-coding RNA transcript activity, our results are encouraging, but the exact means by which these variants are associated with AD remains unclear. Furthermore, our power analysis suggests that with our current sample sizes we are underpowered to detect significant associations with small to moderate effect sizes.

This may also indicate that any significant SNPs are false positives. Additionally, significant variants are not associated with LD peaks, which may further decrease confidence in our results.

Our gene-set enrichment analysis based on the p-value distribution of SNPs and their associated genes indicated 4 gene sets as significant, 2 of which were associated with breast cancer amplicons. However, based on the same SNP p-value distribution, several genes of interest that have no immediate cancer links were found to be statistically significant after multiple testing correction, including *EIF2B5*, which is expressed at high levels across multiple brain tissues (Figure 3.8). It is encouraging to note that temporal lobe atrophy was the main feature encoded in our autoencoder weights, and two temporal lobe structures – the amygdala and hippocampus – have generally high expression values for this gene. Despite this gene's high average expression pattern across multiple tissues – including non-brain tissues – previous studies have observed that mutations of *EIF2B5* can cause features of adult-onset leukodystrophy, a disorder often marked by white matter atrophy [261]. Indeed, leukodystrophy, despite being a distinct clinical entity, is often mistaken for AD given a degree of symptomatic overlap. Despite its relevance, it is unclear how best to interpret this result - we would expect larger single-association test statistics from the *EIF2B5* locus in the GWAS presented in Figure 4.10. However, a recent study found *EIF2B5* expression had a significant causal effect on AD risk from Mendelian randomization experiments in a multi-ancestry cohort (FDR $P = 0.00248$) [262]. However, the authors state that colocalisation signal between the loci is not strong ($R^2 < 0.05$), meaning the direct link is tenuous. Further, in the context of the previous results, a positive causal effect of *EIF2B5* expression on AD risk may also be explained by potential misdiagnosis of leukodystrophy, although this hypothesis is difficult to empirically test (and further presupposes that leukodystrophy-related mutations in the locus cause increased expression as opposed to vise versa).

Interestingly, the substantia nigra is observed as the most significantly enriched tissue in a test of differentially expressed genes from our input set in Table 4.1. Lesions and tau phoshporylation deposits have been a historical observation in the substantia nigra in AD participants [263]. Further, the substantia nigra is an important brain region in the dopaminergic system [264]. Neuronal loss in this region has historically been observed in AD participants, with more recent proteomic studies suggesting several proteins in the region undergo expression changes in AD relative to controls [265, 266]. While neuronal loss is not apparent in all substantia nigra AD studies, a comprehensive meta-analysis found significantly lower dopamine levels in AD participants relative to controls, suggesting that functional alterations in this region (which may not always manifest as atrophy) may have relevance to certain aspects of AD pathology and/or presentation [267]. Despite these encouraging results, it is important to note that the tissue enrichment test did not survive Bonferroni correction for multiple testing; further, recent studies of substantia nigra functional and anatomical variation are limited, with the majority of studies cited more

than 10 years old. In spite of these factors, our tissue results overall indicate brain-related functions are captured by our autoencoder score.

Additionally, we note that *ERBB4* has been shown to mediate amyloid-β neurotoxicity in AD mouse models, but its precise molecular mechanism in human tissues remains to be verified [268]. *ERBB4* is a receptor of neuregulin-1, an important transcription factor influencing early brain development and the development of oligodendrocytes. One study found elevated levels of neuregulin-1 in the cerebrospinal fluid of AD participants, but it is difficult to determine the precise dynamics of the neuregulin-*ERBB4* complex; for instance, it is unknown how variants in this gene can impact neuregulin activity, and its exact relationship with AD [269].

In our ancestrally-constrained GWAS, we find diverging results from that of the main analysis. While results had tissue specificity for the brain based on a *MAGMA* gene property model, we note that of seven identified mapped genes, only two have previous AD-related literature support [245, 246]. This is interesting given that ancestry appears to be captured in our principal components and accounted for in our regression (Supplementary materials). The degree to which results differ based on the exclusion of 33 samples indicates that sample sizes may be too small. This is further supported by our power analysis which suggests that identified significant SNPs may be false positives. These results also contrast from the ancestrally-constrained GWAS carried out in Chapter 3, whereby the exclusion of samples only meant that certain SNPs were not found significant in a smaller cohort. This may call into question the stability of this phenotype given that results differed so drastically.

4.4.5 Method agreements, disagreements, and perspectives

We note that there is little overlap amongst the three GWAS carried out in the same dataset for nominally significant SNPs, with (Figure 4.13). If our results are to be trusted, this may suggest that each method is representing relatively distinct elements of AD presentation, and hence, returning different genetic results. However, our lack of power across all GWAS means that a large portion of SNPs nominally significant may be false positives. The fact that different methods may capture different elements of disorder pathology may be unsurprising in the context of our theoretical framing of the problem space; binary AD status captures a broad range of symptoms, and both CNN and autoencoder methods respectively capture specific elements relating primarly to AD neuroanatomy. This would be of interest to investigate in cohorts of larger sample size.

Interestingly, our CNN nominally significant β's had the highest correlation with its shared SNPs in our binary AD GWAS (Figure 4.14). The interpretation of this observation is non-trivial. If we assume our binary AD GWAS is our 'gold standard', this implies that our CNN GWAS captures more signal

traditionally associated with AD, albeit with the subphenotype derivation resulting in specific feature encoding, thus resulting in significant GWAS hits where they are not present in the baseline. However, we can regard the binary AD GWAS in a more critical fashion, with the absence of previous AD associations implying the dataset is underpowered to act as a comprehensive baseline. Framing our results as a false dichotomy of which method is 'best' ignores the fact that the purpose of using our respective methods is to capture features distinct from binary categorisation; positing that a successful subphenotype is constituted by a near-complete reconstruction of the main phenotype renders our theoretical and practical arguments circular. Rather, we should judge the quality of our derived sub-phenotypes by their consistencies (or lack thereof) with prior results and use them to inform novel investigative studies.

For instance, in the case of our autoencoder GWAS, while our hits mapping directly to long non-coding RNA transcripts are initially confusing, the supposed contradiction relative to the AD literature may inform us as to the properties of a quantity whose activity is robustly correlated with AD status. This raises a further point related to the specificity of our conditioned variable; while we condition on correlation to AD status, we have no guarantee that an interesting sub-phenotype for AD will always be given by statistically significant node values. In practice, it may be interesting to examine nodes with different characteristics – our approach is primarily designed to ensure that we are investigating an autoencoder node with a clear and statistically verifiable relationship to phenotypic categories. From this perspective, we may also consider that combinations of multiple nodes ranked by test statistic may yield more informative phenotypes. However, while optimal node selection strategies remain unknown in this novel field of research, our selection strategy follows a simple and statistically rigorous heuristic – examine the node with the most significant test statistic for phenotype status.

4.4.6 Limitations

While we present results that have molecular concordance with the available body of scientific literature, some of our tissue expression enrichment estimates are based on *post-mortem* brain tissues which may not accurately reflect tissue expression patterns in the living brain [270]. We acknowledge that a more comprehensive understanding of what characterises gene expression profiles *in vivo* may alter the nature of a portion of our results. Further, we note that our sample size is small compared to other GWAS experiments which leaves us underpowered to detect variants with small effects on our endophenotype. As previously mentioned, our GWAS results also differ significantly when constrained to individuals of one ancestry, meaning it is difficult to have confidence in our results.

4.5 Conclusions

In summation, our autoencoder-derived phenotype captured neuroanatomical features relevant to aspects of AD pathology. Signficant genetic loci associated with our phenotype were long ncRNAs, whose role in neurodegenerative disorders is becoming better understood. However, our current sample size leaves us underpowered to detect true genetic effects of small to moderate effect sizes and our GWAS results vary greatly based on the exclusion of a small number of samples. While our method is of theoretical interest, further experiments in larger samples are required to test its utility.

Preamble to Chapter 5

This work was jointly supervised by Dr. Niamh Mullins and Dr. David Knowles and is currently in preparation for manuscript submission. All work was carried out by the candidate at the Icahn Scool of Medicine at Mount Sinai, New York.

Chapter 5

Deriving Causal Networks of Brain Imaging Phenotypes and Bipolar Disorder using Mendelian Randomization

5.1 Introduction

Our previous analyses were primarily concerned with investigating the genetics of sub-phenotypes of a brain disorder. However, our results do not allow us to investigate the causal relationship between brain disorders and neuroimaging phenotypes. This is pertinent given that reconciling our genetic analyses with current bodies of literature are contingent upon the assumption that there is a causal link between our extracted biomarkers and the overall conditions considered. Aside from our investigations which seek to establish frameworks by which to extract intermediate phenotypes from neuroimaging modalities, several candidate biomarkers for psychiatric conditions already exist [59, 71]. For example, the UK Biobank (UKB) is a large prospective epidemiological study of nearly 500 thousand patients. Recent efforts to expand the variables available for each patient have included a retrospective imaging collection strategy, which has so far resulted in just over 42 thousand patients having multi-modal imaging carried out [271]. Additionally, the UKB contains genetic information for every participant, affording researchers the potential to investigate the genetic basis of multiple traits. The linking of these distinct data sources gave rise to a study carrying out genome wide association studies (GWAS) of over 3 thousand imaging-derived-phenotypes (IDPs) across 33 thousand individuals with imaging data in the UKB [232]. The summary statistics for these GWAS

are publicly available, facilitating a host of post-GWAS analyses that can make use of single nucleotide polymorphism (SNP) β values.

Historically, it has been difficult to describe causality in common disorders using observational data due to physical and ethical barriers preventing randomised controlled trial environments. A branch of epidemiological methods, termed Mendelian randomization (MR), can help to facilitate causal reasoning from observational data sources using genetic information [272]. This can afford us the opportunity to estimate the causal relationships between multiple IDPs and a psychiatric outcome.

5.1.1 Mendelian Randomization

Modelling the causal relationship between two variables can be conceptualised as evaluating β in the following equation:

$$Y = X\beta + \gamma, \tag{5.1}$$

where X is a trait of interest, Y is the outcome, β is the effect of X on Y, and γ is a set of covariates that can confound the relationship between X and Y. In observational data, the set of variables in γ cannot be measured and controlled for in its entirety, which is one factor that can complicate causal inference.

Mendelian randomization attempts to bypass this limitation by leveraging the fact that alleles are randomly segregated at birth, are fixed at conception, and can be causal for certain traits. By this intuition, a SNP causal for trait X can remove certain types of unmeasured confounding from the set of variables γ. The classic example of this approach is a case study of populations of East-Asian ancestry that are heterozygous or homozygous for the minor allele of the *ALDH2* gene [273]. This mutation causes adverse reactions to alcohol consumption, meaning carriers are likely to consume less alcohol. This causal relationship between SNP and exposure, the phenotype whose causal effect we are interested in capturing, allows us to quantify the effect of alcohol consumption on a separate phenotype – in this case, blood pressure. As such, we can stratify individuals by their variant status and estimate the effect of alcohol consumption on their blood pressure, with the knowledge that their alcohol consumption values are causally influenced by the variant in question. This means that observing a relationship between a variant that causes lower alcohol consumption and high blood pressure implies a causal effect of alcohol on blood pressure values. Because we understand the genetic determinants of alcohol consumption in this cohort, a host of unmeasured lifestyle factors that would usually need to be controlled for do not need to be considered.

Mendelian randomization relies on a set of key assumptions to support this casual reasoning – firstly, the SNP cannot be associated with any confounding factors that may influence X and Y jointly; secondly,

the SNP must be predictive of X; thirdly, the SNP must be independent of Y when conditioning on X [274].

Mendelian randomization approaches have been extended to allow for the effect of multiple genetic variants and differing sample sources for exposures and outcomes. Most methods rely on combining SNP-outcome/SNP-exposure ratios. In a single-variant example, the following can be used to obtain a ratio estimate:

$$\hat{\beta}_{XY} = \frac{\beta_{Y_j}}{\beta_{X_j}} \tag{5.2}$$

For a causal variant indexed by j. This quantity represents the degree of change expected in the outcome Y per unit increase in the exposure X, expressed here as the degree of change expected in Y per unit increase in SNP j scaled by the degree of change expected in X per unit increase in SNP j. This can also be applied where SNP effects have been measured in different samples, meaning separate sets of summary statistics can be considered. Additionally, multiple ratio estimates can be combined to obtain one causal estimate [275].

5.1.2 Mendelian randomization of brain phenotypes

The availability of summary statistics from GWAS of over 3 thousand brain IDPs and numerous psychiatric conditions presents a unique opportunity to combine these data sources. The authors in [276] examined the causal effect of 587 IDPs on a panel of 10 psychiatric disorders and described several significant associations that survived rigorous multiple testing corrections. For example, decreased volume of the left accumbens was found to have a statistically significant positive causal effect on a bipolar disorder (BD). Other findings of interest include schizophrenia diagnosis causing decreased volume and surface area of the right pars orbalis.

While individual factors can have causal effects, it is unlikely that they work in isolation. Complex conditions such as BD are influenced by genetic risk factors, environmental stressors, and random chance. To model part of this complexity, we may consider using multiple individual causal relationships to build networks for better understanding BD.

5.1.3 Building causal networks using inspre

Creating causal networks from Mendelian randomization estimates is non-trivial for several reasons. The β values estimated using Mendelian randomization approaches are intended to capture the *total causal effect* (TCE) of an exposure on an outcome. This means that building a network from β_{TCE} values will capture the total effect of X on Y, including any mediators of that relationship. We instead seek to

capture the likely topology of a network of exposures and outcomes, meaning we seek the *direct causal effect* (DCE) matrix. This differs in the sense that we want to capture the causal effect of exposure X on outcome Y independent of any mediating factors, and we also seek to capture the effect of X and Y on their respective mediators.

Calculating a DCE matrix from TCE entries is a complex task – in effect, we seek the inverse of a partial-correlation-like matrix. The inverse in this context will yield estimates of the direct unmediated causal effect of one variable on another. Because we may have zero entries in the TCE matrix, calculating the exact inverse may not be possible as the determinant of the causal matrix may be zero, or the computation may be computationally intractable. To address this problem and generalise the derivation of the DCE matrix from a set of TCE entries, the authors of [277] developed *inspre*, a sparse-inverse regression algorithm that leverages the assumption that the underlying DCE matrix has many zeroes. Broadly, it applies a graphical lasso approach to calculate direct unmediated causal effects from a matrix of total causal effects. The full derivation of this method can be found in the supplemental note of [277].

Therefore, using *inspre* and the associated *bimmer* package [277], we seek to derive the DCE of a TCE input matrix of BD and a set of IDPs. We focus on BD given that its pathology is not thought to be marked by obvious neuroanatomical structural differences, making it of great interest to consider from a network-level approach. Examining the resultant network could offer the potential to understand individual causal relationships independent of included mediating factors and better understand the significance of graph topology in the context of a complex psychiatric disorder.

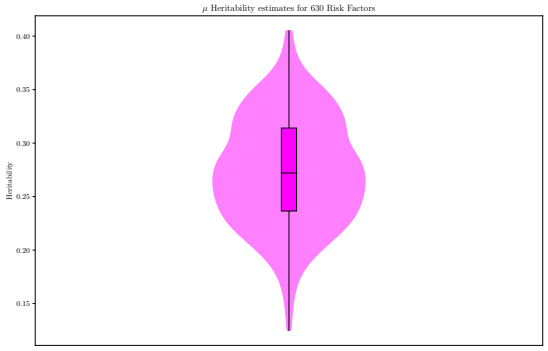
5.2 Methods

Initially, we sought to utilise the data provided from [276], but encountered several complicating factors which limited inference of a direct TCE matrix. Firstly, the set of 587 IDPs from [276] were chosen based on phenotypes that satisfied $\alpha < 0.05$, where α is the p-value of the genetic correlation (rg) test provided by the *ldsc* package between psychiatric disorder and IDP [278]. Genetic correlation seeks to capture how similar the genetic architectures of two traits are based on two sets of GWAS summary statistics. More details on this method can be found in [278]. Significant test statistics imply that there is high correlation between traits based on their summary statistics. This filtering strategy produced near-private sets of IDPs per respective psychiatric disorder, meaning that the same brain regions were not considered across disorders. Secondly, the inclusion threshold of $\alpha < 0.05$ was not corrected for multiple testing, which may introduce bias in phenotype inclusion. Thirdly, there is no guarantee that phenotypes with significantly overlapping genetic architectures are likely to have causal relationships with the outcome, so

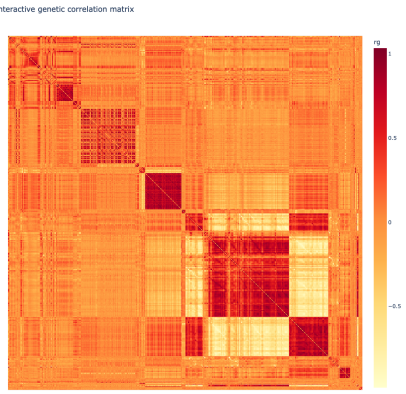
it may not make sense to condition phenotype inclusion based on this quantity. For these reasons, we sought to follow [277]'s methodologies to derive a novel TCE using a subset of 3929 IDPs given by [232].

5.2.1 Filtering

We downloaded 3929 sets of summary statistics from `https://open.win.ox.ac.uk/ukbiobank/big40/` and kept the phenotypes with more than 5 genome significant loci, leaving a set of 630 IDPs. We plotted the heritabilities of the remaining phenotypes to verify that significant amounts of phenotypic variation were explained by additive genetic effects, finding the median heritability to be 0.27 (Figure 5.1a). We next used *ldsc* to measure the rg values for every phenotype against each other and yielded a matrix of rg values per pair (Figure 5.1b). We filtered the set of 630 IDPs based on rg values, whereby $rg > 0.85$ caused a pair to be flagged for removal. The phenotype in the offending pair with less genome wide significant loci was removed. This was performed to mitigate the risk of phenotypic redundancy, and is akin to multicollinearity check, resulting in a final set of 159 phenotypes plus bipolar disorder.



(a) Heritability estimates of 630 IDPs with greater than 5 genome wide significant loci estimated using *ldsc*. The estimate is plotted on the y-axis.



(b) rg estimates for every phenotype against each other and bipolar disorder. Magnitude is denoted by the intensity of the color, whereby pairs of phenotypes constitute the entries of the matrix.

Figure 5.1

5.2.2 MR experiments

Next, we used the protocol outlined in [276] to carry out bidirectional MR tests of every pair of phenotypes. Firstly, we clumped SNPs in every phenotype with an r^2 threshold of 0.05, a clump window of 500kb,

and a p-value threshold of $5e-6$. We next applied the *RadialMR* package to remove SNPs violating MR assumptions in every phenotype, followed by pruning of SNPs based on the p-value of the intercept term in an Egger regression test of pleioptropy [279, 280]. With our pruned SNPs for each phenotype, we carried out MR experiments using the *TwoSampleMR* package for every pair of phenotypes in the forward and reverse direction, yielding $\binom{160}{2} \cdot 2$ MR tests [281]. We applied 5 MR methods: simple mode, weighted mode, weighted median, inverse variance weighted approach, and MR Egger. Each method varies in their details, but they broadly follow a procedure similar to 5.2. Briefly, the simple mode procedure estimates the modal causal effect for a group of SNPs; weighted mode scales this modal causal effect by the inverse variance of the weights used to compute the mode; the weighted median takes the median causal effect weighted by the inverse variance of the weights; the inverse variance weighted procedure combines ratio estimates weighted by the inverse of the variance of the estimate; finally, Egger regression combines ratio estimates of multiple SNPs using a meta-regression framework including an intercept parameter for estimates of pleiotropy. We corrected every MR estimate p-value for multiple testing using the Benjamini-Hochberg method and visualised significant pair overlap using an UpSet plot. 9 Exposure-outcome pairs containing BD with FDR-corrected p-values < 0.01 in at least 2 methods were visualised in forest plots. For DCE derivation, we used the estimates from the weighted mode estimator, owing to its low false positive rate and theoretical ability to handle correlated horizontal pleiotropy by taking the mode of β values [277].

5.2.3 Network experiments

We constructed a TCE matrix of every phenotype's causal effect on each other in two MR methods, yielding a 160×160 matrix. We applied *inspre* through the *bimmer* package with a 10-fold cross validation schema, retaining the network with optimal graph stability under resampling. This is motivated by the intuition that stable graphs are usually yielded for smaller values of λ, the penalty parameter, under random resamplings of input data. For more details on this this metric, see the supplemental note of the bimmer paper [277]. We developed a diagnostic plot to visualise the differences between the derived DCE and the input TCE. Specifically, we sought to ensure that input zero values were not significantly inflated by the graphical lasso regression and given large values by *inspre*. A degree of inflation is expected owing to the bias terms in iterative graphical lasso operations, but offending input zero-value TCE entries outside of 1.5 standard deviations in the DCE matrix were set to zero for subsequent experiments.

We visualised networks from both tools using *Cytoscape*, and retained the weighted mode estimator values for further testing owing to its diagnostic plots [282]. We stratified phenotypes on their designation as either describing gray matter structural features or describing diffusion tensor features. Gray matter

structural features include measures such as volume or surface area, whereas diffusion features include metrics relating to white matter microstructural hindrances on molecule motion in specific regions. We fit the following models to examine if the BD DCE coefficients were significantly different in structural and diffusion phenotypes:

$$\left|\beta^{i}_{BD}\right| = \beta D + \epsilon \tag{5.3}$$

where D is the phenotype designation and $|\beta_{BD}|$ is the absolute BD coefficient for that phenotype indexed as either an exposure or outcome by i.

By setting all elements within 1 standard deviation of the mean to zero, we obtained the degree of each phenotype in the network. We ran the following linear models using *statsmodels* to examine the differences in outcome/exposure degree between structural and diffusion phenotypes:

$$Y = \beta D : i + \epsilon \tag{5.4}$$

where Y is the degree and D is the phenotype designation (coded as 0 or 1) stratified by i, denoting the phenotype as either outcome or exposure. Because every phenotype can be an exposure or outcome in the network it was desirable to understand how status as a diffusion phenotype affected the number of connections in both directions (ie., row-wise sum for exposure degree and column-wise sum for outcome degree).

Next, we fit the following models:

$$Y_i = \beta \left|\beta_{BD}\right| : D + \epsilon \tag{5.5}$$

where Y_i is the exposure/outcome degree indexed by i and $|\beta_{BD}|$ is the absolute centered DCE BD estimate across phenotype designation category status D. This model examined if differences in exposure or outcome degree were correlated with the absolute magnitude of DCE BD coefficients (in the forward or reverse direction) in diffusion and structural phenotypes.

We examined whether or not diffusion phenotypes were enriched in the top 20 exposures with the greatest absolute effect on bipolar disorder, calculating the probability mass function of a hypergeometric distribution parameterised by our inputs. This gave us the p-value for the probability of observing a specific number of diffusion phenotypes from a selection of 20 against a background of 159 phenotypes total (107 of which were diffusion-based). We also plotted networks of the top 20 exposures acting on BD and the top 20 outcomes causally affected by BD, where we ordered graph orientation by degree. For calculating the degree of a node, we set all entries within one standard deviation of the mean β_{DCE} value to zero, and ordered our network from highest degree to lowest from left to right starting at the BD node.

For every phenotype in the network, we sought to trace its most influential causal path to a certain depth and examine the phenotypic designations of its members. We recursively extracted the most influential exposure for every phenotype to a depth of 5 nodes, noting the phenotype's categorisation as either structural or diffusion-based. We summed the number of occurrences of diffusion phenotypes and ran a 1-sample proportion test with continuity correction to test the null hypothesis that the proportion of structural and diffusion phenotypes in respective significant causal paths was equal to 0.5. We compared the real counts of diffusion phenotypes to 160 randomly generated lists of binomial random variables of length 5 and conducted a t-test to examine if the real count distribution differed significantly from the null.

Further, for every pair of phenotypes in the network, we plotted its $\beta_{exposure}$ against its $\beta_{outcome}$ to test for the presence of feedback loops. Pairs containing phenotypes with differing signs $> 1\sigma$ outside the global mean that feature in the top 20 absolute interactions with BD were investigated by plotting subnetworks. Similarly, the $\beta_{exposure}$ effect of BD on every phenotype was plotted against the same phenotype's $\beta_{exposure}$ effect on BD and pairs of interest were plotted using *Cytoscape*.

5.3 Results

5.3.1 Genetic correlation matrix of IDPs and BD

Our $\binom{630}{2}$ rg tests yielded a matrix of every phenotype against each other using *ldsc* (Figure 5.1). This revealed widespread patterns of positive genetic correlation between brain regions and was used to inform subsequent filtering of imaging-derived phenotypes.

5.3.2 Causal relationships between BD and brain regions

We found ≈ 2000 exposure-outcome pairs identified by all 5 considered methods as significant after FDR correction of over 25 thousand MR tests ($FDR < 0.01$, Figure 5.2). We found 9 exposure-outcome pairs containing BD as a term that were found to be FDR significant at $P < 0.01$ in at least 2 MR methods, including the three pairs highlighted in Figures 5.3, 5.4, and 5.5. These pairs include the effect of BD on the area of the lateral orbitofrontal cortex in the left hemisphere ($\hat{\beta} = -0.0897 \pm 0.03$), the effect of the surface area of the left hemispheric anterior transverse sulcus on BD ($\hat{OR} = 1.24 \pm 0.05$), and the effect of the mean intracellular volume fraction (ICVF) in the pontine crossing tract's on BD ($\hat{OR} = 1.25 \pm 0.12$) (Figures 5.3, 5.4, and 5.5). Here, $\hat{\beta}$ is the average β value across the five methods with the standard deviation of this mean denoted after the $\pm$ sign. Each of the five methods have their own associated standard errors

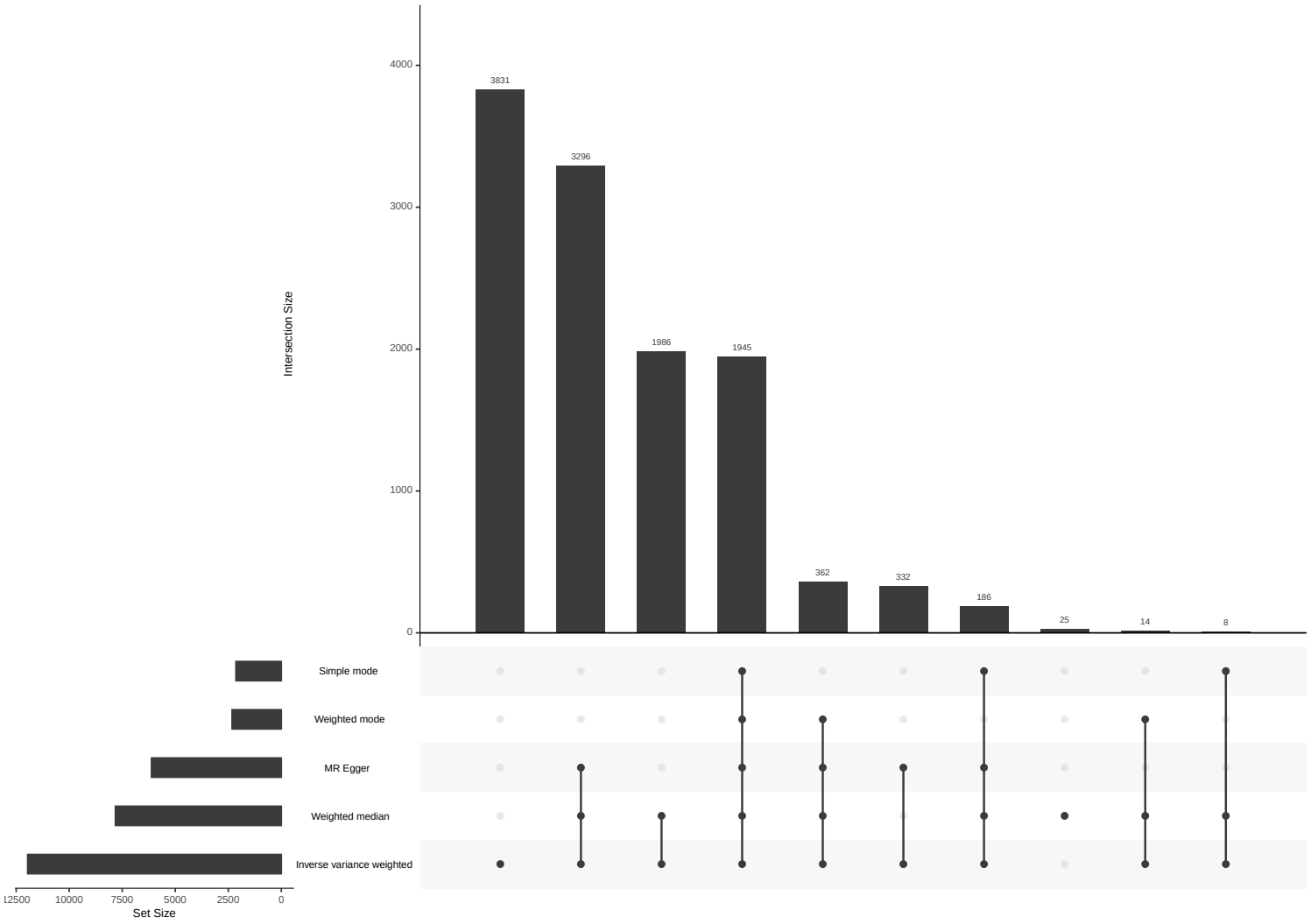


Figure 5.2: UpSet plot of all method total set size and intersection size for FDR-corrected pair p-values less than 0.01. Here, the bottom panel dotplot represents the set in question, and the barplot in the top panel represents the size. Individual dots refer to private sets, ie, the number of p-values identified by only the inverse variance weighted method is 3831, and the overall set size is larger. The total set sizes are denoted on the left barplot.

attached to the β values which were used to derive confidence intervals in Figures 5.3, 5.4, and 5.5. Five of these 9 FDR-significant exposure-outcome pairs feature BD as an outcome, including one significant bidirectional relationship with the mean orientation dispersion (OD) index in the left cerebral peduncle. The only association of the 9 FDR-significant pairs whose confidence intervals do not overlap with 0 (β) or 1 (OR) in all methods is the effect of BD on the surface area of the lateral orbitofrontal cortex in the left hemisphere (Figure 5.4), although not all method p-values reached our FDR-corrected significance threshold.

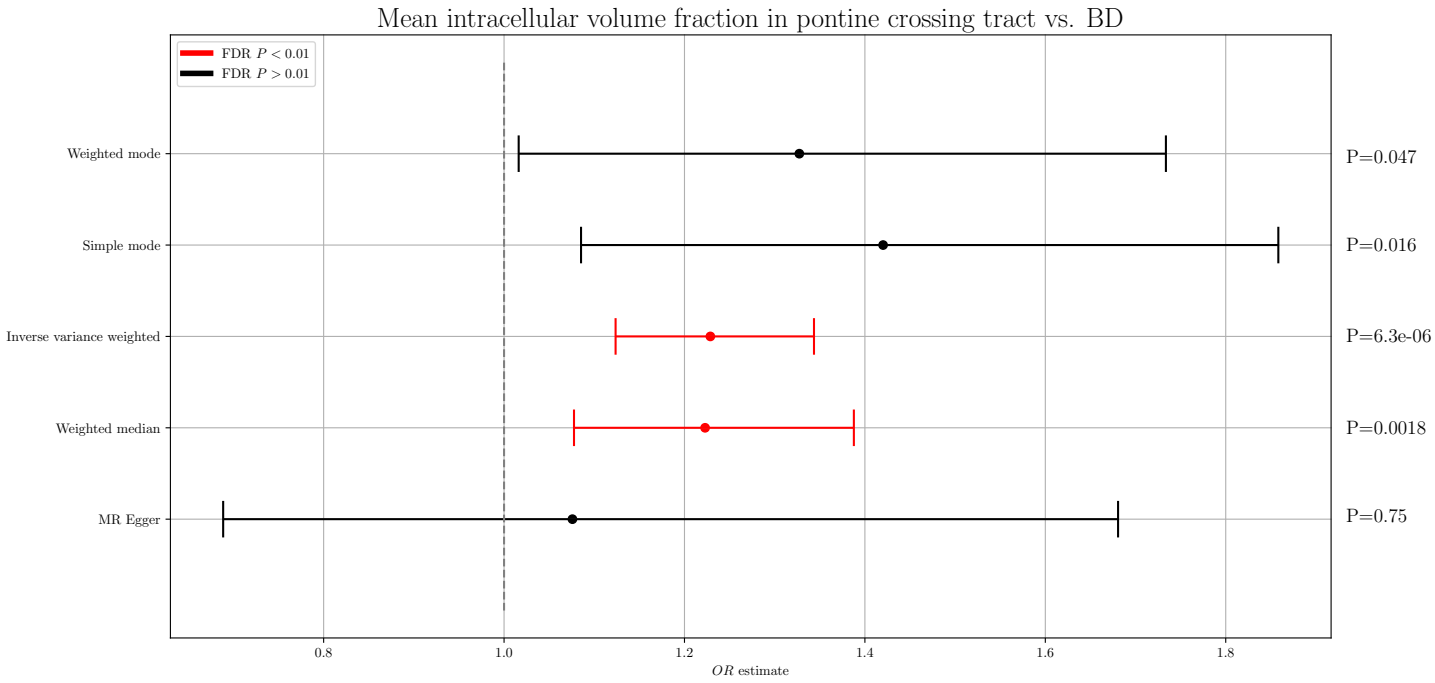


Figure 5.3: Forest plot of the effect of increased mean intracellular volume fraction in the pontine crossing tract on BD.

5.3.3 Network dynamics of IDPs and BD

We considered the weighted mode DCE for further network investigation owing to its competitive false positive rate in simulations [277] and its custom diagnostic plot performances (Figure 5.6). We observed that zero values in the TCE matrix mostly remained zero in the DCE matrix, which can be seen by examining Figure 5.6. Figure 5.8 shows the top 20 exposures with the greatest absolute effect on BD with edge bundling applied. Edge bundling attempts to reduce visual clutter by bending edges together where

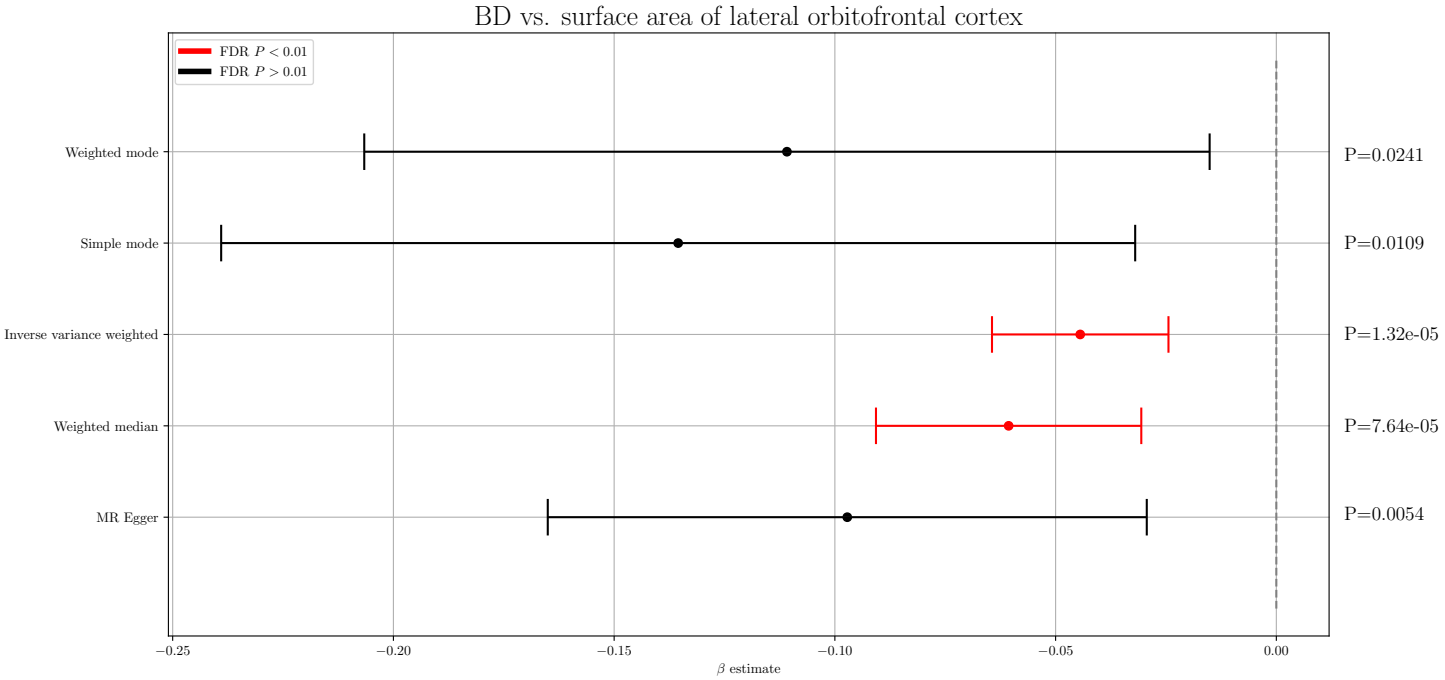


Figure 5.4: Forest plot of the effect of BD on lateral orbitofrontal surface area.

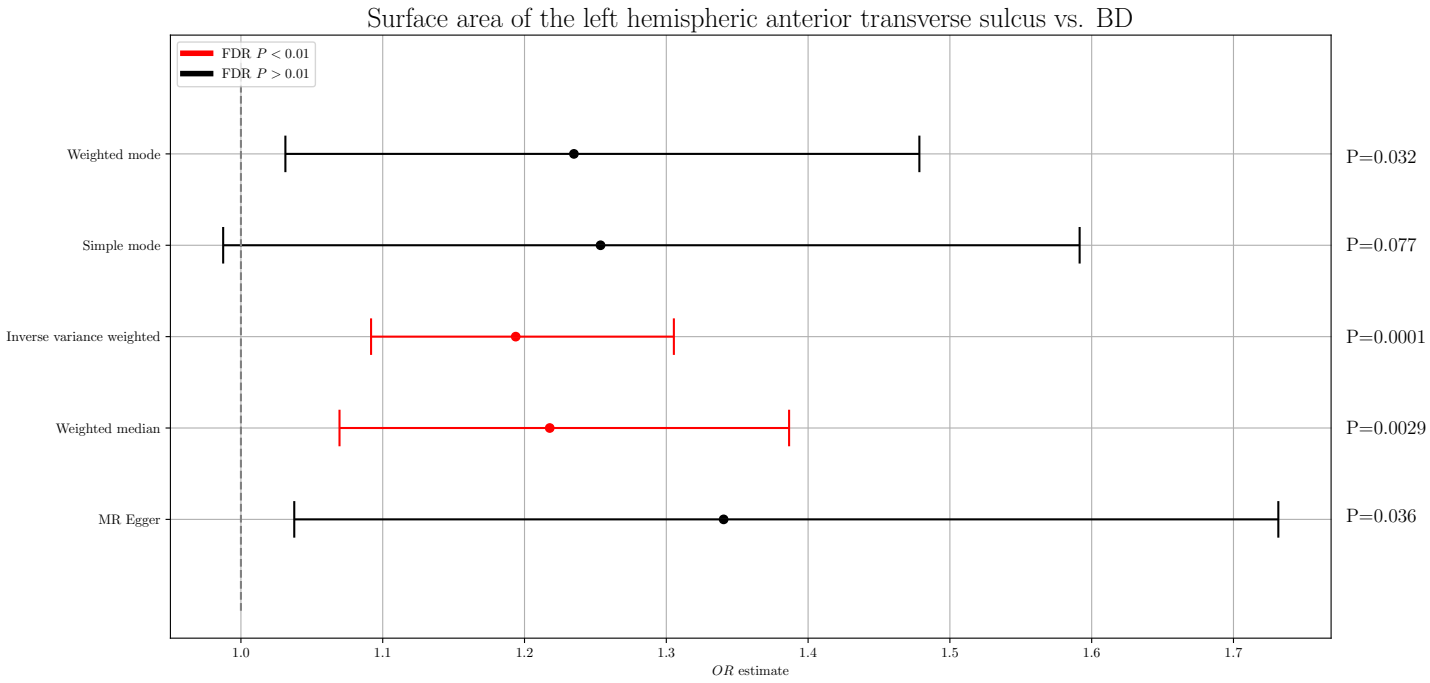


Figure 5.5: Forest plot of the effect of increased left hemispheric sulcal surface area on BD.

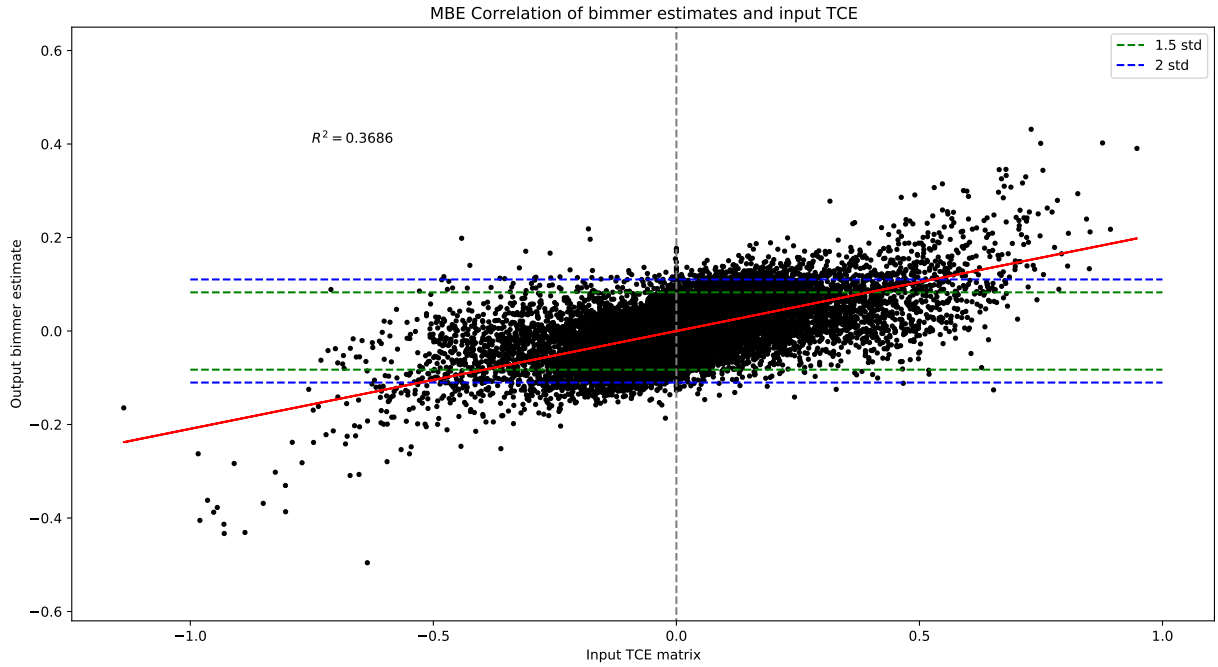


Figure 5.6: Scatterplot of the TCE (x-axis) and the *inspre*-transformed output DCE y-axis. Standard deviation lines are drawn at $1.5 \times \sigma$ and $2 \times \sigma$ in green and blue dashed lines respectively.

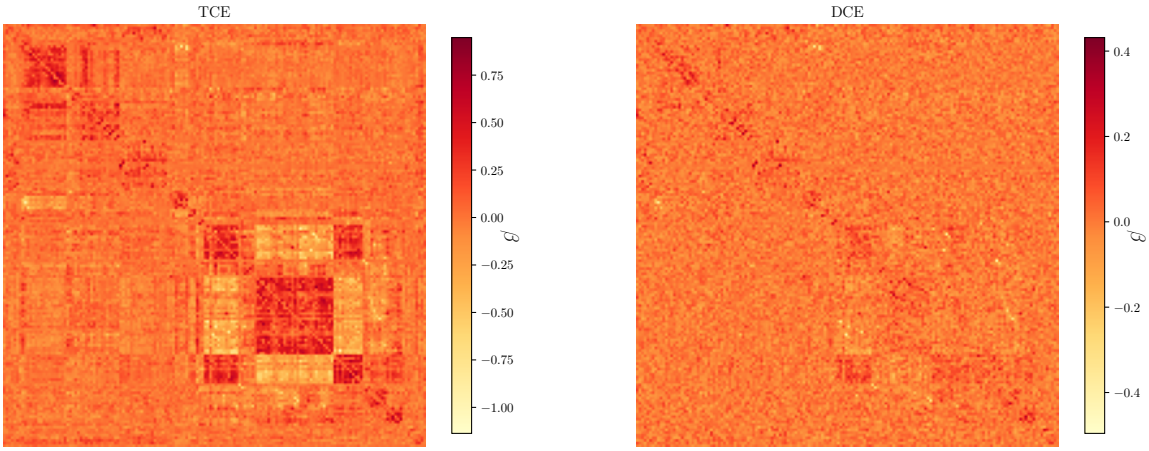


Figure 5.7: Visualisation of the TCE (left) and the *inspre*-transformed output DCE (right).

exposures affect similar outcomes, creating parsimonious visual representations of paths. Amongst these exposures, 12 were diffusion tensor phenotypes ($P = 0.67$, hypergeometric test of enrichment). Figure 5.9 shows BD's top 20 highest absolute weight outcomes with edge bundling; that is, the IDPs BD exerts the strongest absolute causal effects on in the network. This subset contained 15 diffusion phenotypes ($P = 0.13$, hypergeometric test of enrichment).

We found no significant relationships from the regression detailed in 5.3 ($P = 0.73, p = 0.232$ respectively), meaning that BD was not observed to have significantly differing effects on diffusion tensor or gray matter structural phenotypes in absolute magnitude. Conversely, diffusion tensor phenotypes do not have a significantly different absolute effect on BD when compared to gray matter structural phenotypes (Figure 5.10).

We found diffusion tensor outcomes were on average affected by 3 more phenotypes than gray matter structural outcomes ($P = 0.0171$, Figure 5.11), while the same effect was not found for diffusion exposures ($P = 0.4157$). There were no significant differences in global degree between gray matter structural and diffusion tensor phenotypes ($P = 0.283$). We found that structural outcomes affected with greater

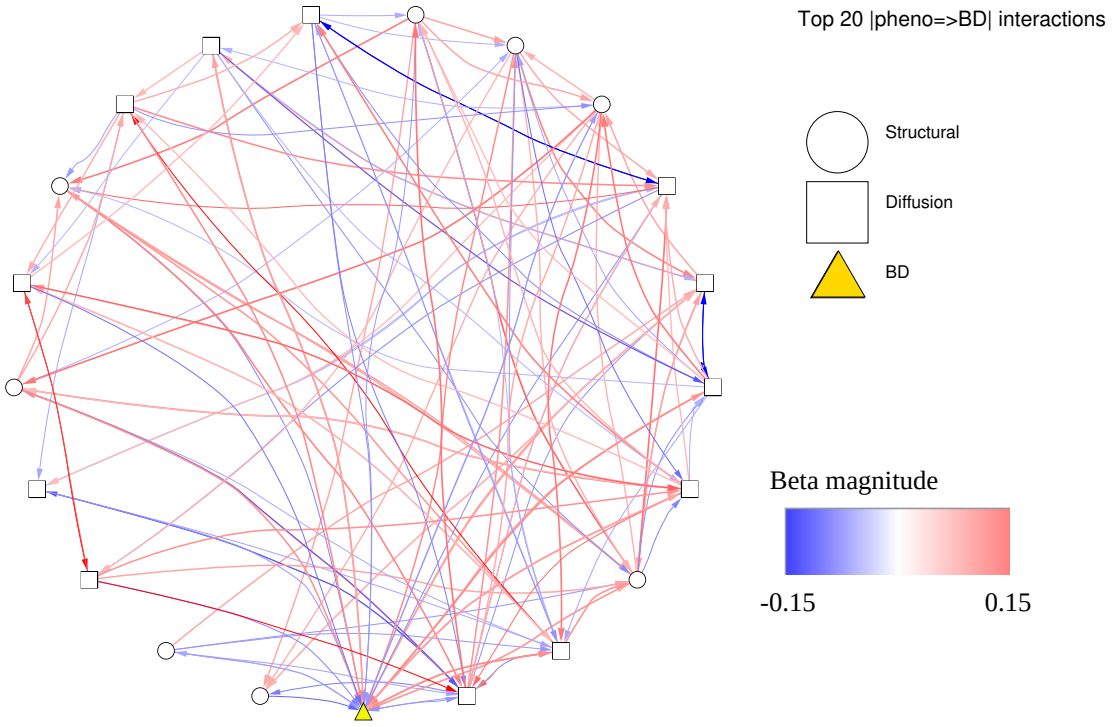


Figure 5.8: The top 20 IDPs with the largest absolute effect on BD ordered by degree from left to right starting at BD (high to low). This is to say that the IDP immediately to the right of BD has the highest degree apart from BD; the IDP immediately to its right has the next highest degree, etc. Color intensity denotes the magnitude of the β effect.

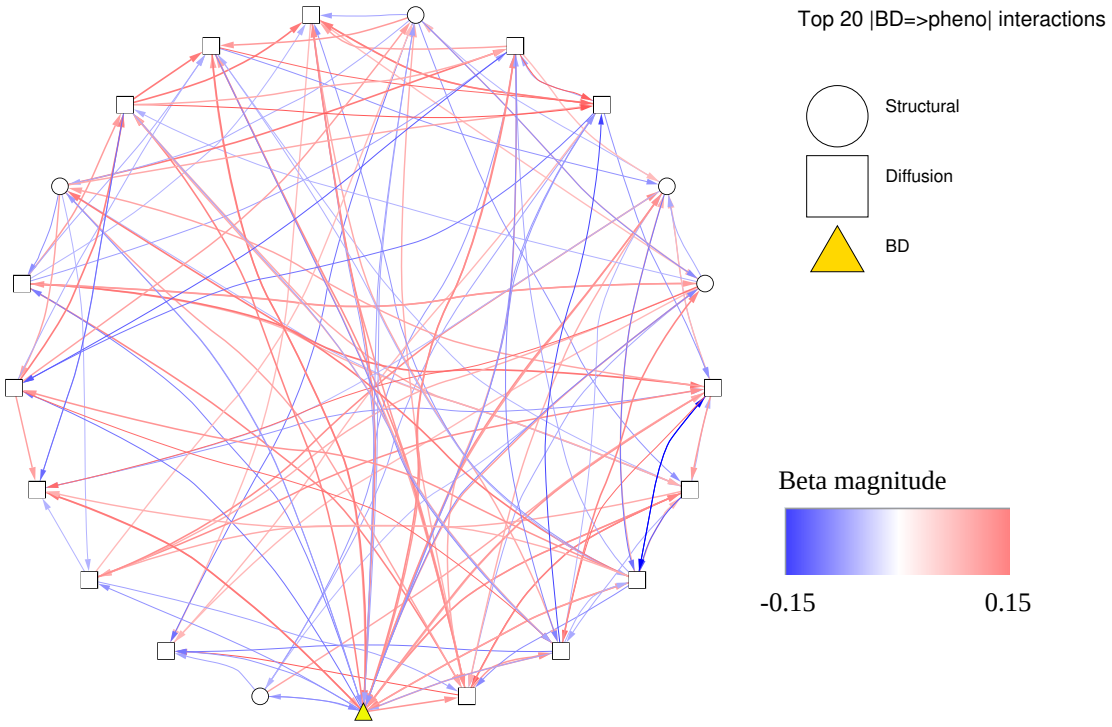


Figure 5.9: The top 20 IDPs that BD affects ranked by absolute β. Entries are sorted from left to right based on degree starting at the BD node (high to low) in the same manner as Figure 5.8. Color intensity denotes the magnitude of the effect.

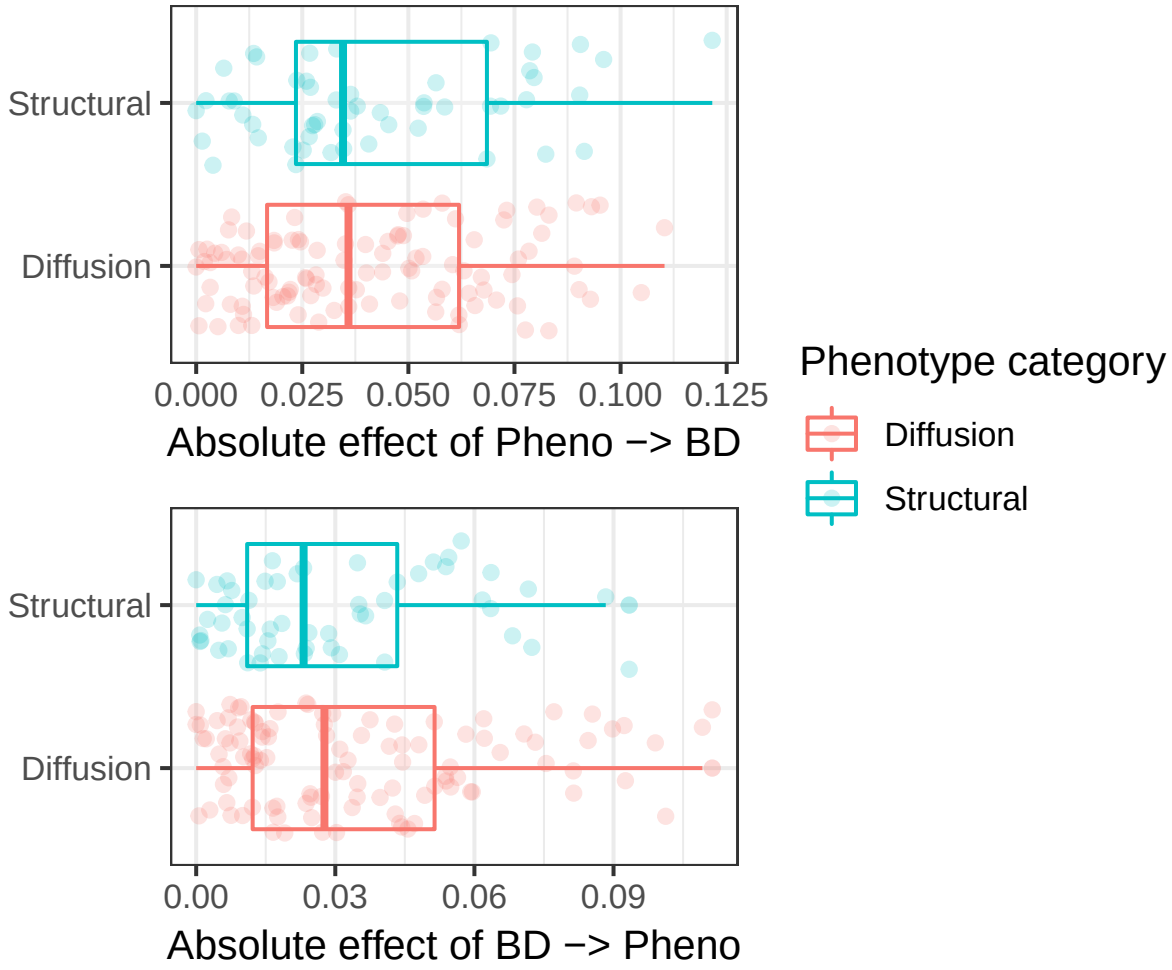


Figure 5.10: Boxplots of $|\beta|$ values of exposures and outcomes stratified by phenotype categorisation.

magnitude by BD are affected by 3 more phenotypes on average compared to other structural phenotypes ($P = 0.008$, Figure 5.12), while the same effect was not observed within diffusion outcomes ($P = 0.07$) . However, an interactions effect model of outcome degree against absolute BD effect on the phenotype multiplied by phenotypic category yielded a non-significant p-value (0.244), suggesting that this result does not hold when considering all phenotypes affected by BD. We found no significant correlations between node degree and larger phenotype effects on BD in stratified models or interaction models (Figure 5.12).

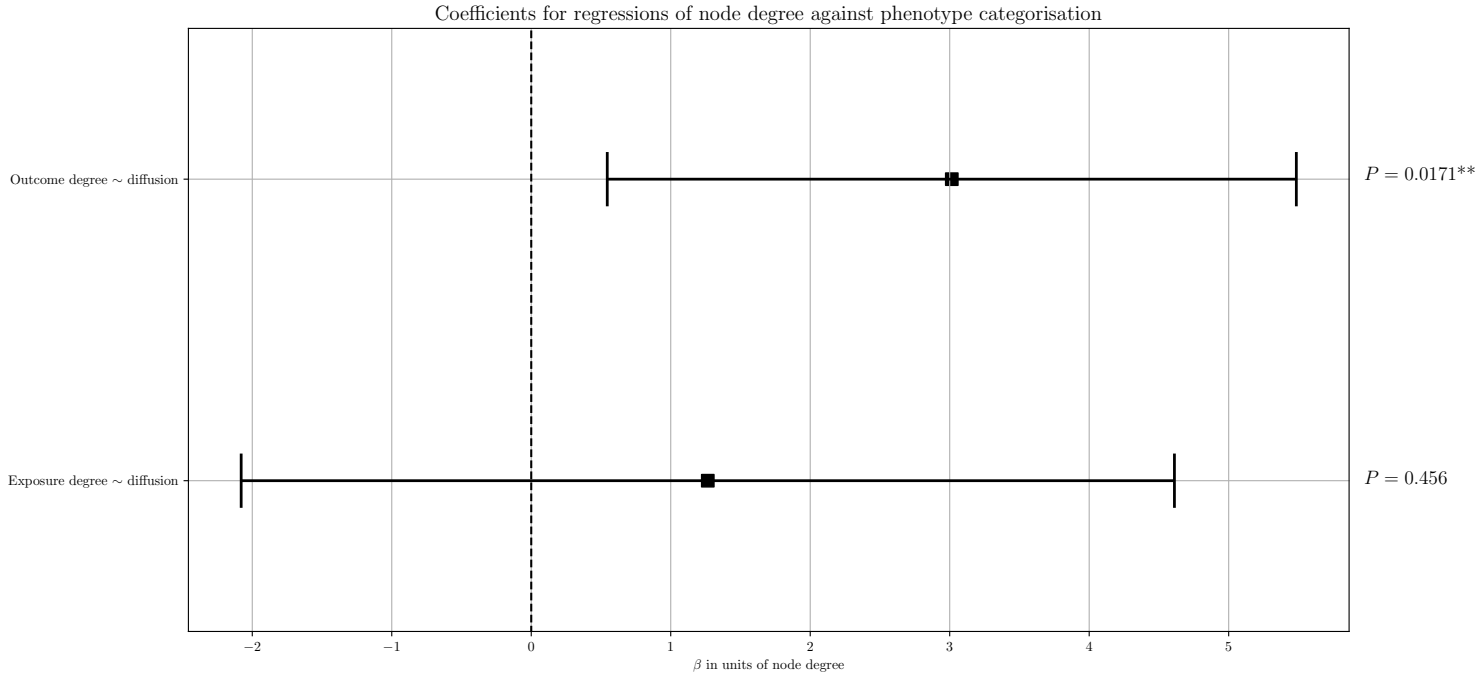


Figure 5.11: Coefficients from a regression of node degree against phenotype categorisation in exposures and outcomes. The β coefficient and aassociated 95 % confidence intervals are displayed on the x-axis, with the y-axis denoting the relationship tested.

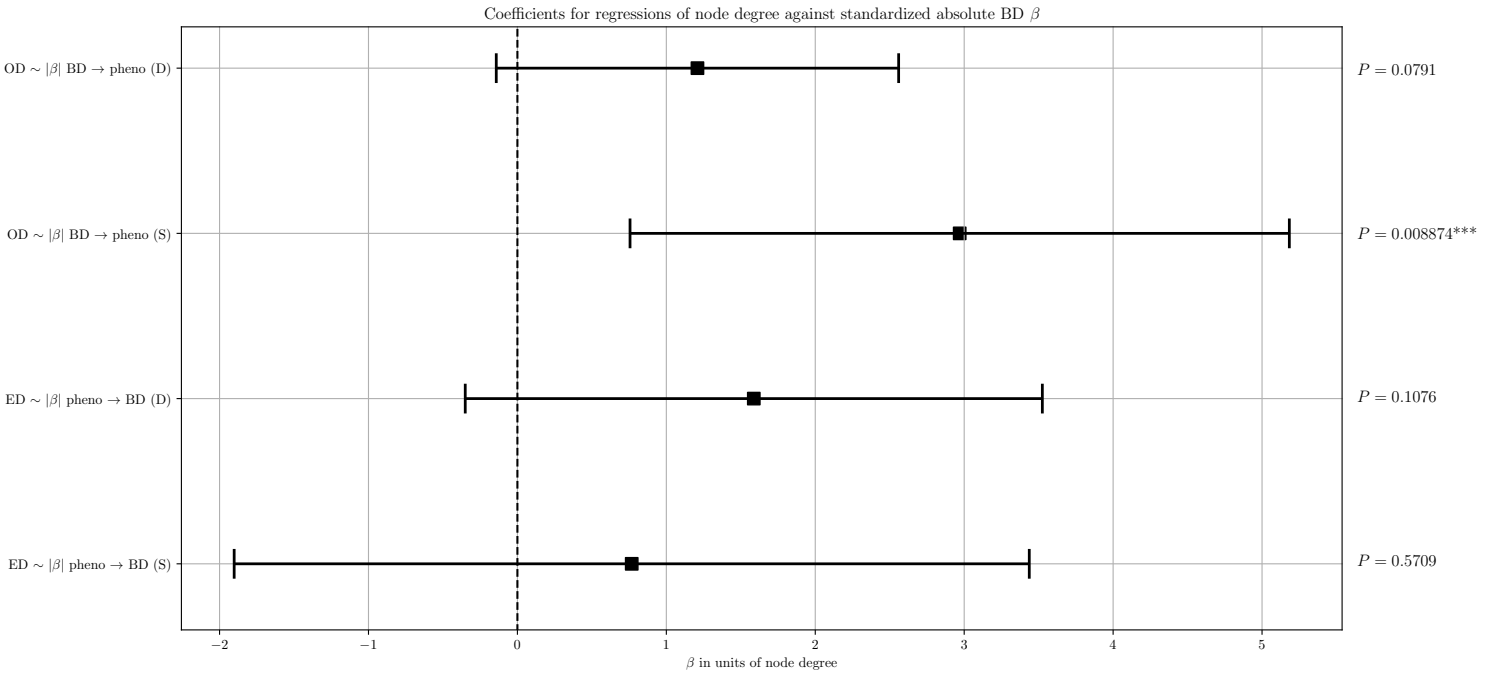


Figure 5.12: Coefficients from regressions of node degree against absolute β^{BD}_{DCE} ($|\beta|$) in exposures and outcomes stratified by phenotype categorisation. Here, ED stands for exposure degree (number of outgoing causal effects outside of one standard deviation in magnitude), OD stands for outcome degree (number of incoming causal effects outside of one standard deviation in magnitude), (S) stands for structural (gray matter structural phenotypes), and (D) stands for diffusion (white matter microstructural phenotypes).

For every phenotype, we identified its highest weight exposure. We repeated this operation 5 times to create lists of phenotypes and their highest weight exposures. For example, phenotype 1's highest weight exposure may be phenotype 2, and phenotype 2's highest weight exposure may be phenotype 3, and so on. This was carried out non-redundantly to ensure that causal path lists contained unique entries. We took note of the phenotypic designation of each member of the causal path. We found that diffusion phenotypes were over-represented in the 160×5 ranked paths, occurring 515 times out of 800, reaching significance in a test of equal proportions with continuity correction ($P = 5.6e - 16$). When compared to 160 simulated causal paths, a t-test indicated significant differences between the observed causal paths and simulations ($P = 9e - 4$, Figure 5.13).

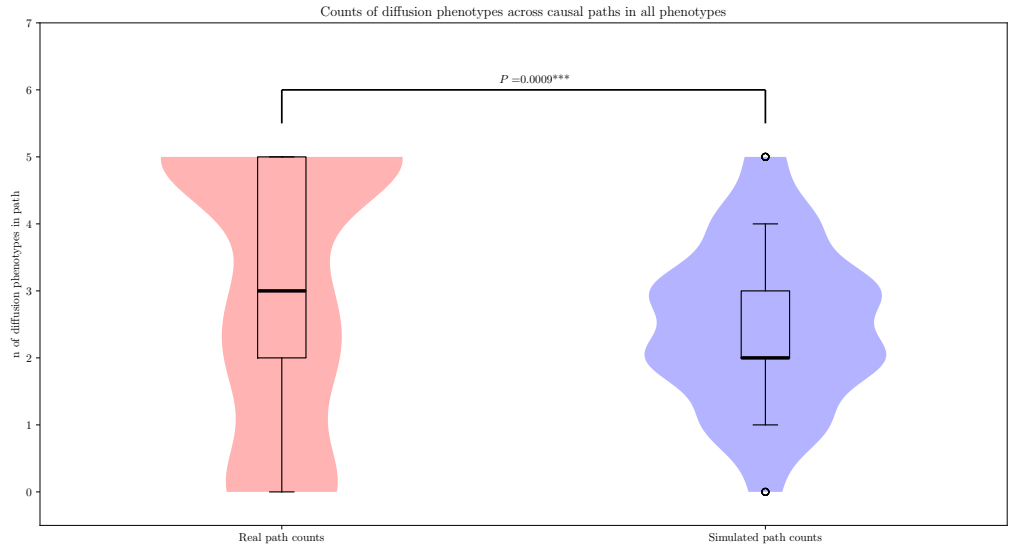
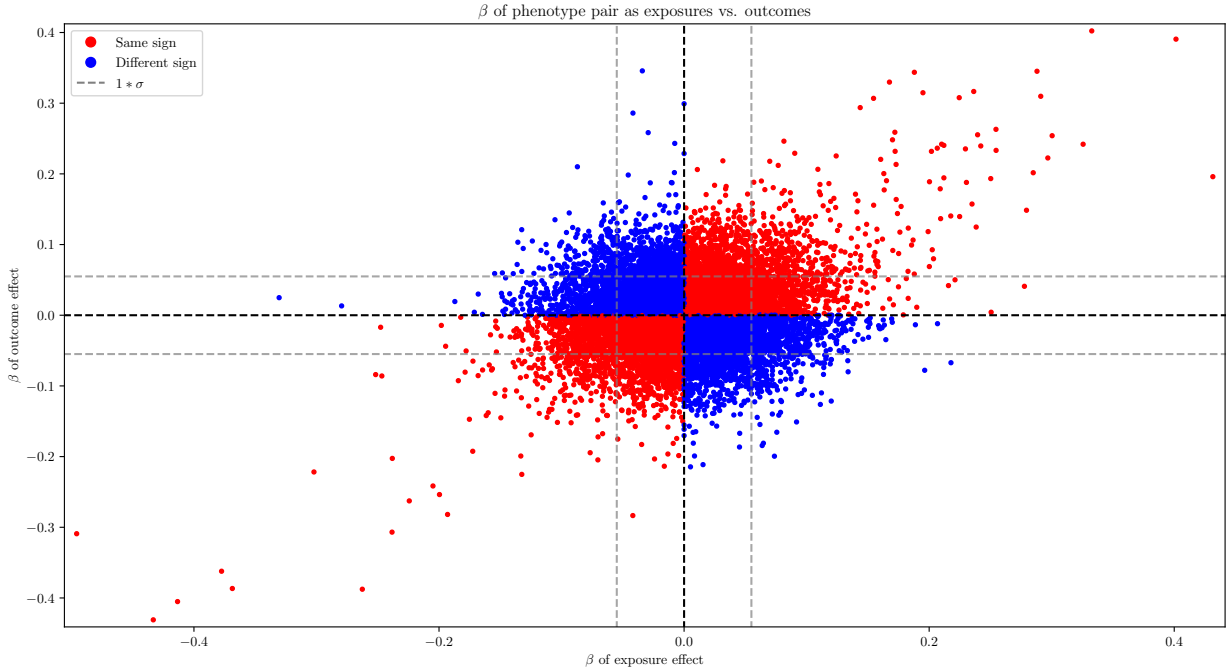


Figure 5.13: Real recursive causal path sums for all phenotypes (left) vs. simulated list sums (right) with associated p-value of test statistic from t-test.

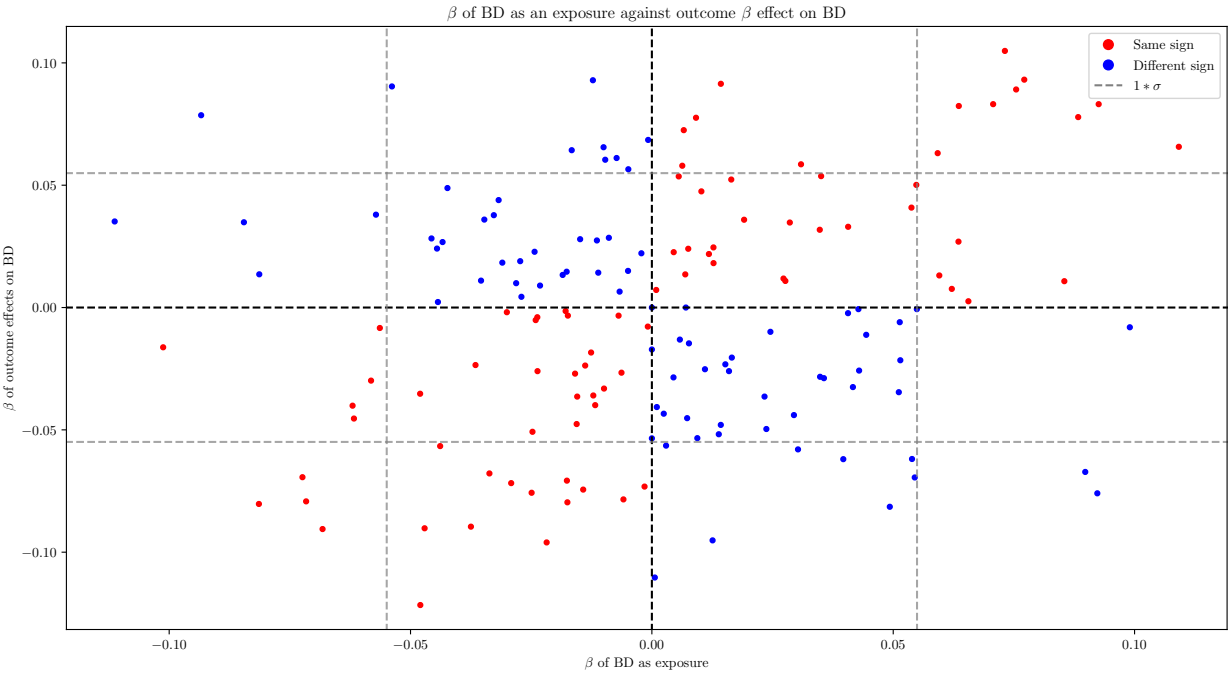
We plotted the effect of every phenotype on every other phenotype for all exposure-outcome pairs (Figure 5.14a). We noted that half of all pairs (50.97%) had the same direction of effect, while the other half had differing signs when exposure/outcome status was swapped. We observed that 3% of pairs had differing directions of effect with magnitudes outside 1σ of the mean. In investigating this subset, we found that 185 out of 383 pairs with differing directions of effect $> 1\sigma$ were diffusion-diffusion pairs (two white matter microstructural phenotypes), 152 were mixed pairs, and 46 were structural-structural pairs (two gray matter structural phenotypes). Further, we found 16 pairs of this subset where both phenotypes were in the top 20 highest absolute weight outcomes of BD (BD=>pheno) or the top 20 highest absolute weight exposures on BD (pheno=>BD). We visualised 1 pair and their relationships with BD in subnetworks. Figure 5.16 shows the mean intracellular volume in the right tapetum and the mean FA in the right anterior corona radiata.

We observed that half of direct BD interactions were characterised by bidirectional relationships of consistent effect direction, meaning that $sign(\beta_{ij}) = sign(\beta_{ji})$ (Figure 5.14b, 80/160 pairs). Three pairs of BD interactions had differing signs $> 1\sigma$ outside the mean. We investigated these interactions by visualising their relationships with BD and each other in subnetworks (Figure 5.17). Briefly, we visualised

the effect of the intensity of the right hemispheric cerebellum cortex, the mean diffusivity in the right external capsule, and the mean ICVF in the fornix. We also found that the absolute effect of phenotypes on BD was greater than the absolute effect of BD on phenotypes in a two-sided t-test of means ($P = 0.0047$, Figure 5.15). However, this difference was non-significant when considering non-absolute β values.



(a) Scatterplot of $\beta_{exposure}$ vs $\beta_{outcome}$ for every pair in the network; ie, the effect of phenotype i in phenotype j plotted against the effect of phenotype j on phenotype i where $i, j \in P; i \neq j$, and P is the set of all phenotypes.



(b) Scatterplot of all IDP's effect on BD against their respective effect on IDPs.

Figure 5.14

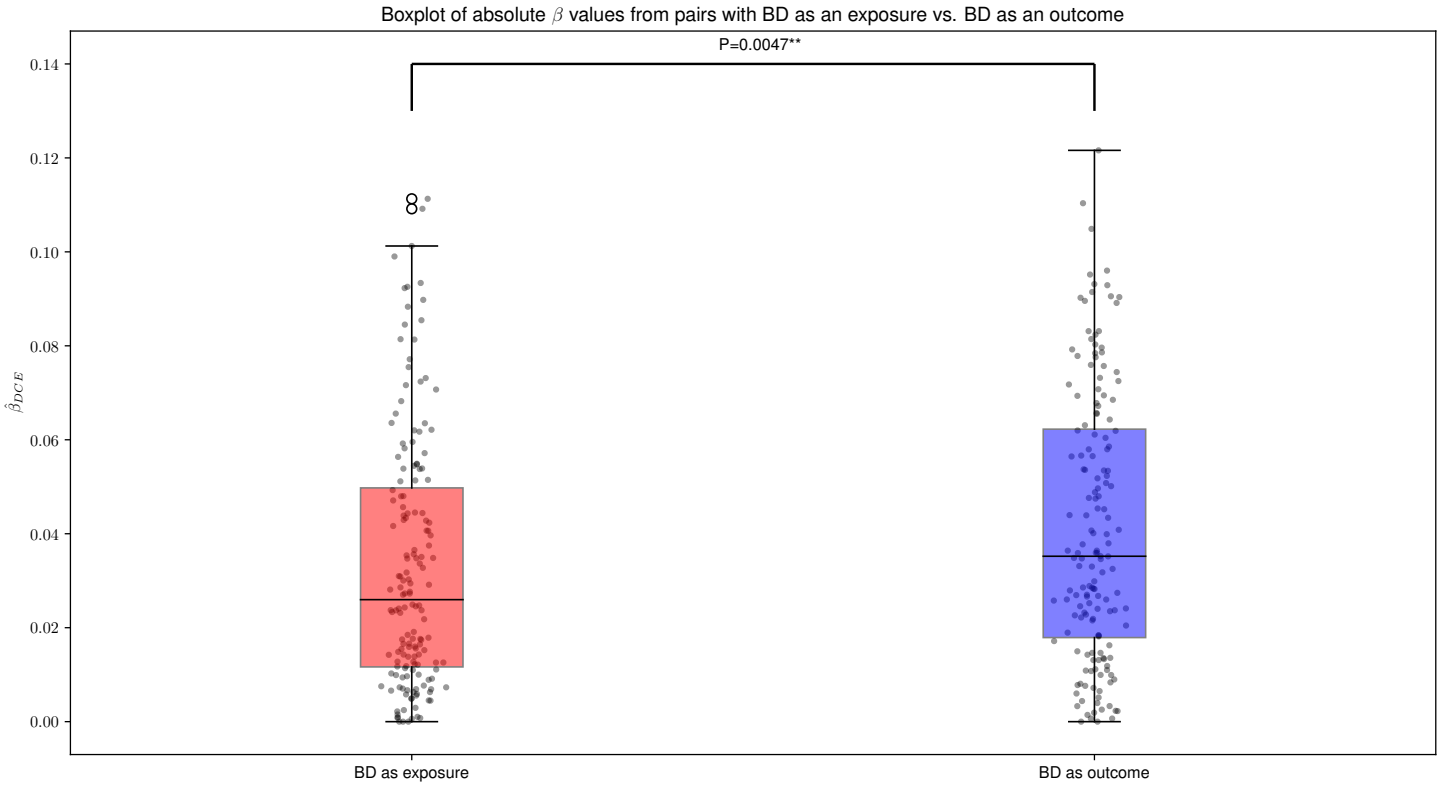


Figure 5.15: Boxplot of absolute β values with BD as an exposure vs. absolute β values of phenotypes effect on BD (where BD is an outcome). P-values were calculated using a two-sided related samples t-test.

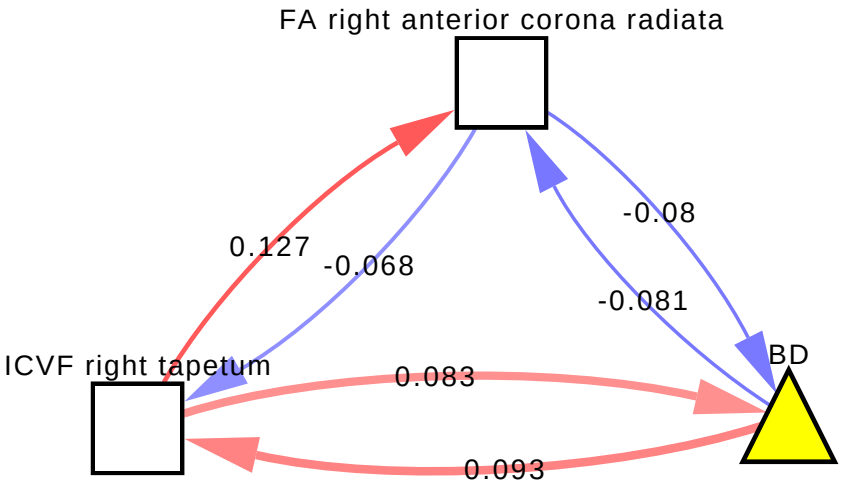


Figure 5.16: DCE subnetwork of BD, the mean ICVF in the right tapetum, and FA in the anterior corona radiata in the right hemisphere. Edge values are β values following the same schema as previous network figures, where red denotes positive and blue denotes negative.

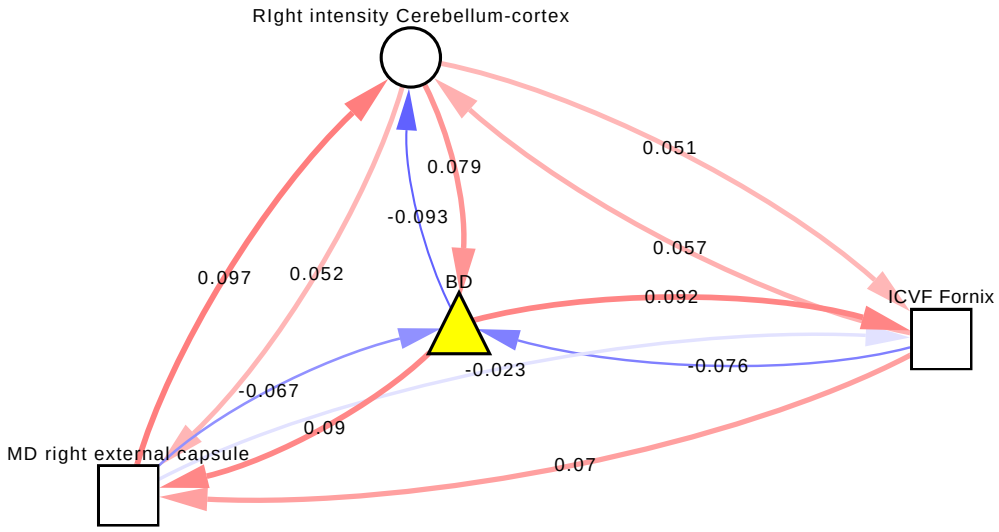


Figure 5.17: Subnetwork of BD interactions with bidirectional relationships significantly differing in maginitude and sign. Edge values are β values following the same schema as previous network figures, where red denotes positive and blue denotes negative.

IDP	$\lvert TCE\beta_{xy}\rvert$	$\lvert TCE\beta_{yx}\rvert$	$\lvert DCE\beta_{xy}\rvert$	$\lvert DCE\beta_{yx}\rvert$	Status
MSOCC	**-0.136**	-0.070	**-0.111**	0.035	Top BD => pheno IDP
RCI4	-0.091	0.009	-0.101	-0.016	Top BD => pheno IDP
ARIC	-0.088	0.179	-0.093	0.078	Top BD => pheno IDP
ITFRLT	-0.091	0.030	-0.084	0.034	Top BD => pheno IDP
FACRR	-0.058	0.030	-0.081	-0.080	Both top BD exposure and outcome
ITFRLCI	-0.057	0.021	-0.081	0.013	Top BD => pheno IDP
AGV4	-0.111	-0.082	-0.072	-0.069	Top BD => pheno IDP
ARAS	-0.085	-0.197	-0.071	-0.079	Both top BD exposure and outcome
TRVL	-0.081	-0.208	-0.068	-0.090	Both top BD exposure and outcome
ALAS	-0.097	0.210	-0.053	0.090	Top pheno => BD IDP
HRVH	-0.044	-0.194	-0.047	**-0.121**	Top pheno => BD IDP
OCPL	-0.083	-0.149	-0.047	-0.090	Top pheno => BD IDP
MCPL	4e-6	-0.073	-0.037	-0.089	Top pheno => BD IDP
DRAI	<1e-06	-0.160	-0.021	-0.096	Top pheno => BD IDP
PRAT	-0.061	-0.138	-0.017	-0.079	Top pheno => BD IDP
IDPIIR	-0.008	0.144	-0.012	0.092	Top pheno => BD IDP
IMCP	0.025	-0.044	6e-4	-0.110	Top pheno => BD IDP
IICPR	0.033	-0.005	0.012	-0.095	Top pheno => BD IDP
WLIT	0.057	0.188	0.014	0.091	Top pheno => BD IDP
MPLOICL	0.028	**-0.217**	0.049	-0.081	Top pheno => BD IDP
PLAS	0.067	0.127	0.063	0.082	Top pheno => BD IDP
OSLFL	0.042	0.126	0.070	0.083	Both top BD exposure and outcome
FMCP	0.080	0.211	0.073	**0.104**	Both top BD exposure and outcome
ITFRVCV	0.127	0.103	0.075	0.089	Both top BD exposure and outcome
LSLFL	0.059	0.092	0.077	0.093	Both top BD exposure and outcome
ITFRLCV	0.097	0.046	0.085	0.010	Top BD => pheno IDP
AGVB	0.045	0.055937	0.088	0.077	Top BD => pheno IDP
MECR	0.051	-0.153	0.089	-0.067	Top BD => pheno IDP
IF	0.104	-0.104	0.092	-0.075	Top BD => pheno IDP
ITR	0.099	**0.268**	0.092	0.083	Both top BD exposure and outcome
FF	0.070	0.067	0.099	-0.008	Top BD => pheno IDP
IDPFSL	**0.102**	0.149	**0.109**	0.065	Top BD => pheno IDP

Table 5.1: Summary table of exposures and outcomes with the highest absolute $DCE\beta$ values where BD is a term (top 20); β_{xy} refers to the effect of BD on the phenotype and β_{yx} refers to the effect of the phenotype on BD in both the TCE and DCE matrices respectively, status refers to whether or not the IDP is a top exposure or outcome of BD. Bold values denote the highest values in respective columns with the exception of the mean ICVF of pontine crossing tract for $|TCE\beta_{yx}|$, which was not a top 20 exposure/outcome for BD in the DCE. The abbreviations correspond to the following codes: **MSOCC**: Mean diffusivity in splenium of corpus callosum, **RCI4**: rfMRI connectivity ICA-features 4, **ARIC**: aseg rh intensity Cerebellum-Cortex, **ITFRLT**: IDP T1 FAST ROIs L thalamus, **FACRR**: FA Anterior corona radiata R, **ITFRLCI**: IDP T1 FAST ROIs L cerebellum IX, **AGV4**: aseg global volume 4th-Ventricle, **ARAS**: a2009s rh area S-parieto-occipital, **TRVL**: ThalamNuclei rh volume LP, **ALAS**: a2009s lh area S-collat-transv-ant, **HRVH**: HippSubfield rh volume Hippocampal-tail, **OCPL**: OD Cerebral peduncle L, **MCPL**: MO Cerebral peduncle L, **DRAI**: DKTatlas rh area insula, **PRAT**: pial rh area TotalSurface, **IDPIIR**: IDP dMRI ProbtrackX ISOVF ilf r, **IMCP**: ICVF Middle cerebellar peduncle, **IICPR**: ICVF Inferior cerebellar peduncle R, **WLIT**: wg lh intensity-contrast transversetemporal, **MPLOICL**: MD Posterior limb of internal capsule L, **PLAS**: pial lh area superiortemporal, **OSLFL**: OD Superior longitudinal fasciculus L, **FMCP**: FA Middle cerebellar peduncle, **ITFRVCV**: IDP T1 FAST ROIs V cerebellum VI, **LSLFL**: L1 Superior longitudinal fasciculus L, **ITFRLCV**: IDP T1 FAST ROIs L cerebellum V, **AGVB**: aseg global volume-ratio BrainSegVol-to-eTIV, **MECR**: MD External capsule R, **IF**: ICVF Fornix, **ITR**: ICVF Tapetum R, **FF**: FA Fornix, **IDPFSL**: IDP dMRI ProbtrackX FA str l

5.4 Discussion

5.4.1 Bidirectional effects of BD on IDPs in the TCE space

We performed over 25 thousand tests of the causal relationship between 159 brain phenotypes and bipolar disorder. After strict FDR corrections, we found that a diagnosis of BD had a negative effect on the standardised surface area of the lateral orbitofrontal cortex in the left hemisphere, a result identified by all MR methods (Figure 5.4). This finding is consistent with the meta analysis performed in [58], which detailed widespread frontal cortex thinning in BD cases versus controls. Additionally, cortical thinning of areas of the prefrontal cortex have been consistently observed in BD patients across multiple studies [283, 284]. The left lateral orbitofrontal cortex is thought to be involved in reward systems and emotional regulation, a core system thought to be disrupted in BD patients [285]. While widespread frontal cortex thinning has long been observed in BD patients, we report causal estimates of BD affecting gray matter structure. Because our causal estimation captures 'all-cause' effects of one phenotype on another, it is possible that the mechanism by which BD affects cortical structure is via another unmeasured mediating factor. For example, it is possible that a diagnosis of BD accompanied by subsequent therapeutic treatment could be a cause of this effect, especially given that certain antipsychotic medications can cause changes in

brain structural volumes [286]. However, the primary treatment of BD, lithium, has not been observed to be correlated with cortical thinning in the left lateral orbitofrontal cortex, and moreover has been correlated with cortical thickening [287]. However, it is difficult to understand how exactly BD may cause cortical thinning.

Further, we found that increased surface area of the left hemispheric anterior transverse collateral sulcus had a significant positive causal effect on the diagnosis of BD, increasing the odds ratio by ≈ 1.2 on average. This region is encompassed by the ventricles, whose increased volume sizes have been previously observed in BD patients [55]. Further, this region delineates the boundaries of the larger hippocampal complex, which is thought to be responsible for learning and memory development tasks, an executive system of great interest to BD pathology [288]. Interestingly, patients with major depressive disorder in smaller studies have been described as having reduced sulcal depth in this region, which may correlate with lower surface area [289]. While depression is a clinically distinct entity to BD, we expect a degree of neuroanatomical overlap, making these results difficult to reconcile. Secondly, how such a structural difference in practice may have a causal effect on BD diagnosis is a point of contention. As previously mentioned, it is possible this causal effect may be mediated by an unmeasured factor. However, it is also possible that aspects of ventricle structure have some impact on cognitive systems related to BD, thus meaning that changes to a feature effecting that system may increase the chances of developing BD.

We also described the positive causal effect of the mean ICVF in the pontine crossing tract on BD, finding that a unit increase in this quantity resulted in an average odds ratio greater than 1 in 4 methods. Higher ICVF values correspond to greater concentrations of water between axonal and dendritic boundaries. The pontine crossing tract is a series of nerve fibres involved with the pons, a key component of motor and sensory function. However, the MR Egger estimates of this effect have very large confidence intervals that overlap with 1, meaning this effect may be less robust than the previously mentioned pairs. A previous study also observed white matter abnormalities along the pontine crossing tract in BD patients, with lower fractional anisotropy (FA) reported vs. control subjects (which corresponds to less constrained water molecule flow in a region) [290]. However, this effect was described in a sample size of 68 and it is not known whether increased ICVF is correlated with lower FA in all cases. Intuitively, we may expect that the quantities should instead be positively correlated, with increased amounts of intracellular volume in a region acting as hindrances upon isotropic molecule motion, thus corresponding to increased FA values. However, the interpretation of this finding is difficult as this reasoning will not always hold for complex white matter fibers [291]. This is because FA is a global measure of hindrances to small molecule motion in a region, and the degree of anisotropic molecule motion may not be an accurate reflection of localised microstructural pattern differences. Nevertheless, this result implicates white matter microstructural

differences as causal factors increasing the chances of BD diagnosis, specifically the average amount of intracellular fluid in the pontine crossing tract. Microstructural alterations impeding molecule motion between regions through white matter tracts may have downstream implications for the function of executive systems overall, which may be of particular note in the context of BD pathology.

5.4.2 *inspre*-derived direct causal matrix: BD effects on IDPs

Using the *bimmer* and *inspre* packages, we calculated an approximate sparse inverse of our input TCE, yielding a DCE of unmediated direct causal effects. The top 20 highest weight pairs where BD was the exposure term contained 15 diffusion phenotypes ($P = 0.13$, Table 5.1), with the largest affected IDP being the mean diffusivity in the splenium of the corpus callosum ($\beta_{DCE} = -0.11$). Differences in white matter mircostructure in the corpus callosum have been previously investigated as a potential endophenotype of BD [59] The mean diffusivity of a region represents the degree to which small molecules can move without hindrance in any direction, and is thus distinct from FA, which captures the amount of directional molecule motion hindrance in the primary direction relative to orthogonal direction axes. Broadly, higher mean diffusivity captures the amount of directed and undirected molecule motion within a region, whereas FA speaks to the amount of directional diffusivity. As such, this result suggests BD directly causes microstructural alterations in the splenium of the corpus callosum which result in restricted molecule motion in every direction. What exactly we can infer about the properties of the corpus callosum in BD patients from this result is contentious; theoretically, increased hindrance in every direction of possible molecule motion may imply increased white matter density in this region, a result contradictory to findings of reduced volume in the corpus callosum of BD patients [292]. However, decreased mean diffusivity only captures the amount of restricted molecular movement in a region, which may not always correspond to increased white matter density - rather, it may capture a reorganisation of existing white matter that affects the diffusivity of molecules in that space.

The second highest weight BD outcome is the weighted mean FA in superior thalamic radiations ($\beta_{DCE} = 0.109$), a nerve fiber involved in transmitting information from the thalamus to the cerebral cortex and thus an important part of the sensory-motor system. Interestingly, decreased FA in the anterior thalamic radiations has been previously described in BD and schizophrenia [293, 294]. Our results indicate the opposite direction of effect, with BD causing an increased amount of FA corresponding to more anisotropic molecule motion constrained by white matter microstructural differences. It is worth noting that previous literature is concerned with the anterior thalamic radiations as opposed to the superior part of the structure.

Further, most BD-outcome pair coefficients were given lower values in the DCE relative to the TCE. This implies that BD's causal effects as described by TCE coefficients may have been mediated by other nodes included in the network, thus resulting in downweighting in the DCE, an estimate of the direct unmediated causal effect. Given the strong effects of certain IDPs on each other, and the relatedness of certain phenotype constructs, this is unsurprising. However, most of the TCE top 20 outcomes of BD are present in the DCE top 20 outcomes of BD (13), albeit with coefficients of a lesser magnitude, suggesting that most targets of BD do not have especially strong included mediators in their BD-outcome relationships.

Interestingly, the phenotype with the second largest negative causal effect from BD in the DCE is an independent components analysis summary feature for resting state functional networks ($\beta_{DCE} = -0.10$). This phenotype was estimated by applying independent components analysis to all resting state functional network measures to derive 6 components representing the activity of multiple networks in a more parsimonious fashion [232]. Further details on the derivation of these features can be found in [232]. Briefly, this phenotype captures the functional correlations of multiple independent resting state functional networks. An examination of the high-weight components of feature 4 reveals a landscape of inter-related brain regions, including the relationship between the posterior cerebellar regions and frontal/dorsal parietal lobules. The overall relationship observed in resting state networks is a net negative correlation between the blood oxygenation levels in these regions; this implies BD may cause negative correlation between brain regions in terms of blood oxygenation levels, a result consistent with literature that observes the same effect in limbic regions [295]. However, the authors in [295] posit that the functional network disruptions have a theoretical causal effect on BD as opposed to vise versa; here, we report evidence that BD may also cause disrupted connectivity in resting state networks involving the limbic system.

Interestingly, the top gray matter structural target of BD in the TCE, the volume of the 4th ventricle, is transformed from $\beta_{TCE} = -0.11$ to $\beta_{DCE} = -0.07$, suggesting that BD's all-cause effect on reduced volume in this region is mediated by effects through other brain regions. Lateral ventricle volume increase has been previously described in BD by multiple studies [55], meaning this direction of effect runs contrary to the literature. However, the volume of the 4th ventricle, involved in the production of cerebrospinal fluid, has not been investigated or reported individually.

5.4.3 *inspre*-derived direct causal matrix: Exposures acting on BD

While diffusion tensor imaging phenotypes are frequent targets of BD in our network, the phenotype exerting the greatest absolute effect on BD is the right hemispheric volume of the hippocampal tail

($\beta_{DCE} = -0.12, OR = 0.88$, Table 5.1). This indicates that increased volume of this region decreases the chances of BD diagnosis. Volumetric reductions in the hippocampuses of BD patients have been described by [58], but its causal effect has been previously unreported. The hippocampus is a key component of the limbic system, which has important roles in emotional regulation from a functional standpoint [296, 297]. A unit increase in the standardised hippocampal tail volume has a larger negative effect on BD than the reverse association ($\beta_{DCE} = -0.047$), meaning that while this relationship has evidence of bidirectional causality, there is a clear magnitude difference in favour of a phenotype affecting BD. It is possible that this causal effect is mediated by other nodes in the network, but tracing the most influential causes of increased hippocampal tail volumes is difficult given that every phenotype has a path to another phenotype in the network. Our causal paths analysis for this phenotype highlighted 3 other gray matter structural phenotypes as its most influential variables. This can be explained given that the size of brain regions can inform the size of other neighbouring brain regions. This finding is interesting given that white matter microstructural variation is thought to be a candidate endophenotype of BD [59].

We also observe that this phenotype primarily effects other gray matter structural phenotypes - the top 5 IDPs affected by right hemispheric tail volume contain 3 T1-weighted structural phenotypes, the amygdala lateral nucleus volume, left hemispheric volume of whole hippocampus, and the right thalamic nuclei volume. These regions are all considered important to limbic system function [297]. Furthermore, the lone diffusion tensor phenotype in the top 5 targets affected by this IDP - the FA in the right cingulum cingulate gyrus - is also known to be involved in the limbic system and has been previously characterised as structurally abnormal in BD patients [298]. These results describe a network of interactions involving converging cognitive systems and different phenotypes with moderate causal effects on BD diagnosis. These results indicate that information on the effect of variables on other phenotypes in a network can contextualise results, as in the example of the limbic system.

The highest weight positive causal effector of BD in the DCE is the FA in the middle cerebellar peduncle ($\beta_{DCE} = 0.10, OR = 1.1$). This indicates that increased directional hindrance informed by white matter microstructural constraints in a white mattter fiber connecting the cerebellum to the brain stem has a positive causal effect on BD diagnosis. As previously mentioned, we must be cautious in interpreting this result as a finding of increased white matter density, although this may still be the case in this specific region [291]. The major causal effectors of this phenotype are proximally close in location, with 2 diffusivity measurements of the pontine crossing tract featuring in its most significant causal path. The middle cerebellar peduncle is a large set of fibers connecting the pons and the cerebellum, and as such plays a role in sensory and motor function. A review carried out by [299] of neuroimaging studies of major depressive disorder and BD indicates that the cerebellum appears to interact with areas of the limbic

system to regulate mood. However, it is unknown what the significance of this phenotype is in relation to the pathology of BD, although areas of the cerebellum have been previously described as functionally disrupted in BD [300, 301]. Despite this, differences in functional connectivity are not explicitly correlated with FA values in all cases, meaning it is difficult to relate this finding to the existing literature. While certain regions associated with the cerebellum were tested in MR experiments from [276], the causal effects of BD acting on those IDPs were either non-significant or not performed owing to the phenotype selection strategy employed by the authors. We also found no previous studies reporting FA differences in cerebellar fibers, making this result of particular interest for its novelty.

Interestingly, we found that the absolute effects of phenotypes on BD were statistically larger than the absolute effect of BD on phenotypes. Of note when interpreting this result is the fact that most phenotypes have bidirectional causality even when exact magnitudes differ. Furthermore, the unmediated causal effect of BD may differ from its all-cause effect. However, this result is not supported by a t-test of TCE BD exposures and outcomes which finds the same effect ($P < 0.001$). It is possible that BD influences on brain phenotypes are lesser in magnitude and not necessarily specific to the phenotypes presented in this analysis. Additionally, the causal estimates of certain exposures' effects on BD may be influenced by weaker instruments owing to differences in sample size and ancestry effects.

5.4.4 The role of diffusion phenotypes and network trends

While the regression of 5.10 for coefficient differences between diffusion tensor and T1-weighted structural phenotypes did not yield significant results, we found that diffusion phenotypes may play important roles in our DCE networks by less direct means. For example, diffusion tensor outcomes were found to have on average 3 more exposure effects from different phenotypes with $\beta > 1\sigma$ compared to gray matter structural phenotypes (Figure 5.11, $P = 0.0171$). While higher degree is not a direct proxy of overall node importance, this result suggests that white matter microstructural phenotypes can be affected more often, and as such contribute to more causal paths. This is especially pertinent given the fact that every node is reachable by another, meaning nodes affected more often are likely to mediate more causal interactions. Interestingly, we noted that the same effect is not observed in exposures, meaning that phenotype categorisation is not correlated with an increased number of phenotype targets.

Interestingly, we found that gray matter structural phenotypes affected in greater magnitudes by BD were more likely to be significantly affected by other phenotypes relative to gray matter structural phenotypes with lower BD causality estimates (Figure 5.12, $P = 0.008$). This suggests that, similar to the overall trend of higher diffusion node outcome degree, gray matter structural outcomes affected with greater BD β values are likely involved in more causal paths relative to other structural outcomes affected by

BD. However, these results were non-significant in an interaction model of all phenotypes, meaning that absolute BD causal effects in either direction are not predictive of the number of phenotypes affecting a node. Given that brain regions appear more likely to causally influence each other than to be influenced by BD, this is not entirely surprising, and further suggests that not all IDPs have strong relationships with BD.

The importance of diffusion tensor phenotypes overall was underscored by the non-redundant causal paths analysis in Figure 5.13, occurring as recursive top exposures 515 times out of a possible 800 in paths of length 5 for every phenotype. This means that diffusion IDPs appear to be overrepresented in causal paths in the network. This speaks to the general importance of white matter microstructural alterations in the context of bipolar disorder. White matter fiber tracts, primarily captured by diffusion tensor imaging modalities, are less neuron-heavy than gray matter brain regions, and as such these regions are primarily involved in information transfer from one region to another. This result overall recalls aspects of the dysconnectivity hypothesis, which implies that functional connectivity disruptions are a central feature of BD and other psychiatric conditions [78]. However, microstructral differences in white matter fiber tracts do not strictly speak to connectivity disruptions between regions; rather, they describe the local constraints on molecule motion by multiple metrics that are informed by white matter microstructural organisation. However, our results indicate that metrics describing microstructural white matter integrity are of importance at a network level due to their occurrence in causal paths. This is especially interesting from a gray matter/white matter perspective, with gray matter structural features having greater magnitude direct effects on BD and white matter microstructural features having overrepresented in causal paths.

5.4.5 Feedback loops and bidirectional causality

In examining every causal pair in the network and its reverse association, we found that just half of interactions have the same direction of effect in both cases (Figure 5.14a). These relationships can be thought of as feedback loops, but these results nevertheless underscore the difficulty of definitively resolving a single causal actor in a pair of phenotypes. Our understanding of causality suggests that single causal factors with large effects are unlikely. Instead we construct a network of complex bidirectional causal relationships of differing magnitudes and effect directions. This complexity is further highlighted when considering pairs of phenotypes with differing signs of effects in the forward and reverse direction with significantly different magnitudes (Figures 5.16,5.17). In considering the visualised examples, we can attempt to reason with their causal structure.

For example, in Figure 5.16, we can see the relationship between FA in the corona radiata, the mean ICVF in the tapetum, and BD. Coefficients represent the effect of a standardised unit increase on the value of

another phenotype. If we assume that a unit decrease in a phenotype causes the same magnitude of effect in the opposite direction, we can see that from every starting node, there appears to be a feedback inhibition-like loop. Starting from BD, a diagnosis causes higher mean ICVF in the tapetum, which in turn causes increased FA in the corona radiata. However, increased FA in the corona radiata has a negative effect on BD diagnosis, leading to an apparently inconsistent empirical estimate. This effect is repeated no matter the starting point in the subnetwork, ultimately resulting in a decreased causal effect on the input. This is initially difficult to resolve but there are several factors to consider. Firstly, BD is usually diagnosed once throughout a person's life and no factors can cause a decrease in its odds once the phenotype is present. Therefore, in considering causal paths beginning with BD, we must be conscious that signal returning the BD node cannot meaningfully effect diagnosis odds in any tangible sense. Secondly, we can consider the possibility of non-symmetry about the coefficient estimates, meaning that the magnitude effect of a unit increase is not exactly equal to the magnitude effect of a unit decrease on another phenotype. This may be an example of Simpson's paradox, whereby certain stratum of the data may describe the opposite effect to the same variables in a different strata. However, we must also consider that not all causal estimates are true due to potential violations of MR assumptions. The inclusion of weak instruments or ancestry effects may lead to an imprecise causal estimate with differing directions of effect.

Figure 5.17 visualises the three pairs of BD associations with differing signs of significant magnitude from Figure 5.14b. Again, we see that BD diagnosis results in decreased odds of a BD diagnosis through every node in the subnetwork. Of note is the biological significance of these phenotypes, with the fornix and the cerebellar cortex involved directly or indirectly with the limbic system, and the external capsule acting as a route for cholinergic fibers, an important focus of recent BD research efforts [302]. The fornix is a major output tract of the hippocampus and its structural alterations have also been previously described in BD compared to controls [303]. The exact interpretation of these contradictory β effects is difficult to determine, but we note the occurrence of another limbic system subnetwork in our results. We further note that the three phenotypes in question relate to white matter microstructure - intensity increase can be a proxy for increased white matter density or fat density, and ICVF/mean diffusivity have been previously explained.

5.4.6 Global trends

These results indicate that both diffusion tensor and gray matter structural phenotypes are important in the derived DCE network. Moreover, their effects can manifest as differences in node degree and over-representation in causal paths. Applying a graphical lasso framework in this domain grants us two advantages – firstly, we can estimate approximate unmediated causal effects conditional on other included

nodes in the graph; secondly, we can derive networks of unmediated causal estimates that can facilitate reasoning on a larger scale using degree and path metrics. It is of particular note that the highest weight effectors and targets of BD are not over-represented for any category of phenotypes (Figures 5.8,5.9). A possible explanation is that certain mediators of BD's causal effect may not be included in the graph, such as medication usage post-diagnosis, meaning that macro/microstructural changes as a result of BD may be represented as all-cause coefficients. However, the regions in question do not have strong literature support as candidates affected by BD treatment drugs, suggesting that other mediators may be involved in this interaction. Our results may indicate that certain structural changes, such as reduced volume of the hippocampal right tail, may be involved in influencing systems that contribute to BD pathology. We also report several strains of evidence across imaging modalities implicating disruptions to the limbic system as a causal effector of BD.

However, we note that interpreting how brain imaging phenotypes may affect BD is difficult. This is made especially challenging given the difficulty in ensuring that all MR assumptions are satisfied. For example, an estimate of BD causally effecting increased FA in an arbitrary white matter tract is reliant on ensuring that associated variants are causal for BD, that there are no ancestry effects between the two phenotypes, and also that there is no pleiotropy whereby the variants are associated with the outcome as well as the exposure. Even when all assumptions are met, MR captures the all-cause effect of an exposure on an outcome, making it difficult to resolve how exactly one phenotype may affect another. Deriving direct causal effect estimates is dependent on the variables included.

5.5 Conclusion

We performed over 25 thousand bidirectional MR tests and applied a graphical lasso framework to derive estimates of unmediated causal effects between BD and brain phenotypes. Our results implicated several previously described brain systems and individual regions as important causal actors in the network, which may provide future insight for studying the overall neuroanatomical pathology of BD.

Chapter 6

Discussion

6.1 Research Findings

The field of brain disorder imaging and genetics research is predominantly focused on understanding the biological bases of a myriad of debilitating conditions. Our work attempts to contribute to this goal via several distinct and interdisciplinary methodologies. In sum, we sought to characterise the genetic architecture of neuroimaging biomarkers discriminative for psychiatric phenotypes identified using flexible and interpretable deep learning approaches and investigate the causal relationship between neuroimaging and psychiatric conditions. Below, we summarise our main findings.

In Chapter 2, we systematically reviewed the literature of convolutional neural network models applied to neuroimaging collections of patients with brain disorders. We found that predictive performances were high in testing sets, with an average reported accuracy of $89.36 \pm 8.694\%$ ($\mu \pm STD$). However, we also observed several methodological practices which may limit the clinical utility of such models, including a lack of code availability, variable application of repeat experiments, and interpretability issues. This study marked the first time to date that studies in this problem domain were assessed through the lens of modelling practices, transparency, and interpretability. In the context of our overall goal, our results were encouraging – structural neuroimaging data can be used to classify brain disorders. These data modalities therefore have potential as endophenotype sources. Despite this, the aforementioned limiting factors made it difficult to make use of existing model architectures. Specifically, six studies made code available, and seventeen studies considered interpretability via a variable range of methods. From a code-sharing perspective, none of the considered studies made their full models available, which complicates our general task. Concurrently, while there exist drawbacks in the usage of certain interpretability methods, examination of these approaches in the seventeen studies applying interpretative methods for CNNs informed a

careful consideration of the best means by which to empirically express an understandable quantity representing region importance. In a larger sense, we found that repeat experiments were applied in over half of all studies, and interpretability in some form was considered in seventeen of our 55 considered papers. These trends are encouraging and will no doubt increase the clinical potential of such models moving forward. While interpretability has historically remained of less importance than predictive outputs in the field of deep learning [304], it is important to bear in mind the scope of our presented studies. For instance, our ultimate goal of biomarker derivation necessitates a different set of experimental conditions that must be fulfilled; because we are concerned with capturing the genetic architecture of deep-learning derived features, we must understand their constituent elements and express those features numerically. When the study goal is to examine the predictive capabilities of modelling approach, a different set of conditions become important, including a focus on model tuning and optimisation. Despite our aspirations to better understand brain disorders using these approaches, we should not imply that one goal is more important than the other. Indeed, the motivation for our research goal was informed by an observation that various brain disorder phenotypes can be predicted with high accuracies using deep learning models. Here, our interest in inferential dynamics was reliant on an emphasis on predictive performance.

Subsequently, we attempted to use the knowledge gained in our systematic literature review to guide model construction in Chapter 3. Using data from the ADNI consortium, we trained a CNN predictive model on structural neuroimaging data from participants with and without AD with the aim of extracting an interpretable quantity whose genetic architecture can be queried. We obtained a mean accuracy across three repeat experiments of $80.3\% \pm 1.24$. Examining Table 2.1 shows that our highest performing model (82%) is in the bottom 7 of studies considered. However, our model construction was motivated by a desire for interpretation, meaning our architecture may not have been optimal for prediction. Future studies could investigate how best to balance these two motivations.

Our parsimonious modelling approach informed an architecture with a penultimate global average pooling operation followed by a final sigmoid layer. Our global average pooling layer took the mean of the output from multiple feature maps and our sigmoid layer returned the sum-weighted combination of those averages. This allowed us to directly examine the final layer weights to determine which feature map average had the highest contribution to a classification of Alzheimer's disease. This had the useful property of providing a single number summarising the importance of a particular feature map to the output. Furthermore, this allowed us to apply several straightforward interpretability methods to better understand the features encoded in the feature map score. For instance, we regressed 40 principal components derived from tabular summary information of participants against our score, finding that principal component 2 was significantly associated with score after multiple testing corrections. Examining the

loadings of this component revealed that multiple ventricle measures (Table 3.3) were top loadings in our correlated principal component. Additionally, we found that eleven brain variables had p-values less than 0.05 in full regression models of every variable against our CNN score; the two with the highest R^2 values were visualised in Figure 3.12 (volume of left lateral ventricle and thickness of right inferior parietal gyrus). These results indicated consistency with previous literature noting ventricular enlargement in AD [204] and atrophy in parietal lobe regions [235]. Additionally, these findings were consistent with those of the authors in [305], which we considered for our systematic literature review. This is encouraging, as our model can recapitulate associations previously reported in the literature despite using an entirely different approach.

Furthermore, we also obtained the gradient of score output with respect to small changes in the input image to yield a visual map of feature importance in Figure 3.11. This confirmed our empirical findings by implicating parietal and ventricular regions as important to model representation. Our visual gradient experiments coupled with our regression analyses provided substantial evidence that the aforementioned regions were important to model representation.

Our multi-method approach to understanding model representation built upon our findings in the review of Chapter 2, where we identified the potential of simple regression models for adding an extra layer of depth to our interpretability considerations. This result underscored the importance of empirical grounding in interpretability considerations, as we manipulated patient tabular information to supplement our gradient-based findings.

Further, our GWAS of this phenotype yielded several plausible genetic associations with AD, identifying several genome-wide significant findings related to molecular pathways previously associated with the disorder, including *PSME4*. This gene encodes a subunit of the proteasomal complex which is responsible for cellular homeostasis. The primary molecular feature of AD is the deposition of amyloid-β plaques, meaning that impeded proteasomal recycling activity may be a factor contributing to this cellular phenotype. Adding to this evidence is the observation that differential expression of *PSME4* is observed in AD individuals relative to controls [227]. Additionally, lowered *PSME4* mRNA levels have been previously described in patients with Parkninson's disease, a separate neurodegenerative disorder [226]. This disrupted cellular homeostasis finding is given further weight by *ERLEC1*, which forms part of the endoplasmic reticulum degradation pathway. In mouse brains, AD-like features can be simulated by downregulating membralin, a protein involved with the endoplasmic reticulum degradation pathway; further, *ERLEC1* was found to be a hub gene in differentially expressed human AD gene sets in the brain [224, 230]. These results suggest that disrupted cellular homeostasis via perturbed degradation of extracellular compounds is a key molecular system associated with our CNN score.

In general, we found that most of our identified genetic variants fell in intronic or intergenic regions, which is of particular interest when considering the analyses of Chapter 4 (which will be discussed shortly). In understanding these findings, it is important to note the phenotype encoding; these variants are associated with global neuroanatomical changes associated with AD, specifically in certain parietal and ventricular regions. Therefore, we can posit that our molecular results are associated with these specific neuroanatomical changes which are observed in AD patients. However, even with an endophenotype, it is still difficult to directly link our molecular results to the larger phenotype.

While these results are encouraging, we faced a limitation in the application of this method. Namely, we trained a supervised predictive model to extract our phenotypes, meaning that our genetic results are dependent on the representation of a model trained on an arbitrary data split, which may introduce bias or inflated test statistic association scores. Additionally, our CNN method requires intensive preprocessing owing to the data splitting procedure, meaning it may be difficult to apply this approach to new domains. Therefore, in Chapter 4, we sought to derive endophenotypes using an interpretable deep learning model with a different approach that could mitigate some of the aforementioned issues. We trained an unsupervised shallow de-noising autoencoder on tabular brain summary information from AD participants and controls in the same dataset. Specifically, we trained a single-layer de-noising autoencoder with 60 latent nodes on corrupted input data with the aim of creating a robust information embedding in our encoding layer. We achieved a reconstructive loss of $6.21e{-}9$ after an iterative noise reduction schema (whereby the amount of noise added to the data was reduced by a constant factor every 250 epochs). After training, we found nodes with differential scores in AD participants and controls using an independent t-test; we then corrected the resultant p-values using a Benjanini-Hochberg procedure and examined the node with the most significant p-value (node 51, $adj.P = 2.47e-07$). This procedure was motivated primarily by a desire for statistical rigor, despite the fact that other nodes may contain relevant information. However, if we decide to examine other nodes, or even combinations of nodes, we must be conscious of how to select relevant nodes. As such, focusing on just one node with stringent statistical thresholds ensures that our analysis is specific for one signal associated with AD.

Interestingly, we found that the variables with the greatest impact on node output score had little overlap with those identified in our CNN analysis. This may be due to different methodologies capturing different neuroanatomical relationships. Namely, we found that atrophy of several temporal lobe regions was found to influence node output (Figure 4.9). This was in agreement with several studies of AD neuroanatomy describing volumetric loss in temporal lobe structures, and an encouraging confirmatory finding [203, 248, 251]. Further, we found agreement with our CNN results in a finding of increased ventricle size driving node 51 activity. We then carried out a GWAS of our autoencoder-derived quantity

and found that our three significant variants were mapped to long non-coding RNA transcripts (*RP11-239H6.2, RP11-509J21.1, and RP11-707M1.1*). This is of particular interest in the context of the quantity they are correlated with – global neuroanatomical changes in AD participants, with a focus on pronounced temporal lobe structural atrophy. The role of long non-coding RNAs in mediating gene expression in brain tissues is a topic of growing relevance [306]. In particular, the expression of certain long non-coding RNA transcripts has been posited as candidate biomarker for AD [255]. These molecules are thought to affect regulation of gene expression via various means, including through acting as precursors to other regulatory molecules (micro RNA), 'sponges' for other gene transcripts via sequestration of transcription factors, and direct gene silencing via binding and degradation [306]. However, the correlation of these loci with gene expression levels of long non-coding RNAs was not investigated in these studies, which may add further plausibility to our findings. Additionally, our gene-based results implicated *EIF2B5* as a statistically significant gene; this was encouraging in light of a recent Mendelian randomization study of AD, which found that its expression had a significant causal effect on AD risk in a multi-ancestry cohort [262]. This gene is a eukaryotic translation initiation factor essential for translatory regulation; mutations in this gene can also result in leukodystrophy, a neurodegenerative condition marked by widespread white matter atrophy [261]. Our non-coding transcript findings in combination with *EIF2B* significance at the gene level present a set of molecular results distinct from that of our CNN analyses. We will discuss the implications of this further in later sections.

We also observed that genes found to be statistically significant in our *MAGMA* gene-based test had tissue-specific expression for the substantia nigra (Figure 4.12). Although this result did not survive Bonferroni correction for multiple testing, this tissue in particular is of interest for AD; previous studies have described β-amyloid and tau phosphorylation accumulation in this brain region [263]. Of particular note is the substantia nigra's role in the dopaminergic system, posing a possible link between AD and dopamine; indeed, neuronal loss has been observed in this region in AD participants [264, 266]. In summation, we find that genes significant in a gene-based test for neuroanatomical variation associated with AD have tissue-specific expression for a brain region involved in the dopaminergic system. Concurrently, long non-coding RNA transcripts and translation regulation machinery components are implicated by variants found to be significantly associated with the same node quantity. This speaks to a range of diverse molecular findings arising from an association study of a latent space component summarising neuroanatomical variation across multiple brain regions.

However, our power analyses indicated that both of our studies were underpowered to detect small to moderate genetic effects. It is unknown what effect sizes we expect for these derived endophenotypes, but larger sample sizes are required to have more confidence in results. Furthermore, ancestrally-constrained

GWAS of our autoencoder score introduced 12 more significant independent genetic loci. These results are in spite of controlling for ancestry at the principal component level, which may suggest that our autoencoder endophenotype is unstable. There also exist issues with Manhattan plot outputs, with the presence of broad peaks and significant variants not associated with LD peaks, which may decrease confidence in our results owing to possible artefacts at the genotype level. Larger sample sizes in a different cohort and more stringent quality control procedures may be required in future studies. Overall, these issues mean we should be cautious in the interpretation of the full analysis results.

The implicit assumption underlying our endophenotype analyses is that there is a causal relationship between neuroimaging measures and brain disorders. In Chapter 5, we sought to investigate this assumption in greater detail for a separate psychiatric disorder with subtle neuroanatomical correlates – BD. We obtained summary statistics from multiple GWAS of imaging phenotypes from the UK Biobank and the latest PGC BD GWAS to carry out bidirectional MR experiments and construct causal networks using inverse sparse regression [38, 232, 277]. Briefly, we filtered a set of 3929 brain imaging phenotypes by the number of genome-wide significant loci in their association results, retaining those with greater than 5 variants. This resulted in a candidate set of 630 phenotypes which were further filtered for a multicollinearity-like quantity – genetic correlation (rg). We flagged pairs with rg values greater than 0.85 as potentially redundant, and removed the entry from the pair with less significant genetic instruments. This resulted in 159 brain imaging phenotypes plus BD. We carried out instrument quality control for pleiotropy or violations of MR assumptions and carried out over 25 thousand bidirectional MR experiments to obtain a matrix of total causal effects, whereby the causal effect of every phenotype on every other phenotype is estimated. We found that 9 BD-imaging phenotype pairs remained significant after Benjanini-Hochberg multiple testing correction in at least two of the five considered MR methods. These included the effect of BD on the surface area of the lateral orbitofrontal cortex, the effect of the surface area of the left hemispheric anterior transverse sulcus on BD, and the effect of the mean ICVF in the pontine crossing tract's on BD (Figures 5.3-5.5). These findings suggest that BD may have a causal effect on decreased brain region volume. In particular, the lateral orbitofrontal cortex is thought to be involved in reward systems and emotional regulation, with previous studies describing cortical thinning in prefrontal regions of individuals with BD [283–285]. Our work suggests that BD may affect neuroanatomical variation in prefrontal cortex regions. Further, we also estimate that certain neuroanatomical measures may have a causal effect on BD diagnosis, including decreased mean ICVF in the pontine crossing tract and increased surface area of the left anterior transverse sulcus. To our knowledge, the causal effect of neuroanatomical variation on BD has not been estimated previously. In relation to our previous results implicating $EIF2B$ as an associated gene of AD and the work of the authors in [262] describing that its

expression has a causal effect on AD risk, we provide evidence of a factor causally influencing the risk of a brain disorder. Indeed, our network analysis allowed us to derive a matrix of unmediated direct causal effects of every phenotype against each other; this allowed us to examine the nature of multiple causal relationships between BD and neuroanatomy. Specifically, we found suggestive evidence that white matter microstructural abnormalities constraining the direction of molecular movement were of importance, featuring as the most influential causal factors in causal paths (Figure 5.13). This result overall is consistent with previous findings of white matter abnormalities observed in BD [303]. Strikingly, we also found that the absolute effect of other imaging phenotypes on BD was on average larger than that of the absolute effect of BD on other phenotypes ($P = 0.0047$, Figure 5.15). While we may be able to understand how individual IDPs have larger effects on BD than *vice versa*, this result on an average level is surprising, especially considering that the brain imaging phenotypes used in this study were not selected based on any BD-specific inclusion criteria. Furthermore, our graphical lasso procedure allowed us to construct a network of unmediated causal effects, meaning that we can identify the direct causal relationship in several BD-IDP pairs, including the effect of BD on the mean diffusivity in the splenium of the corpus callosum ($\beta = -0.11$) and the effect of right hemispheric tail volume on BD ($OR = 0.88$).

6.2 Novel lines of enquiry: an integrated overview of results

Our results in totality provide several contributions to the study of brain disorders using neuroimaging. In the context of AD, we provide confirmatory results of temporal lobe atrophy and ventricular enlargement across both CNN and autoencoder approaches. Interestingly, these methodologies appear to capture different elements of AD neuroanatomical variation; our CNN method implicated parietal region volumetric reduction and overall brain size differences, whereas our autoencoder results indicated atrophy of temporal lobe regions as the main contributing factor to node 51 output. Both methods converge on ventricular enlargement as an important neuroanatomical feature of AD. This overlap is interesting given the differences in both modelling strategies, with one relying on the representation of a supervised predictive model trained on raw neuroimaging collections and the other trained on unsupervised tabular summary information. Thus, increased ventricle size in AD can therefore be thought of as a method-invariant important feature. The magnitude of difference in methodological approaches is apparent in the results of our respective GWAS experiments; we report no overlapping genome-wide significant variants and no overlapping significant *MAGMA*-based genes. However, these findings may suggest a role for disrupted cellular homeostasis as an important molecular phenotype of AD [307]. This is because we find that *ERLEC1* of the endoplasmic reticulum degradation complex and *EIF2B* of the eukaryotic translation ini-

tiation complex are significant in our CNN and autoencoder studies respectively. Concurrently, we also find that there is modest tissue specific expression in the substantia nigra for our autoencoder-based genes, an area of β-amyloid plaque accumulation. Such accumulations can be attenuated by efficient cellular recycling which can be coordinated by the endoplasmic reticulum and faithful translation control machinery. In a larger sense, while the specifics of genetic association results differ across our AD-based analyses, we find that in both studies, SNPs were enriched for intergenic and intronic functional categories in terms of effects on genes. This may suggest that neuroanatomical variation related to AD is associated with regulatory genomic regions via non-coding transcripts and cellular homeostasis machinery [308]. This is the first such result linking AD-related neuroanatomical variation to cellular homeostasis phenotypes. We believe this may be a promising avenue of future research, perhaps by examining the correlation between brain imaging measures and features of decreased autophagic fidelity.

However, this may be inconsequential owing to the fact that the majority of the genome is intergenic or intronic. Furthermore, larger sample sizes would be required to validate that the results presented here are not false positives.

The results of Chapter 5 indicate that there are multiple possible causal relationships between IDPs and BD. In the context of the overall thesis goal, this is an important extension of our biomarker derivation approaches. In particular, our results also suggest that BD may have direct causal influence on brain structural variation. For example, BD causing a volumetric decrease in the left hemispheric lateral orbitofrontal cortex provides evidence that a long-standing observational association of BD neuroanatomy is in fact a result of BD. This can aid our interpretation of the systems which volumetric loss in these region may give rise to. Importantly, this finding can help to inform our reasoning about these cognitive systems; for instance, we may next consider investigating the mechanism by which BD causes this effect, either through independent experiments or through the mediating factors included in our analysis. This is a subtle re-framing of the research question informed by our network analysis. Additionally, our network-based approach allowed us to understand how white matter microstructural variables differ from gray matter structural variables in a causal context. Namely, we provide novel indications that white matter microstructure may be causally influential in the context of BD through differing means to gray matter structural variables; this is evidenced by our causal paths analysis where diffusion-tensor imaging outcomes are overrepresented among the top exposures acting on phenotypes. In the context of our overall thesis goal, these network-based approaches are of great relevance; derived biomarkers are unlikely to act or be acted on in isolation, and our work can help to understand those dynamics in relation to other phenotype-disorder relationships. Furthermore, we demonstrate that the absolute causal effect of IDPs on BD is larger than that of BD on IDPs. Our findings suggest for the first time that causal estimates of

neuroanatomical variation on BD are on average greater than *vice versa* in BD. This may inform revised hypothesis formulation for future experiments to refute or confirm these results.

However, given the difficulty in ensuring that MR assumptions are met, these results should be cautiously interpreted. Even if assumptions are met, it is worth considering if any statements concerning the effect of a brain region on BD are possibly falsifiable or testable in practice. It would be exceedingly difficult to design any experiment that could test the results of any MR study on complex disorders. However, our results may point to lines of inquiry that may help to shed light on the molecular and neuroanatomical correlates of BD. Even if causality cannot be proven, these findings may help to build a more complete picture of BD.

Generally, we seek to define novel biomarkers of brain disorders and investigate causal links between neuroanatomy and conditions, and we may be tempted to judge their quality by the concordance with the overall body of literature. However, this reasoning immediately limits the discovery potential of this approach owing to an over-dependence on assumptions of 'similarity'. As previously noted, we are capturing different quantities using distinct approaches, a fact which also applies when attempting to compare our genetic results across studies. An awareness of the explicit question being asked will facilitate a more reasonable interpretation of the results, which is to say we should bear in mind what our biomarkers represent and what their associated genetic results are associated with. This can help us to generate new lines of enquiry into the relationships between neuroimaging, genetics, and brain disorders.

6.3 Concluding remarks

Hannah Arendt's seminal work of political theory, *The Human Condition*, draws attention to an important characteristic of modern science that sets it apart altogether from antiquated scientific thought - the importance of process [309]. Various social and technologicial factors in centuries prior to the enlightenment facilitated a consideration of the world's natural phenomenon as mostly inscrutable, or ends to be regarded in their own right. Modern science, in contrast, is marked indelibly by a focus on describing the means by which products of the world arise. Arendt's lengthy treatise on the historical circumstances that gave rise to this ontological shift are not of direct relevance to our research interests, but her observation is correct - the process character of phenomena is of primary scientific concern in the modern age. This imperative makes interpretability of opaque models and systems of particular importance. The application of deep learning predictive models to brain disorder neuroimaging gives rise to end products whose constituent elements may be illuminating as to the processes by which those conditions progress. We

should therefore endeavour to not treat the internal representations of such models as inscrutable, and continue centering the process character at the outset in our experimental efforts.

However, modern research has conceived of algorithms whose mathematical properties are often inexpressible in human terms, making this seemingly untenable to fully realise. Our research demonstrates that these conflicting forces can be reconciled, thus enabling us to take advantage of complex and usually opaque models. As deep learning approaches become increasingly widespread and sophisticated, we envision that embracing the principles of interpretability will yield novel insights across a range of domains.

Our results estimated bidirectional causality between every pair, and every node was connected to each other via paths of different magnitude. Crucial to this discussion, and conspicuously absent from empirical measurement, is the context in which a causal agent acts upon another. We derive point estimates with confidence intervals (at the total causal effects level) describing an overall causal effect of one phenotype on another. However, what this means in practice can be difficult to understand. For instance, it may be difficult to conceive of a coherent hypothesis in which a highly plastic brain measurement, which varies throughout life and is impacted by several other factors, can increase the lifetime odds for the diagnosis of BD. We may posit that the function of certain cognitive systems related to this disorder are affected by neuroanatomical variation, but it is unlikely that any causal factor acts in isolation divorced from context. Rather, it is more likely that neuroanatomical variation in the correct context may affect the odds of BD diagnosis, and conversely that BD presentation given the necessary circumstances may affect neuroanatomical variation. The significance of statements speaking to average trends is underscored by [310], in which the importance of causal context is explored. In future work, it would be interesting to consider how best to resolve causal context in our domain, with a potential approach being the use of longitudinal brain imaging data.

Coupled with this fact is the fidelity of the genetic instruments used to anchor causal reasoning. For example, while the UK biobank imaging resource is a useful data source, it reflects approximately 42 thousand volunteers aged between 40 and 69 years old of mostly European ancestry. This means that our genetic instruments are associated with neuroanatomical variation in a relatively narrow subset of one population of European ancestry. Additionally, we must expect that a certain proportion of genetic variants will not replicate across populations, meaning that our findings may not generalise in different ancestral contexts.

With this in mind, we suggest a cautious interpretation of our results in a network context. While the significance of individual phenotypes in relation to BD may vary based on context, there are likely higher-order systems that are as important as individual factors. For example, the limbic system is comprised of multiple constituent brain elements, and our findings implicate disruption to multiple aspects of this

system. While we expect that results at the element level differ, the system at large should theoretically still be associated with the phenotype. This recalls our finding that cellular homeostasis factors are genetically associated with distinct neuroanatomical biomarkers of AD identified using different deep learning approaches.

6.4 Final thoughts

Sean Eddy discussed the discipline-agnostic origins of computational biology and what it means for the future of the field in his 2005 essay, *"Antedisciplinary" Science* [311]. We too, in the field of brain disorder neuroimaging, must consider what the discipline of the future will look like. This requires careful examination of our research goals, an idea of where we are going, and a willingness to embrace the potentially useful elements of seemingly disparate research disciplines. By its nature, this work has drawn from multiple disciplinary standards, without specific expertise in any one area. However, as demonstrated in this work, a scientific approach to integrating multiple approaches can yield interesting novel findings. The 'antedisciplinary' character of scientific progress demands a needs-must attitude whereby we do not relegate branches of research to those with the 'correct' credentials. Furthermore, we as computational researchers must think carefully about what comes next if we are successful in fully characterising the genetic basis of brain disorders and identifying their associated neuroanatomical profiles. The computational biologist of the future must be ready to defy credentialism and embrace the "coherence, clarity, and glorious idiosyncrasy that can only come from a single mind" [311].

Appendix A

A.1 A note on tools

Python and *R* programming languages were used for analyses [312, 313]. Schematic diagrams were made using `draw.io`. Analyses relied heavily on the usage of *NumPy* [314] and *pandas* [315]. Visualisation of results throughout this work was carried out using *matplotlib* [316] and *ggplot* [317].

A.2 Supplementary methods and figures

Power was calculated using base *R* functions. It can be obtained with the following code:

```
N = N #sample size
alpha = 5e-8 # detection threshold
ES = 0.1 # effect size

threshold = qchisq(alpha, df = 1, lower.tail = FALSE)
power = pchisq(threshold, df = 1, lower.tail = FALSE, ncp = N * ES)
```

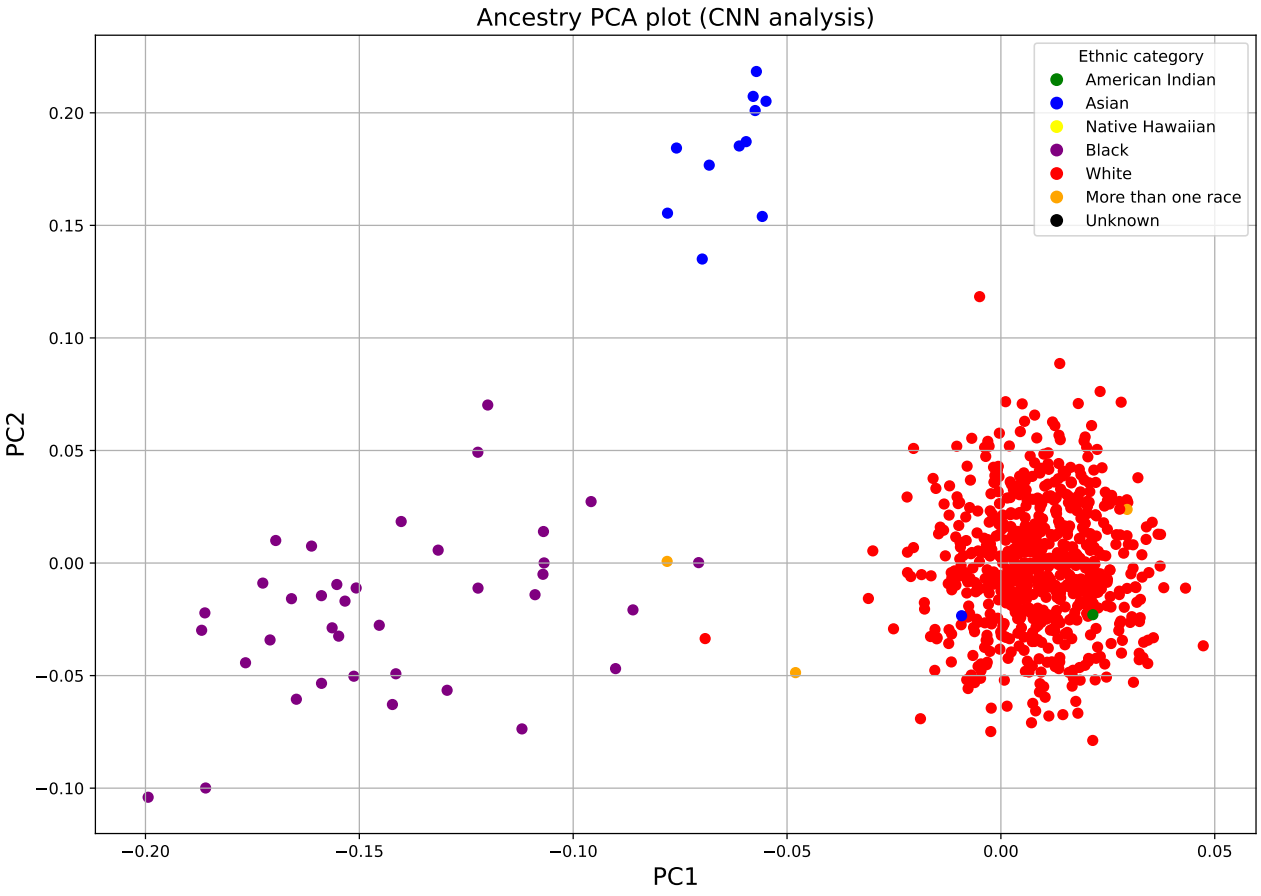


Figure A.1: Scatterplot of PC1 and PC2 colored by ancestry. Individuals colored red were taken forward for an ancestrally-constrained GWAS.

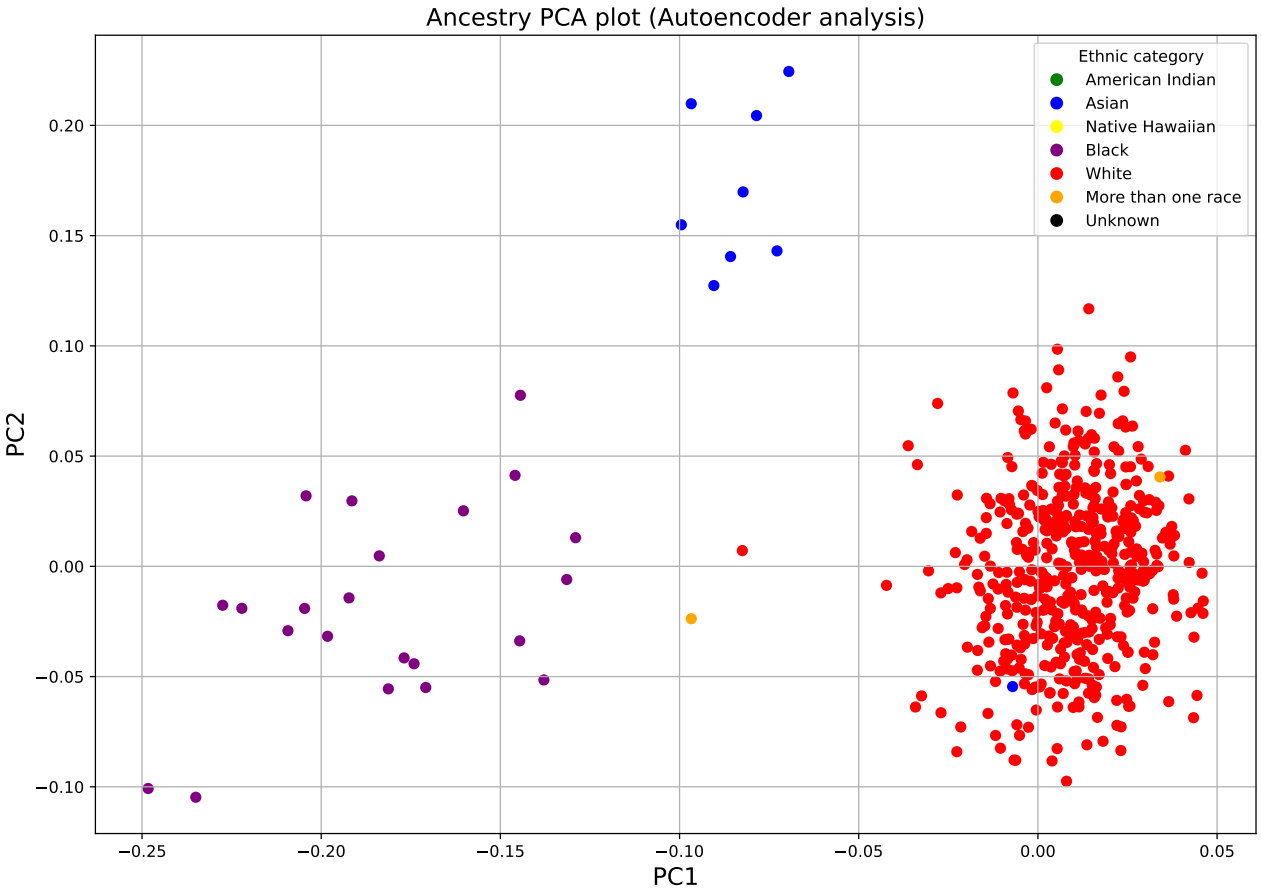


Figure A.2: Scatterplot of PC1 and PC2 colored by ancestry. Individuals colored red were taken forward for ancestrally-constrained GWAS.

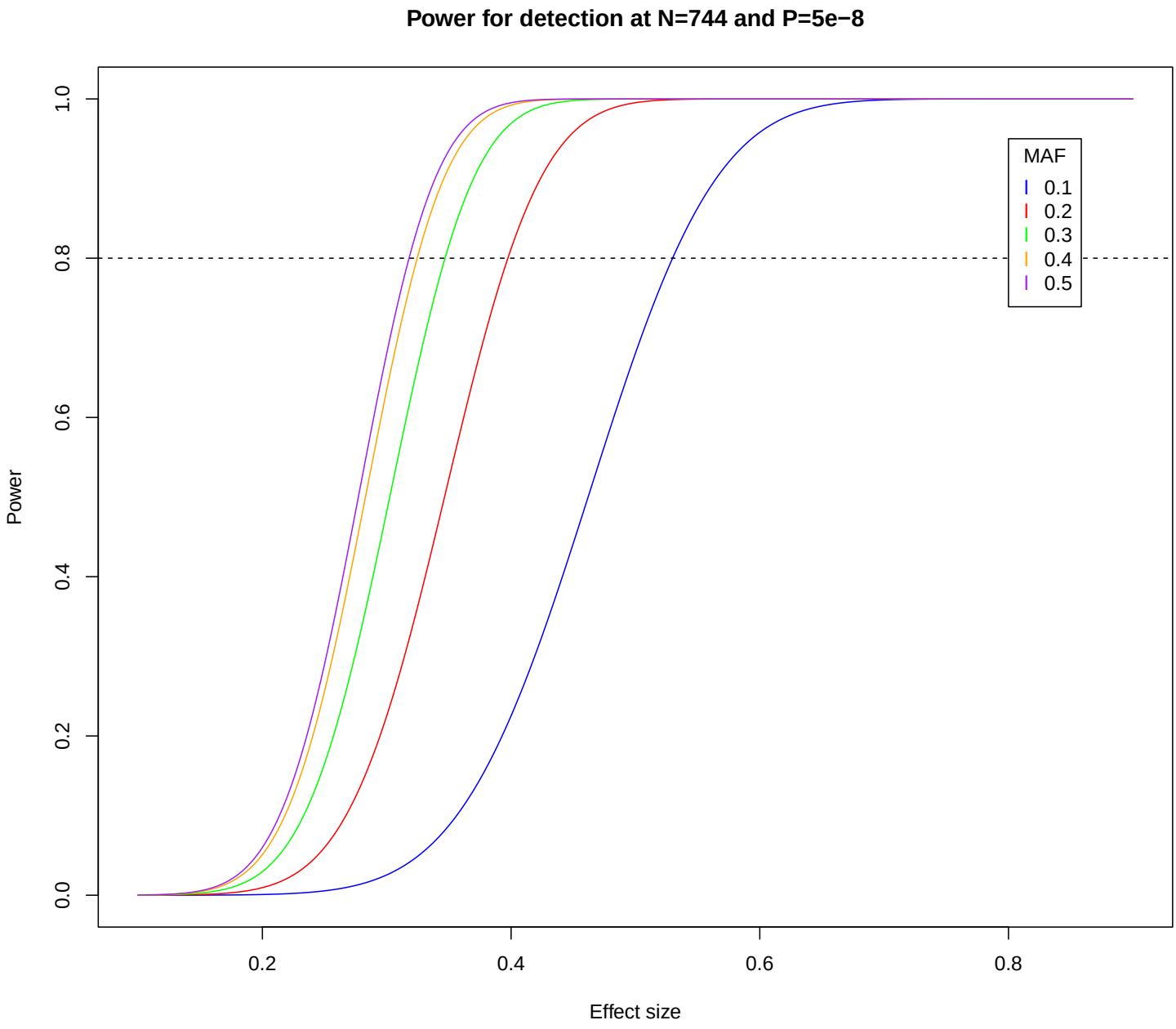


Figure A.3: Power curves for N=744 and genome wide significance at $P = 5e{-}8$. Effect size is denoted on the x-axis and power on the y-axis. Different minor allele frequencies are colored lines.

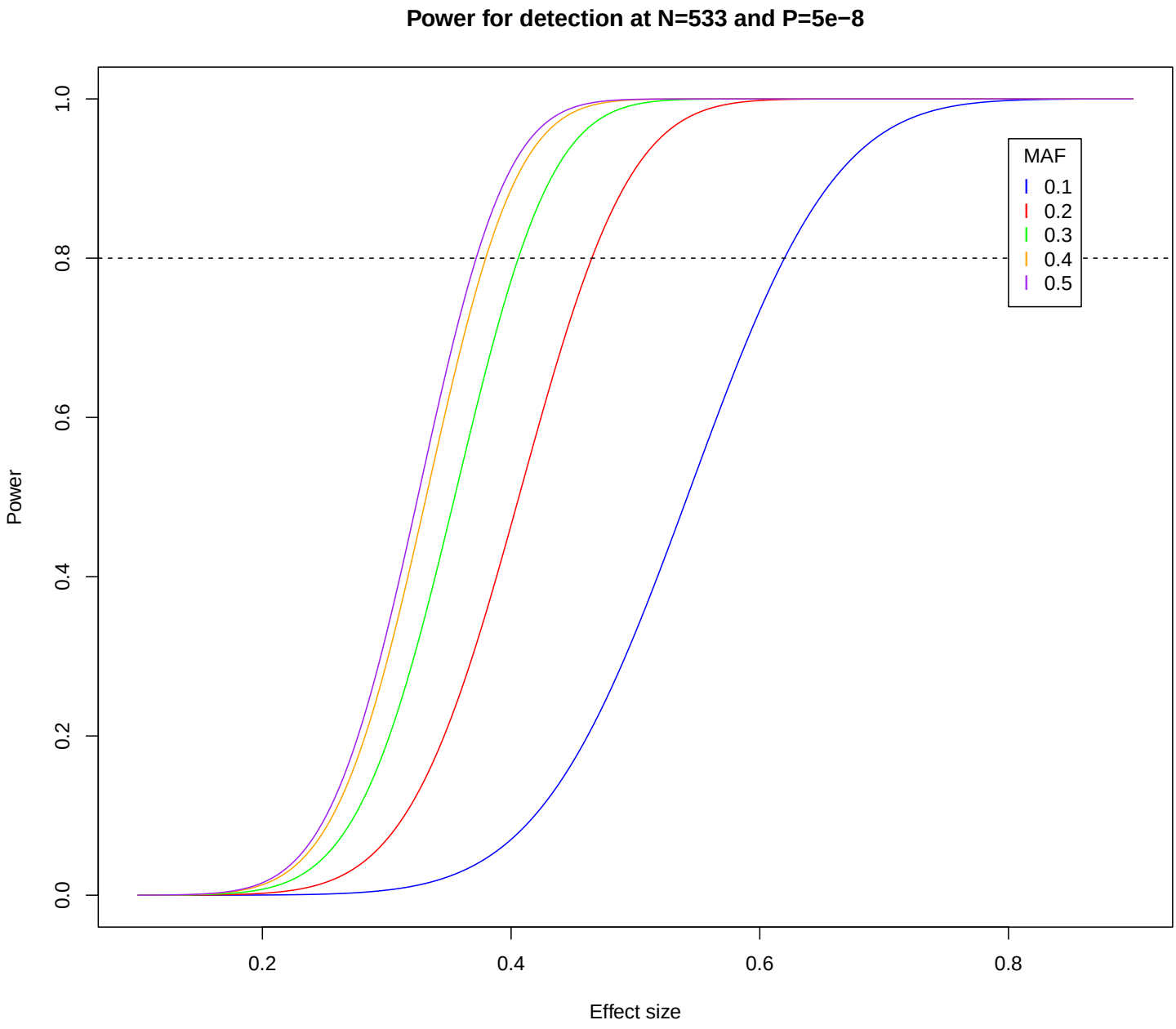


Figure A.4: Power curves for N=533 and genome wide significance at $P = 5e{-}8$. Effect size is denoted on the x-axis and power on the y-axis. Different minor allele frequencies are colored lines.

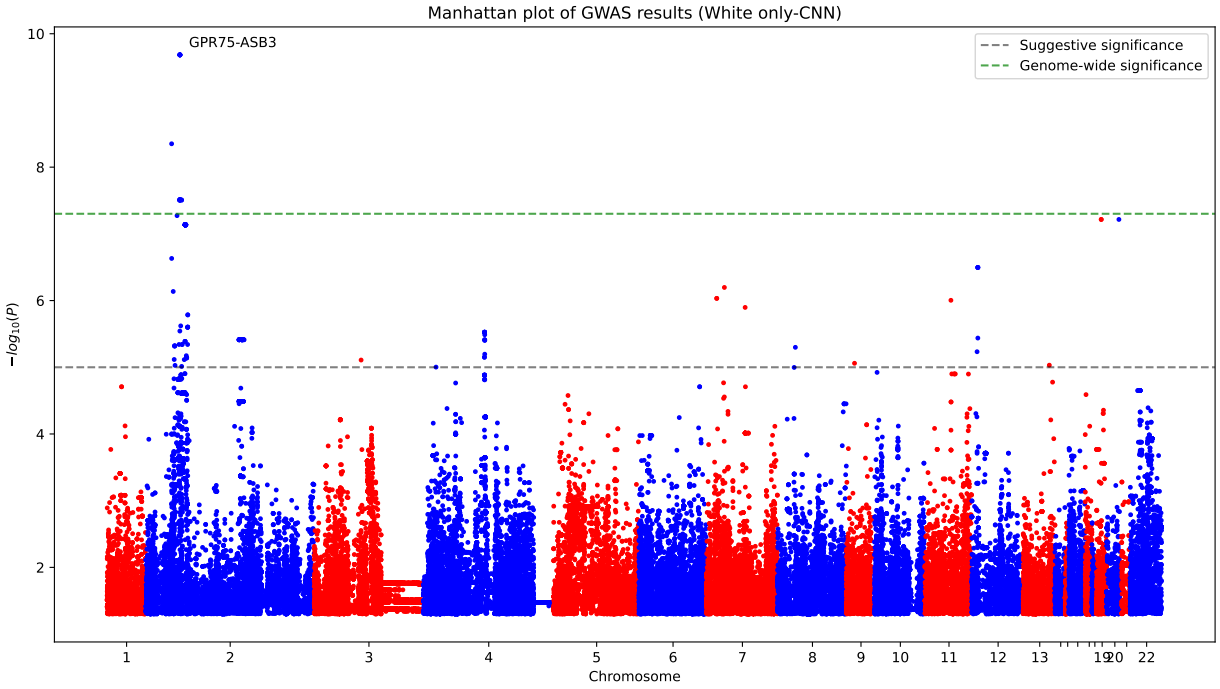


Figure A.5: Manhattan plot of GWAS results of CNN score in white ethnicity samples only.

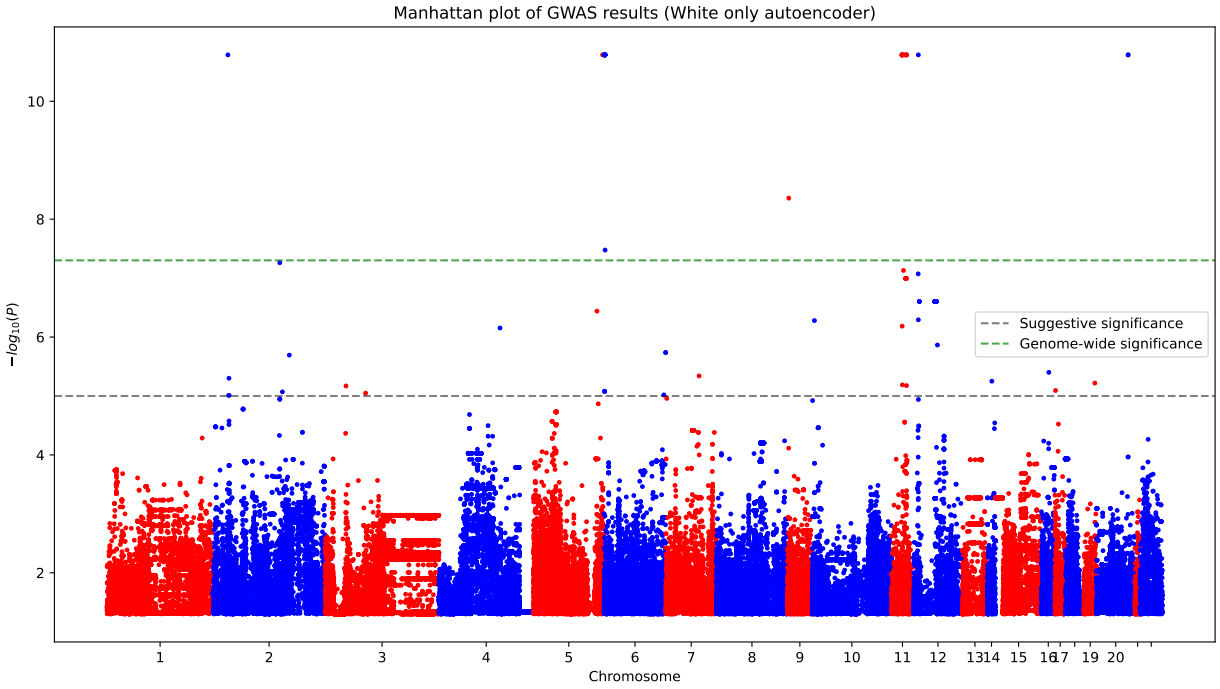


Figure A.6: Manhattan plot of GWAS results of autoencoder score in white ethnicity samples only.

Bibliography

1. *The American Psychiatric Associtation: What is psychiatry?* https://www.psychiatry.org/patients-families/what-is-psychiatry. Accessed: 2023-09-29.
2. Association, A. P. *Diagnostic and statistical manual of mental disorders: DSM-5* 5th ed. (Autor, Washington, DC, 2013).
3. *The Neurological Alliance, UK: What is a neurological condition?* https://www.neural.org.uk/about-us/about-neurological-conditions/. Accessed: 2023-09-25.
4. *National Institute on Aging: How is Alzheimer's Disease diagnosed?* https://www.nia.nih.gov/health/how-alzheimers-disease-diagnosed. Accessed: 29-09-2023.
5. *World Health Organisation: Disability-adjusted life years* https://www.who.int/data/gho/indicator-metadata-registry/imr-details/158. Accessed: 29-09-2023.
6. Ding, C. *et al.* Global, regional, and national burden and attributable risk factors of neurological disorders: The Global Burden of Disease study 1990–2019. *Frontiers in Public Health* **10,** 952161 (2022).
7. Arias, D., Saxena, S. & Verguet, S. Quantifying the global burden of mental disorders and their economic value. *EClinicalMedicine* **54** (2022).
8. WHO, A. World health statistics 2016: monitoring health for the SDGs sustainable development goals. *World Health Organization* (2016).
9. George, P., Jones, N., Goldman, H. & Rosenblatt, A. Cycles of reform in the history of psychosis treatment in the United States. *SSM-Mental Health* **3,** 100205 (2023).
10. Gardner-Thorpe, D. C. A SHORT HISTORY OF NEUROLOGY. *Brain* **123,** 2573–2575 (2000).
11. Yang, H. D., Lee, S. B., Young, L. D., *et al.* History of Alzheimer's disease. *Dementia and neurocognitive disorders* **15,** 115–121 (2016).

12. Van Hoesen, G. W., Augustinack, J. C., Dierking, J., Redman, S. J. & Thangavel, R. The parahippocampal gyrus in Alzheimer's disease: clinical and preclinical neuroanatomical correlates. *Annals of the New York Academy of Sciences* **911,** 254–274 (2000).

13. Hajek, T., Carrey, N. & Alda, M. Neuroanatomical abnormalities as risk factors for bipolar disorder. *Bipolar disorders* **7,** 393–403 (2005).

14. Watson, J. D. & Crick, F. H. *The structure of DNA* in *Cold Spring Harbor symposia on quantitative biology* **18** (1953), 123–131.

15. Collins, F. S., Morgan, M. & Patrinos, A. The Human Genome Project: lessons from large-scale biology. *Science* **300,** 286–290 (2003).

16. Behjati, S. & Tarpey, P. S. What is next generation sequencing? *Archives of Disease in Childhood-Education and Practice* **98,** 236–238 (2013).

17. Verweij, K. J., Mosing, M. A., Zietsch, B. P. & Medland, S. E. Estimating heritability from twin studies. *Statistical human genetics: methods and protocols,* 151–170 (2012).

18. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nature reviews genetics* **6,** 95–108 (2005).

19. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273,** 1516–1517 (1996).

20. Edvardsen, J. *et al.* Heritability of bipolar spectrum disorders. Unity or heterogeneity? *Journal of affective disorders* **106,** 229–240 (2008).

21. Egeland, J. A. *et al.* Bipolar affective disorders linked to DNA markers on chromosome 11. *Nature* **325,** 783–787 (1987).

22. Roses, A. D. On the discovery of the genetic association of Apolipoprotein E genotypes and common late-onset Alzheimer disease. *Journal of Alzheimer's Disease* **9,** 361–366 (2006).

23. Martens, Y. A. *et al.* ApoE Cascade Hypothesis in the pathogenesis of Alzheimer's disease and related dementias. *Neuron* **110,** 1304–1317 (2022).

24. *Brain Disorder Research* https://research-and-innovation.ec.europa.eu/research-area/health/brain-research-en. Accessed: 31-01-2024.

25. Gibson, G. Rare and common variants: twenty arguments. *Nature Reviews Genetics* **13,** 135–145 (2012).

26. Smith, E. N. *et al.* Genome-wide association study of bipolar disorder in European American and African American individuals. *Molecular psychiatry* **14,** 755–763 (2009).

27. Ferreira, M. A. *et al.* Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nature genetics* **40,** 1056–1058 (2008).

28. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nature genetics* **43,** 977–983 (2011).

29. Liu, Y. *et al.* Meta-analysis of genome-wide association data of bipolar disorder and major depressive disorder. *Molecular psychiatry* **16,** 2–4 (2011).

30. Grupe, A. *et al.* Evidence for novel susceptibility genes for late-onset Alzheimer's disease from a genome-wide association study of putative functional variants. *Human molecular genetics* **16,** 865–873 (2007).

31. Li, H. *et al.* Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. *Archives of neurology* **65,** 45–53 (2008).

32. Harold, D. *et al.* Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nature genetics* **41,** 1088–1093 (2009).

33. Bycroft, C. *et al.* Genome-wide genetic data on˜ 500,000 UK Biobank participants. *BioRxiv,* 166298 (2017).

34. *Psychiatric Genomics Consortium: About us* `https://pgc.unc.edu/about-us/`. Accessed: 03-02-2024.

35. Wightman, D. P. *et al.* A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nature genetics* **53,** 1276–1282 (2021).

36. Howard, D. M. *et al.* Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nature neuroscience* **22,** 343–352 (2019).

37. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604,** 502–508 (2022).

38. Mullins, N. *et al.* Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. *Nature genetics* **53,** 817–829 (2021).

39. Stahl, E. A. *et al.* Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nature genetics* **51,** 793–803 (2019).

40. Koromina, M. *et al.* Fine-mapping genomic loci refines bipolar disorder risk genes. *medRxiv,* 2024–02 (2024).

41. Griciuc, A. & Tanzi, R. E. The role of innate immune genes in Alzheimer's disease. *Current opinion in neurology* **34,** 228–236 (2021).

42. Wiltse, L. L. & Pait, T. G. Herophilus of Alexandria (325-255 BC): The father of anatomy. *Spine* **23,** 1904–1914 (1998).

43. Hamilton, L. W. & Hamilton, L. W. A Brief History of the Study of Neuroanatomy. *Basic Limbic System Anatomy of the Rat,* 1–6 (1976).

44. Geva, T. Magnetic resonance imaging: historical perspective. *Journal of cardiovascular magnetic resonance* **8,** 573–580 (2006).

45. Stoll, A. L., Renshaw, P. F., Yurgelun-Todd, D. A. & Cohen, B. M. Neuroimaging in bipolar disorder: what have we learned? *Biological Psychiatry* **48,** 505–517 (2000).

46. Scheltens, P. & Korf, E. S. Contribution of neuroimaging in the diagnosis of Alzheimer's disease and other dementias. *Current opinion in neurology* **13,** 391–396 (2000).

47. Ascoli, G. A. Progress and perspectives in computational neuroanatomy. *The Anatomical Record: An Official Publication of the American Association of Anatomists* **257,** 195–207 (1999).

48. Fischl, B. & Dale, A. M. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences* **97,** 11050–11055 (2000).

49. McDonald, C. *et al.* Meta-analysis of magnetic resonance imaging brain morphometry studies in bipolar disorder. *Biological psychiatry* **56,** 411–417 (2004).

50. Strakowski, S., Delbello, M. & Adler, C. The functional neuroanatomy of bipolar disorder: a review of neuroimaging findings. *Molecular psychiatry* **10,** 105–116 (2005).

51. Kantarci, K. & Jack, C. R. Neuroimaging in Alzheimer disease: an evidence-based review. *Neuroimaging Clinics* **13,** 197–209 (2003).

52. Masdeu, J. C., Zubieta, J. L. & Arbizu, J. Neuroimaging as a marker of the onset and progression of Alzheimer's disease. *Journal of the neurological sciences* **236,** 55–64 (2005).

53. Jack Jr, C. R. *et al.* The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* **27,** 685–691 (2008).

54. Ching, C. R. *et al.* What we learn about bipolar disorder from large-scale neuroimaging: Findings and future directions from the ENIGMA Bipolar Disorder Working Group. *Human brain mapping* **43,** 56–82 (2022).

55. Hibar, D. *et al.* Subcortical volumetric abnormalities in bipolar disorder. *Molecular psychiatry* **21,** 1710–1716 (2016).

56. Van Gils, M. *et al. Discovery and use of efficient biomarkers for objective disease state assessment in Alzheimer's disease* in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology* (2010), 2886–2889.

57. Cox, S. R., Ritchie, S. J., Fawns-Ritchie, C., Tucker-Drob, E. M. & Deary, I. J. Structural brain imaging correlates of general intelligence in UK Biobank. *Intelligence* **76,** 101376 (2019).

58. Hibar, D. *et al.* Cortical abnormalities in bipolar disorder: an MRI analysis of 6503 individuals from the ENIGMA Bipolar Disorder Working Group. *Molecular psychiatry* **23,** 932–942 (2018).

59. Guglielmo, R., Miskowiak, K. W. & Hasler, G. Evaluating endophenotypes for bipolar disorder. *International Journal of Bipolar Disorders* **9,** 1–20 (2021).

60. Halliday, G. Pathology and hippocampal atrophy in Alzheimer's disease. *The Lancet Neurology* **16,** 862–864 (2017).

61. Perl, D. P. Neuropathology of Alzheimer's disease. *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine: A Journal of Translational and Personalized Medicine* **77,** 32–42 (2010).

62. Weiner, M. W. *et al.* The Alzheimer's disease neuroimaging initiative: progress report and future plans. *Alzheimer's & Dementia* **6,** 202–211 (2010).

63. Littlejohns, T. J. *et al.* The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nature Communications* **11,** 1–12 (2020).

64. Smith, S. M., Vidaurre, D., Alfaro-Almagro, F., Nichols, T. E. & Miller, K. L. Estimation of brain age delta from brain imaging. *Neuroimage* **200,** 528–539 (2019).

65. Harris, M. A. *et al.* Structural neuroimaging measures and lifetime depression across levels of phenotyping in UK biobank. *Translational psychiatry* **12,** 157 (2022).

66. Elliott, L. T. *et al.* Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature* **562,** 210–216 (2018).

67. Thompson, P. M. *et al.* The enhancing NeuroImaging genetics through meta-analysis consortium: 10 Years of global collaborations in human brain mapping. *Human brain mapping* **43,** 15–22 (2022).

68. Glahn, D. C. *et al.* Arguments for the sake of endophenotypes: examining common misconceptions about the use of endophenotypes in psychiatric genetics. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **165,** 122–130 (2014).

69. Gottesman, I. I. & Shields, J. in *Acad. Press, New York, NY* (1972).

70. Hasler, G., Drevets, W. C., Gould, T. D., Gottesman, I. I. & Manji, H. K. Toward constructing an endophenotype strategy for bipolar disorders. *Biological psychiatry* **60,** 93–105 (2006).

71. Reitz, C. & Mayeux, R. Endophenotypes in normal brain morphology and Alzheimer's disease: a review. *Neuroscience* **164,** 174–190 (2009).

72. Braskie, M. N., Ringman, J. M., Thompson, P. M., *et al.* Neuroimaging measures as endophenotypes in Alzheimer's disease. *International journal of Alzheimer's disease* **2011** (2011).

73. Diehl, C. K., Rockstroh, B., Yee, C. M. & Miller, G. A. Endophenotypes in psychiatric genomics: A selective review of their status and a call to action. *Psychiatric Genomics,* 361–384 (2022).

74. Hu, B. *et al.* Genetic and environment effects on structural neuroimaging endophenotype for bipolar disorder: a novel molecular approach. *Translational Psychiatry* **12,** 137 (2022).

75. Bharthur Sanjay, A. *et al.* Characterization of gene expression patterns in mild cognitive impairment using a transcriptomics approach and neuroimaging endophenotypes. *Alzheimer's & Dementia* **18,** 2493–2508 (2022).

76. Guimond, S., Mothi, S. S., Makowski, C., Chakravarty, M. M. & Keshavan, M. S. Altered amygdala shape trajectories and emotion recognition in youth at familial high risk of schizophrenia who develop psychosis. *Translational Psychiatry* **12,** 202 (2022).

77. Brier, M. R., Thomas, J. B. & Ances, B. M. Network dysfunction in Alzheimer's disease: refining the disconnection hypothesis. *Brain connectivity* **4,** 299–311 (2014).

78. O'Donoghue, S., Holleran, L., Cannon, D. M. & McDonald, C. Anatomical dysconnectivity in bipolar disorder compared with schizophrenia: A selective review of structural network analyses using diffusion MRI. *Journal of Affective Disorders* **209,** 217–228 (2017).

79. Abdi, H. A neural network primer. *Journal of Biological Systems* **2,** 247–281 (1994).

80. LeCun, Y. *et al.* Backpropagation applied to handwritten zip code recognition. *Neural computation* **1,** 541–551 (1989).

81. LeCun, Y., Touresky, D., Hinton, G. & Sejnowski, T. *A theoretical framework for back-propagation* in *Proceedings of the 1988 connectionist models summer school* **1** (1988), 21–28.

82. Lawrence, S., Giles, C. L., Tsoi, A. C. & Back, A. D. Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks* **8,** 98–113 (1997).

83. Lo, S.-C. B. *et al.* Artificial convolution neural network for medical image pattern recognition. *Neural networks* **8,** 1201–1214 (1995).

84. Platt, J. & Nowlan, S. A convolutional neural network hand tracker. *Proc. Adv. Neural Inf. Process. Syst,* 901–908 (1995).

85. Waldrop, M. M. The chips are down for Moore's law. *Nature News* **530,** 144 (2016).

86. Pandey, M. *et al.* The transformational role of GPU computing and deep learning in drug discovery. *Nature Machine Intelligence* **4,** 211–221 (2022).

87. Sarraf, S. & Tofighi, G. Classification of alzheimer's disease using fmri data and deep learning convolutional neural networks. *arXiv preprint arXiv:1603.08631* (2016).

88. Qureshi, M. N. I., Oh, J. & Lee, B. 3D-CNN based discrimination of schizophrenia using resting-state fMRI. *Artificial intelligence in medicine* **98,** 10–17 (2019).

89. Li, Z. *et al.* Deep learning based automatic diagnosis of first-episode psychosis, bipolar disorder and healthy controls. *Computerized Medical Imaging and Graphics* **89,** 101882 (2021).

90. Taheri Gorji, H. & Kaabouch, N. A Deep Learning approach for Diagnosis of Mild Cognitive Impairment Based on MRI Images. *Brain Sciences* **9,** 217. ISSN: 2076-3425. `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6770590/` (2021) (Aug. 2019).

91. Spasov, S. E., Passamonti, L., Duggento, A., Lio, P. & Toschi, N. A Multi-modal Convolutional Neural Network Framework for the Prediction of Alzheimer's Disease. eng. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference* **2018,** 1271–1274. ISSN: 2694-0604 (July 2018).

92. Liu, M. *et al.* A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease. en. *NeuroImage* **208,** 116459. ISSN: 1053-8119. `https://www.sciencedirect.com/science/article/pii/S105381191931050X` (2021) (Mar. 2020).

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

94. Pereira, S., Pinto, A., Alves, V. & Silva, C. A. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE transactions on medical imaging* **35,** 1240–1251 (2016).

95. Simonyan, K. & Zisserman, A. *Very Deep Convolutional Networks for Large-Scale Image Recognition* 2015. arXiv: 1409.1556 [cs.CV].

96. Selvaraju, R. R. *et al.* Grad-CAM: Why did you say that? *arXiv preprint arXiv:1611.07450* (2016).

97. Keane, M. T. & Smyth, B. Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI). *CoRR* **abs/2005.13997.** arXiv: 2005.13997. https://arxiv.org/abs/2005.13997 (2020).

98. Adebayo, J. *et al.* Sanity checks for saliency maps. *arXiv preprint arXiv:1810.03292* (2018).

99. O'Connell, S., Cannon, D. M. & Broin, P. Ó. Predictive modelling of brain disorders with magnetic resonance imaging: A systematic review of modelling practices, transparency, and interpretability in the use of convolutional neural networks. *Human Brain Mapping* **n/a.** eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.26521. https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.26521.

100. James, S. L. *et al.* Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. English. *The Lancet* **392.** Publisher: Elsevier, 1789–1858. ISSN: 0140-6736, 1474-547X. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(18)32279-7/abstract (2021) (Nov. 2018).

101. Kupfer, D. J., First, M. B. & Regier, D. A. *A research agenda for DSM V* (American Psychiatric Pub, 2008).

102. Taber, K. H., Hurley, R. A. & Yudofsky, S. C. Diagnosis and treatment of neuropsychiatric disorders. *Annual review of medicine* **61,** 121–133 (2010).

103. Shah, J. & Scott, J. Concepts and misconceptions regarding clinical staging models. *Journal of psychiatry & neuroscience: JPN* **41,** E83 (2016).

104. Grover, V. P. *et al.* Magnetic resonance imaging: principles and techniques: lessons for clinicians. *Journal of clinical and experimental hepatology* **5,** 246–255 (2015).

105. Milham, M. P., Craddock, R. C. & Klein, A. Clinically useful brain imaging for neuropsychiatry: How can we get there? *Depression and anxiety* **34,** 578–587 (2017).

106. Roh, J. H. *et al.* Volume reduction in subcortical regions according to severity of Alzheimer's disease. *Journal of neurology* **258,** 1013–1020 (2011).

107. Carroll, B. J. Biomarkers in DSM-5: lost in translation. *Australian & New Zealand Journal of Psychiatry* **47,** 676–678 (2013).

108. Furiea, G. S. K. & Gisele, S. Biomarkers in neurology. *Frontiers of neurology and neuroscience* **25,** 55–61 (2009).

109. Reuter, M., Schmansky, N. J., Rosas, H. D. & Fischl, B. Within-Subject Template Estimation for Unbiased Longitudinal Image Analysis. *NeuroImage* **61,** 1402–1418. `http://dx.doi.org/10.1016/j.neuroimage.2012.02.084` (2012).

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

111. Botvinik-Nezer, R. *et al.* Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* **582,** 84–88 (2020).

112. Kamnitsas, K. *et al.* Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical image analysis* **36,** 61–78 (2017).

113. Ueda, M. *et al. An age estimation method using 3D-CNN from brain MRI images* in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)* (2019), 380–383.

114. Zou, L., Zheng, J., Miao, C., Mckeown, M. J. & Wang, Z. J. 3D CNN Based Automatic Diagnosis of Attention Deficit Hyperactivity Disorder Using Functional and Structural MRI. *IEEE Access* **5.** Conference Name: IEEE Access, 23626–23636. ISSN: 2169-3536 (2017).

115. Zhang, Z. *et al.* Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. *Annals of translational medicine* **6** (2018).

116. Yam, J. Y. & Chow, T. W. A weight initialization method for improving training speed in feedforward neural network. *Neurocomputing* **30,** 219–232 (2000).

117. LeCun, Y., Bengio, Y., *et al.* Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks* **3361,** 1995 (1995).

118. LeCun, Y. A., Bottou, L., Orr, G. B. & Müller, K.-R. in *Neural networks: Tricks of the trade* 9–48 (Springer, 2012).

119. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement. *Circulation* **131,** 211–219 (2015).

120. Haibe-Kains, B. *et al.* Transparency and reproducibility in artificial intelligence. *Nature* **586,** E14–E16 (2020).

121. Hosseini-Asl, E. *et al.* Alzheimer's disease diagnostics by a 3D deeply supervised adaptable convolutional network. eng. *Frontiers in Bioscience (Landmark Edition)* **23,** 584–596. ISSN: 1093-4715 (Jan. 2018).

122. Weiss, K., Khoshgoftaar, T. M. & Wang, D. A survey of transfer learning. *Journal of Big data* **3,** 1–40 (2016).

123. Billones, C. D., Demetria, O. J. L. D., Hostallero, D. E. D. & Naval, P. C. *DemNet: a convolutional neural network for the detection of Alzheimer's disease and mild cognitive impairment* in *2016 IEEE region 10 conference (TENCON)* (2016), 3724–3727.

124. Barbaroux, H., Feng, X., Yang, J., Laine, A. F. & Angelini, E. D. *Encoding Human Cortex Using Spherical CNNs - A Study on Alzheimer's Disease Classification* in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)* ISSN: 1945-8452 (Apr. 2020), 1322–1325.

125. Aderghal, K., Benois-Pineau, J., Afdel, K. & Catheline, G. FuseMe: Classification of sMRI images by fusion of Deep CNNs in 2D+e projections. *CBMI* (2017).

126. Pelka, O. *et al.* Sociodemographic data and APOE-ϵ4 augmentation for MRI-based detection of amnestic mild cognitive impairment using deep learning systems. eng. *PloS One* **15,** e0236868. ISSN: 1932-6203 (2020).

127. Walsh, I. *et al.* DOME: recommendations for supervised machine learning validation in biology. *Nature methods* **18,** 1122–1127 (2021).

128. Goldacre, B., Morton, C. E. & DeVito, N. J. *Why researchers should share their analytic code* 2019.

129. Eglen, S. J. *et al.* Toward standard practices for sharing computer code and programs in neuroscience. *Nature neuroscience* **20,** 770–773 (2017).

130. Markowetz, F. Five selfish reasons to work reproducibly. *Genome biology* **16,** 1–4 (2015).

131. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* **267,** 1–38 (2019).

132. Lepri, B., Oliver, N., Letouzé, E., Pentland, A. & Vinck, P. Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology* **31,** 611–627 (2018).

133. Hooker, S. Moving beyond "algorithmic bias is a data problem". *Patterns* **2,** 100241. ISSN: 2666-3899. `https://www.sciencedirect.com/science/article/pii/S2666389921000611` (2021).

134. Page, M. J. *et al.* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Bmj* **372** (2021).

135. Ahmed, S., Kim, B. C., Lee, K. H., Jung, H. Y. & Initiative, f. t. A. D. N. Ensemble of ROI-based convolutional neural network classifiers for staging the Alzheimer disease spectrum from magnetic resonance imaging. en. *PLOS ONE* **15.** Publisher: Public Library of Science, e0242712. ISSN: 1932-6203. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0242712 (2021) (Dec. 2020).

136. Mendoza-Léon, R., Puentes, J., Uriza, L. F. & Hernández Hoyos, M. Single-slice Alzheimer's disease classification and disease regional analysis with Supervised Switching Autoencoders. eng. *Computers in Biology and Medicine* **116,** 103527. ISSN: 1879-0534 (Jan. 2020).

137. Herzog, N. J. & Magoulas, G. D. Brain Asymmetry Detection and Machine Learning Classification for Diagnosis of Early Dementia. *Sensors (Basel, Switzerland)* **21,** 778. ISSN: 1424-8220. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7865614/ (2021) (Jan. 2021).

138. Sarraf, S., Desouza, D. D., Anderson, J., Saverino, C. & Alzheimer's Disease Neuroimaging Initiative. MCADNNet: Recognizing Stages of Cognitive Impairment through Efficient Convolutional fMRI and MRI Neural Network Topology Models. eng. *IEEE access: practical innovations, open solutions* **7,** 155584–155600. ISSN: 2169-3536 (2019).

139. Lian, C., Liu, M., Pan, Y. & Shen, D. Attention-Guided Hybrid Network for Dementia Diagnosis With Structural MR Images. *IEEE Transactions on Cybernetics.* Conference Name: IEEE Transactions on Cybernetics, 1–12. ISSN: 2168-2275 (2020).

140. Li, F. & Liu, M. Alzheimer's disease diagnosis based on multiple cluster dense convolutional networks. en. *Computerized Medical Imaging and Graphics* **70,** 101–110. ISSN: 0895-6111. https://www.sciencedirect.com/science/article/pii/S089561111830199X (2021) (Dec. 2018).

141. Cui, R. & Liu, M. Hippocampus Analysis by Combination of 3-D DenseNet and Shapes for Alzheimer's Disease Diagnosis. eng. *IEEE journal of biomedical and health informatics* **23,** 2099–2107. ISSN: 2168-2208 (Sept. 2019).

142. Liu, M., Zhang, J., Nie, D., Yap, P.-T. & Shen, D. Anatomical Landmark Based Deep Feature Representation for MR Images in Brain Disease Diagnosis. *IEEE journal of biomedical and health informatics* **22,** 1476–1485. ISSN: 2168-2194. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6238951/ (2021) (Sept. 2018).

143. Liu, M., Zhang, J., Adeli, E. & Shen, D. Landmark-based deep multi-instance learning for brain disease diagnosis. en. *Medical Image Analysis* **43,** 157–168. ISSN: 1361-8415. https://www.sciencedirect.com/science/article/pii/S1361841517301524 (2021) (Jan. 2018).

144. Folego, G., Weiler, M., Casseb, R. F., Pires, R. & Rocha, A. Alzheimer's Disease Detection Through Whole-Brain 3D-CNN MRI. English. *Frontiers in Bioengineering and Biotechnology* **8.** Publisher: Frontiers. ISSN: 2296-4185. https://www.frontiersin.org/articles/10.3389/fbioe.2020.534592/full (2021) (2020).

145. Lin, W. *et al.* Convolutional Neural Networks-Based MRI Image Analysis for the Alzheimer's Disease Prediction From Mild Cognitive Impairment. English. *Frontiers in Neuroscience* **12.** Publisher: Frontiers. ISSN: 1662-453X. https://www.frontiersin.org/articles/10.3389/fnins.2018.00777/full (2021) (2018).

146. Böhle, M., Eitel, F., Weygandt, M. & Ritter, K. Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification. English. *Frontiers in Aging Neuroscience* **11.** Publisher: Frontiers. ISSN: 1663-4365. https://www.frontiersin.org/articles/10.3389/fnagi.2019.00194/full (2021) (2019).

147. Qiu, S. *et al.* Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. eng. *Brain: A Journal of Neurology* **143,** 1920–1933. ISSN: 1460-2156 (June 2020).

148. Spasov, S., Passamonti, L., Duggento, A., Liò, P. & Toschi, N. A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. en. *NeuroImage* **189,** 276–287. ISSN: 1053-8119. https://www.sciencedirect.com/science/article/pii/S105381191930031X (2021) (Apr. 2019).

149. Hu, J. *et al.* Deep Learning-Based Classification and Voxel-Based Visualization of Frontotemporal Dementia and Alzheimer's Disease. English. *Frontiers in Neuroscience* **14.** Publisher: Frontiers. ISSN: 1662-453X. https://www.frontiersin.org/articles/10.3389/fnins.2020.626154/full (2021) (2021).

150. Li, F. & Liu, M. A hybrid Convolutional and Recurrent Neural Network for Hippocampus Analysis in Alzheimer's Disease. en. *Journal of Neuroscience Methods* **323,** 108–118. ISSN: 0165-0270. https://www.sciencedirect.com/science/article/pii/S0165027019301463 (2021) (July 2019).

151. Li, F., Cheng, D. & Liu, M. Alzheimer's disease classification based on combination of multi-model convolutional networks. *2017 IEEE International Conference on Imaging Systems and Techniques (IST)* (2017).

152. Marzban, E. N., Eldeib, A. M., Yassine, I. A., Kadah, Y. M. & Initiative, f. t. A. D. N. Alzheimer's disease diagnosis from diffusion tensor images using convolutional neural networks. en. *PLOS ONE* **15.** Publisher: Public Library of Science, e0230409. ISSN: 1932-6203. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0230409 (2021) (Mar. 2020).

153. Gunawardena, K. A. N. N. P., Rajapakse, R. N. & Kodikara, N. D. *Applying convolutional neural networks for pre-detection of alzheimer's disease from structural MRI data* in *2017 24th International Conference on Mechatronics and Machine Vision in Practice (M2VIP)* (Nov. 2017), 1–7.

154. Basaia, S. *et al.* Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. en. *NeuroImage: Clinical* **21,** 101645. ISSN: 2213-1582. https://www.sciencedirect.com/science/article/pii/S2213158218303930 (2021) (Jan. 2019).

155. Tufail, A. B., Zhang, Q.-N. & Ma, Y.-K. Binary Classification of Alzheimer Disease using sMRI Imaging modality and Deep Learning. *Journal of Digital Imaging* **33.** arXiv: 1809.06209, 1073–1090. ISSN: 0897-1889, 1618-727X. http://arxiv.org/abs/1809.06209 (2021) (Oct. 2020).

156. Hu, M., Sim, K., Zhou, J. H., Jiang, X. & Guan, C. Brain MRI-based 3D Convolutional Neural Networks for Classification of Schizophrenia and Controls. eng. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference* **2020,** 1742–1745. ISSN: 2694-0604 (July 2020).

157. Cheng, D., Liu, M., Fu, J. & Wang, Y. Classification of MR brain images by combination of multi-CNNs for AD diagnosis. **10420.** Conference Name: Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 1042042. https://ui.adsabs.harvard.edu/abs/2017SPIE10420E..42C (2021) (July 2017).

158. Nanni, L. *et al.* Comparison of Transfer Learning and Conventional Machine Learning Applied to Structural Brain MRI for the Early Diagnosis and Prognosis of Alzheimer's Disease. English. *Frontiers in Neurology* **11.** Publisher: Frontiers. ISSN: 1664-2295. https://www.frontiersin.org/articles/10.3389/fneur.2020.576194/full (2021) (2020).

159. Yigit, A. & Işik, Z. Applying deep learning models to structural MRI for stage prediction of Alzheimer's disease. *Turkish J. Electr. Eng. Comput. Sci.* (2020).

160. Pan, D. *et al.* Early Detection of Alzheimer's Disease Using Magnetic Resonance Imaging: A Novel Approach Combining Convolutional Neural Networks and Ensemble Learning. *Frontiers in Neuroscience* **14,** 259. ISSN: 1662-4548. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7238823/ (2021) (May 2020).

161. Ortiz-Suárez, J. M., Ramos-Pollán, R. & Romero, E. *Exploring Alzheimer's anatomical patterns through convolutional networks* in *12th International Symposium on Medical Information Processing and Analysis* **10160** (International Society for Optics and Photonics, Jan. 2017), 101600Z. (2021).

162. Lian, C., Liu, M., Zhang, J. & Shen, D. Hierarchical Fully Convolutional Network for Joint Atrophy Localization and Alzheimer's Disease Diagnosis Using Structural MRI. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **42.** Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence, 880–893. ISSN: 1939-3539 (Apr. 2020).

163. Li, A. *et al.* Hippocampal shape and asymmetry analysis by cascaded convolutional neural networks for Alzheimer's disease diagnosis. eng. *Brain Imaging and Behavior.* ISSN: 1931-7565 (Jan. 2021).

164. Cui, R. & Liu, M. *Hippocampus analysis based on 3D CNN for Alzheimer's disease diagnosis* in *Tenth International Conference on Digital Image Processing (ICDIP 2018)* **10806** (International Society for Optics and Photonics, Aug. 2018), 108065O. (2021).

165. Aderghal, K., Afdel, K., Benois-Pineau, J. & Catheline, G. Improving Alzheimer's stage categorization with Convolutional Neural Network using transfer learning and different magnetic resonance imaging modalities. en. *Heliyon* **6,** e05652. ISSN: 2405-8440. https://www.sciencedirect.com/science/article/pii/S2405844020324956 (2021) (Dec. 2020).

166. Liu, M., Cheng, D., Wang, K., Wang, Y. & Alzheimer's Disease Neuroimaging Initiative. Multi-Modality Cascaded Convolutional Neural Networks for Alzheimer's Disease Diagnosis. eng. *Neuroinformatics* **16,** 295–308. ISSN: 1559-0089 (Oct. 2018).

167. Zhang, J. *et al.* Three dimensional convolutional neural network-based classification of conduct disorder with structural MRI. eng. *Brain Imaging and Behavior* **14,** 2333–2340. ISSN: 1931-7565 (Dec. 2020).

168. Lee, B., Ellahi, W. & Choi, J. Using Deep CNN with Data Permutation Scheme for Classification of Alzheimer's Disease in Structural Magnetic Resonance Imaging (sMRI). *IEICE Trans. Inf. Syst.* (2019).

169. Sun, J., Yan, S., Song, C. & Han, B. Dual-functional neural network for bilateral hippocampi segmentation and diagnosis of Alzheimer's disease. eng. *International Journal of Computer Assisted Radiology and Surgery* **15,** 445–455. ISSN: 1861-6429 (Mar. 2020).

170. Oh, J., Oh, B.-L., Lee, K.-U., Chae, J.-H. & Yun, K. Identifying Schizophrenia Using Structural MRI With a Deep Learning Algorithm. *Frontiers in Psychiatry* **11,** 16. ISSN: 1664-0640. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7008229/ (2021) (Feb. 2020).

171. Cui, R. & Liu, M. RNN-based longitudinal analysis for diagnosis of Alzheimer's disease. en. *Computerized Medical Imaging and Graphics* **73,** 1–10. ISSN: 0895-6111. https://www.sciencedirect.com/science/article/pii/S0895611118303987 (2021) (Apr. 2019).

172. Bae, J. *et al.* Transfer learning for predicting conversion from mild cognitive impairment to dementia of Alzheimer's type based on a three-dimensional convolutional neural network. eng. *Neurobiology of Aging* **99,** 53–64. ISSN: 1558-1497 (Mar. 2021).

173. Al-Khuzaie, F. E. K., Bayat, O. & Duru, A. D. Diagnosis of Alzheimer Disease Using 2D MRI Slices by Convolutional Neural Network. en. *Applied Bionics and Biomechanics* **2021.** Publisher: Hindawi, e6690539. ISSN: 1176-2322. https://www.hindawi.com/journals/abb/2021/6690539/ (2021) (Feb. 2021).

174. Zhang, J. *et al.* A 3D densely connected convolution neural network with connection-wise attention mechanism for Alzheimer's disease classification. en. *Magnetic Resonance Imaging* **78,** 119–126. ISSN: 0730-725X. https://www.sciencedirect.com/science/article/pii/S0730725X21000138 (2021) (May 2021).

175. Yee, E. *et al.* Construction of MRI-Based Alzheimer's Disease Score Based on Efficient 3D Convolutional Neural Network: Comprehensive Validation on 7,902 Images from a Multi-Center Dataset. eng. *Journal of Alzheimer's disease: JAD* **79,** 47–58. ISSN: 1875-8908 (2021).

176. Mukhtar, G. & Farhan, S. Convolutional Neural Network Based Prediction of Conversion from Mild Cognitive Impairment to Alzheimer's Disease: A Technique using Hippocampus Extracted from MRI. *Advances in Electrical and Computer Engineering* **20,** 113–122 (2020).

177. Bae, J. B. *et al.* Identification of Alzheimer's disease using a convolutional neural network model based on T1-weighted magnetic resonance imaging. *Scientific Reports* **10,** 1–10 (2020).

178. Nigri, E., Ziviani, N., Cappabianco, F., Antunes, A. & Veloso, A. *Explainable deep CNNs for MRI-based diagnosis of Alzheimer's disease* in *2020 International Joint Conference on Neural Networks (IJCNN)* (2020), 1–8.

179. Kiryu, S. *et al.* Deep learning to differentiate parkinsonian disorders separately using single midsagittal MR imaging: a proof of concept study. *European radiology* **29,** 6891–6899 (2019).

180. Hutson, M. *Artificial intelligence faces reproducibility crisis* 2018.

181. Wen, J. *et al.* Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. *Medical image analysis* **63,** 101694 (2020).

182. Bagdasaryan, E., Poursaeed, O. & Shmatikov, V. Differential privacy has disparate impact on model accuracy. *Advances in Neural Information Processing Systems* **32,** 15479–15488 (2019).

183. Diakopoulos, N. Algorithmic accountability: Journalistic investigation of computational power structures. *Digital journalism* **3,** 398–415 (2015).

184. Buolamwini, J. & Gebru, T. *Gender shades: Intersectional accuracy disparities in commercial gender classification* in *Conference on fairness, accountability and transparency* (2018), 77–91.

185. Stodden, V. C. Trust your science? Open your data and code (2011).

186. *Nature Editorial Policies* https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards. Accessed: 27-10-2021.

187. *Science Editorial Policies* https://www.science.org/content/page/science-journals-editorial-policies. Accessed: 27-10-2021.

188. Google. Colaboratory: Frequently Asked Questions (2018).

189. Kluyver, T. *et al. Jupyter Notebooks – a publishing format for reproducible computational workflows* in *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (eds Loizides, F. & Schmidt, B.) (2016), 87–90.

190. Scheltens, P. *et al.* Alzheimer's disease. *The Lancet* **397,** 1577–1590 (2021).

191. Swerdlow, R. H. Is aging part of Alzheimer's disease, or is Alzheimer's disease part of aging? *Neurobiology of aging* **28,** 1465–1480 (2007).

192. Li, X. *et al.* Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2019. *Frontiers in Aging Neuroscience* **14,** 937486 (2022).

193. Henderson, D. & Maclachlan, S. H. Alzheimer's disease. *Journal of Mental Science* **76,** 646–661 (1930).

194. McMenemey, W. Alzheimer's disease: A report of six cases. *Journal of neurology and psychiatry* **3,** 211 (1940).

195. Jiang, T., Yu, J.-T., Tian, Y. & Tan, L. Epidemiology and etiology of Alzheimer's disease: from genetic to non-genetic factors. *Current Alzheimer Research* **10,** 852–867 (2013).

196. Tariska, I. in *Alzheimer's disease and related conditions* 51–69 (Churchill London, 1970).

197. Brun, A. & Gustafson, L. Distribution of cerebral degeneration in Alzheimer's disease: a clinico-pathological study. *Archiv für Psychiatrie und Nervenkrankheiten* **223,** 15–33 (1976).

198. Ward, A. *et al.* Prevalence of apolipoprotein E4 genotype and homozygotes (APOE e4/4) among patients diagnosed with Alzheimer's disease: a systematic review and meta-analysis. *Neuroepidemiology* **38,** 1–17 (2012).

199. Liu, C.-C., Kanekiyo, T., Xu, H. & Bu, G. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. *Nature Reviews Neurology* **9,** 106–118 (2013).

200. Busche, M. A. & Hyman, B. T. Synergy between amyloid-beta and tau in Alzheimer's disease. *Nature neuroscience* **23,** 1183–1193 (2020).

201. Medeiros, R., Baglietto-Vargas, D. & LaFerla, F. M. The role of tau in Alzheimer's disease and related disorders. *CNS neuroscience & therapeutics* **17,** 514–524 (2011).

202. Sisodia, S. S. & Price, D. L. Role of the β-amyloid protein in Alzheimer's disease. *The FASEB Journal* **9,** 366–370 (1995).

203. Woodworth, D. C. *et al.* Dementia is associated with medial temporal atrophy even after accounting for neuropathologies. *Brain communications* **4,** fcac052 (2022).

204. Luxenberg, J. S., Haxby, J. V., Creasey, H., Sundaram, M. & Rapoport, S. I. Rate of ventricular enlargement in dementia of the Alzheimer type correlates with rate of neuropsychological deterioration. *Neurology* **37,** 1135–1135. ISSN: 0028-3878. eprint: `https://n.neurology.org/content/37/7/1135.full.pdf`. `https://n.neurology.org/content/37/7/1135` (1987).

205. Maj, C. *et al.* Integration of machine learning methods to dissect genetically imputed transcriptomic profiles in Alzheimer's disease. *Frontiers in genetics* **10,** 466810 (2019).

206. Nord, L. I. & Jacobsson, S. P. A novel method for examination of the variable contribution to computational neural network models. *Chemometrics and intelligent laboratory systems* **44,** 153–160 (1998).

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

208. Pasa, F., Golkov, V., Pfeiffer, F., Cremers, D. & Pfeiffer, D. Efficient deep network architectures for fast chest X-ray tuberculosis screening and visualization. *Scientific reports* **9,** 6268 (2019).

209. Lucieri, A. *et al. On interpretability of deep learning based skin lesion classifiers using concept activation vectors* in *2020 international joint conference on neural networks (IJCNN)* (2020), 1–10.

210. Zhang, Y. *et al.* Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging. *Journal of Neuroscience Methods* **353,** 109098 (2021).

211. He, T. *et al.* Medimlp: using grad-cam to extract crucial variables for lung cancer postoperative complication prediction. *IEEE journal of biomedical and health informatics* **24,** 1762–1771 (2019).

212. Moujahid, H. *et al.* Combining CNN and Grad-Cam for COVID-19 Disease Prediction and Visual Explanation. *Intelligent Automation and Soft Computing* **32** (2022).

213. Matsui, T., Taki, M., Pham, T. Q., Chikazoe, J. & Jimura, K. Counterfactual explanation of brain activity classifiers using image-to-image transfer by generative adversarial network. *Frontiers in Neuroinformatics* **15,** 802938 (2022).

214. Shorten, C. & Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *Journal of big data* **6,** 1–48 (2019).

215. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

216. Browning, B. L., Zhou, Y. & Browning, S. R. A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics* **103,** 338–348 (2018).

217. Clarke, L. *et al.* The 1000 Genomes Project: data management and community access. *Nature methods* **9,** 459–462 (2012).

218. Jiang, L. *et al.* A resource-efficient tool for mixed model association analysis of large-scale data. *Nature genetics* **51,** 1749–1755 (2019).

219. Bennett, D., O'Shea, D., Ferguson, J., Morris, D. & Seoighe, C. Controlling for background genetic effects using polygenic scores improves the power of genome-wide association studies. *Scientific Reports* **11,** 19571 (2021).

220. Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nature communications* **8,** 1826 (2017).

221. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature genetics* **42,** 565–569 (2010).

222. Orre, M. *et al.* Reactive glia show increased immunoproteasome activity in Alzheimer's disease. *Brain* **136,** 1415–1431 (2013).

223. Morrow, J. S. & Stankewich, M. C. The spread of spectrin in ataxia and neurodegenerative disease. *Journal of experimental neurology* **2,** 131 (2021).

224. Zhu, B. *et al.* ER-associated degradation regulates Alzheimer's amyloid pathology and memory function by modulating gamma-secretase activity. *Nature Communications* **8,** 1472 (2017).

225. De Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS computational biology* **11,** e1004219 (2015).

226. Yuan, Q. *et al.* Comprehensive analysis of core genes and key pathways in Parkinson's disease. *American Journal of Translational Research* **12,** 5630 (2020).

227. Bessa de Sousa, D. M. *et al.* The platelet transcriptome and proteome in Alzheimer's disease and aging: an exploratory cross-sectional study. *Frontiers in Molecular Biosciences* **10,** 1196083.

228. Satterstrom, F. K. *et al.* Autism spectrum disorder and attention deficit hyperactivity disorder have a similar burden of rare protein-truncating variants. *Nature neuroscience* **22,** 1961–1965 (2019).

229. Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic acids research* **48,** D882–D889 (2020).

230. Bayraktar, A. *et al.* Revealing the molecular mechanisms of Alzheimer's disease based on network analysis. *International Journal of Molecular Sciences* **22,** 11556 (2021).

231. Wilkins, H. M. *et al.* Bioenergetic and inflammatory systemic phenotypes in Alzheimer's disease APOE ε4-carriers. *Aging cell* **20,** e13356 (2021).

232. Smith, S. M. *et al.* An expanded set of genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature neuroscience* **24,** 737–745 (2021).

233. Of Health, U. D., Services, H., *et al.* What happens to the brain in Alzheimer's disease. *National Institute on Aging* (2017).

234. Nestor, S. M. *et al.* Ventricular enlargement as a possible measure of Alzheimer's disease progression validated using the Alzheimer's disease neuroimaging initiative database. *Brain* **131,** 2443–2454 (2008).

235. Greene, S. J., Killiany, R. J., Initiative, A. D. N., *et al.* Subregions of the inferior parietal lobule are affected in the progression to Alzheimer's disease. *Neurobiology of aging* **31,** 1304–1311 (2010).

236. Ball, T. M., Squeglia, L. M., Tapert, S. F. & Paulus, M. P. Double dipping in machine learning: problems and solutions. *Biological psychiatry. Cognitive neuroscience and neuroimaging* **5,** 261 (2020).

237. Gao, L. L., Bien, J. & Witten, D. *Selective Inference for Hierarchical Clustering* 2022. arXiv: 2012.02936 [stat.ME].

238. Tan, J., Hammond, J. H., Hogan, D. A. & Greene, C. S. Adage-based integration of publicly available pseudomonas aeruginosa gene expression data with denoising autoencoders illuminates microbe-host interactions. *MSystems* **1,** e00025–15 (2016).

239. Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)* ISBN: 0387310738 (Springer-Verlag, Berlin, Heidelberg, 2006).

240. Dincer, A. B., Janizek, J. D. & Lee, S.-I. Adversarial deconfounding autoencoder for learning robust gene expression embeddings. *Bioinformatics* **36,** i573–i582 (2020).

241. Grønbech, C. H. *et al.* scVAE: variational auto-encoders for single-cell gene expression data. *Bioinformatics* **36,** 4415–4422 (2020).

242. Li, X. *et al.* Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. *Nature communications* **11,** 2338 (2020).

243. Wichura, M. J. *The coordinate-free approach to linear models* (Cambridge University Press, 2006).

244. You, K., Long, M., Wang, J. & Jordan, M. I. How does learning rate decay help modern neural networks? *arXiv preprint arXiv:1908.01878* (2019).

245. Xue, D. *et al.* Large-scale sequencing studies expand the known genetic architecture of Alzheimer's disease. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring* **13,** e12255 (2021).

246. Soleimani Zakeri, N. S., Pashazadeh, S. & MotieGhader, H. Gene biomarker discovery at different stages of Alzheimer using gene co-expression network approach. *Scientific reports* **10,** 12210 (2020).

247. Prokopenko, D. *et al.* Whole-genome sequencing reveals new Alzheimer's disease–associated rare variants in loci related to synaptic function and neuronal development. *Alzheimer's & Dementia* **17,** 1509–1527 (2021).

248. Dhikav, V., Sethi, M. & Anand, K. Medial temporal lobe atrophy in Alzheimer's disease/mild cognitive impairment with depression. *The British journal of radiology* **87,** 20140150 (2014).

249. Visser, P., Verhey, F., Hofman, P., Scheltens, P. & Jolles, J. Medial temporal lobe atrophy predicts Alzheimer's disease in patients with minor cognitive impairment. *Journal of Neurology, Neurosurgery & Psychiatry* **72,** 491–497 (2002).

250. Pegueroles, J. *et al.* Longitudinal brain structural changes in preclinical Alzheimer's disease. *Alzheimer's & Dementia* **13,** 499–509 (2017).

251. Chan, D. *et al.* Patterns of temporal lobe atrophy in semantic dementia and Alzheimer's disease. *Annals of neurology* **49,** 433–442 (2001).

252. Lee, P.-L. *et al.* Posterior cingulate cortex network predicts alzheimer's disease progression. *Frontiers in aging neuroscience* **12,** 608667 (2020).

253. Lubben, N., Ensink, E., Coetzee, G. A. & Labrie, V. The enigma and implications of brain hemispheric asymmetry in neurodegenerative diseases. *Brain Communications* **3,** fcab211 (2021).

254. Roe, J. M. *et al.* Asymmetric thinning of the cerebral cortex across the adult lifespan is accelerated in Alzheimer's disease. *Nature communications* **12,** 721 (2021).

255. Shobeiri, P. *et al.* Circulating long non-coding RNAs as novel diagnostic biomarkers for Alzheimer's disease (AD): A systematic review and meta-analysis. *Plos one* **18,** e0281784 (2023).

256. Faghihi, M. A. *et al.* Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nature medicine* **14,** 723–730 (2008).

257. Zhang, Y. *et al.* The role of non-coding RNAs in Alzheimer's disease: from regulated mechanism to therapeutic targets and diagnostic biomarkers. *Frontiers in aging neuroscience* **13,** 654978 (2021).

258. Shen, H. *et al.* 6p22. 3 amplification as a biomarker and potential therapeutic target of advanced stage bladder cancer. *Oncotarget* **4,** 2124 (2013).

259. Olsson, A. *et al.* Role of E2F3 expression in modulating cellular proliferation rate in human bladder and prostate cancer cells. *Oncogene* **26,** 1028–1037 (2007).

260. O'Connell, K. S. *et al.* Identification of genetic loci shared between attention-deficit/hyperactivity disorder, intelligence, and educational attainment. *Biological psychiatry* **87,** 1052–1062 (2020).

261. Sassi, C. *et al.* Mendelian adult-onset leukodystrophy genes in Alzheimer's disease: critical influence of CSF1R and NOTCH3. *Neurobiology of Aging* **66,** 179–e17 (2018).

262. Lake, J. *et al.* Multi-ancestry meta-analysis and fine-mapping in Alzheimer's Disease. *Molecular Psychiatry,* 1–12 (2023).

263. Burns, J., Galvin, J., Roe, C., Morris, J. & McKeel, D. The pathology of the substantia nigra in Alzheimer disease with extrapyramidal signs. *Neurology* **64,** 1397–1403 (2005).

264. Cheramy, A., Leviel, V. & Glowinski, J. Dendritic release of dopamine in the substantia nigra. *Nature* **289,** 537–543 (1981).

265. Rinne, J. *et al.* Neuronal loss in the substantia nigra in patients with Alzheimer's disease and Parkinson's disease in relation to extrapyramidal symptoms and dementia. *Progress in clinical and biological research* **317,** 325–332 (1989).

266. Chen, S., Lu, F. F., Seeman, P. & Liu, F. Quantitative proteomic analysis of human substantia nigra in Alzheimer's disease, Huntington's disease and Multiple sclerosis. *Neurochemical research* **37,** 2805–2813 (2012).

267. Pan, X. *et al.* Dopamine and dopamine receptors in Alzheimer's disease: A systematic review and network meta-analysis. *Frontiers in aging neuroscience* **11,** 175 (2019).

268. Zhang, H., Zhang, L., Zhou, D., Li, H. & Xu, Y. ErbB4 mediates amyloid beta-induced neurotoxicity through JNK/tau pathway activation: Implications for Alzheimer's disease. *Journal of Comparative Neurology* **529,** 3497–3512 (2021).

269. Mouton-Liger, F. *et al.* CSF levels of the BACE1 substrate NRG1 correlate with cognition in Alzheimer's disease. *Alzheimer's research & therapy* **12,** 1–10 (2020).

270. Liharska, L. E. *et al.* A study of gene expression in the living human brain. *medRxiv.* eprint: `https://www.medrxiv.org/content/early/2023/08/01/2023.04.21.23288916.full.pdf`. `https://www.medrxiv.org/content/early/2023/08/01/2023.04.21.23288916` (2023).

271. Littlejohns, T. J. *et al.* The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nature communications* **11,** 2624 (2020).

272. Sanderson, E. *et al.* Mendelian randomization. *Nature Reviews Methods Primers* **2,** 6 (2022).

273. Chen, L., Davey Smith, G., Harbord, R. M. & Lewis, S. J. Alcohol intake and blood pressure: a systematic review implementing a Mendelian randomization approach. *PLoS medicine* **5,** e52 (2008).

274. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human molecular genetics* **23,** R89–R98 (2014).

275. Burgess, S., Dudbridge, F. & Thompson, S. G. Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Statistics in medicine* **35,** 1880–1906 (2016).

276. Guo, J. *et al.* Mendelian randomization analyses support causal relationships between brain imaging-derived phenotypes and risk of psychiatric disorders. *Nature Neuroscience* **25,** 1519–1527 (2022).

277. Brown, B. C. & Knowles, D. A. Phenome-scale causal network discovery with bidirectional mediated Mendelian randomization. *bioRxiv.* eprint: `https://www.biorxiv.org/content/early/2020/06/20/2020.06.18.160176.full.pdf`. `https://www.biorxiv.org/content/early/2020/06/20/2020.06.18.160176` (2020).

278. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics* **47,** 291–295 (2015).

279. Bowden, J. *et al.* Improving the visualization, interpretation and analysis of two-sample summary data Mendelian randomization via the Radial plot and Radial regression. *International journal of epidemiology* **47,** 1264–1278 (2018).

280. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International journal of epidemiology* **44,** 512–525 (2015).

281. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7,** e34408. `https://elifesciences.org/articles/34408` (2018).

282. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13,** 2498–2504 (2003).

283. Abé, C. *et al.* Longitudinal structural brain changes in bipolar disorder: a multicenter neuroimaging study of 1232 individuals by the ENIGMA bipolar disorder working group. *Biological psychiatry* **91,** 582–592 (2022).

284. Abé, C. *et al.* Cortical thickness, volume and surface area in patients with bipolar disorder types I and II. *Journal of Psychiatry and Neuroscience* **41,** 240–250 (2016).

285. Cheng, W. *et al.* Medial reward and lateral non-reward orbitofrontal cortex circuits change in opposite directions in depression. *Brain* **139,** 3296–3309 (2016).

286. Moncrieff, J. & Leo, J. A systematic review of the effects of antipsychotic drugs on brain volume. *Psychological medicine* **40,** 1409–1422 (2010).

287. Abé, C. *et al.* Longitudinal cortical thickness changes in bipolar disorder and the relationship to genetic risk, mania, and lithium use. *Biological Psychiatry* **87,** 271–281 (2020).

288. Chepenik, L. G. *et al.* Structure–function associations in hippocampus in bipolar disorder. *Biological psychology* **90,** 18–22 (2012).

289. Shin, S.-J., Kim, A., Han, K.-M., Tae, W.-S. & Ham, B.-J. Reduced sulcal depth in central sulcus of major depressive disorder. *Experimental Neurobiology* **31,** 353 (2022).

290. Mahon, K. *et al.* A voxel-based diffusion tensor imaging study of white matter in bipolar disorder. *Neuropsychopharmacology* **34,** 1590–1600 (2009).

291. Figley, C. R. *et al.* Potential pitfalls of using fractional anisotropy, axial diffusivity, and radial diffusivity as biomarkers of cerebral white matter microstructure. *Frontiers in Neuroscience* **15,** 799576 (2022).

292. Walterfang, M. *et al.* Corpus callosum size and shape in established bipolar affective disorder. *Australian & New Zealand Journal of Psychiatry* **43,** 838–845 (2009).

293. Sussmann, J. E. *et al.* White matter abnormalities in bipolar disorder and schizophrenia detected using diffusion tensor magnetic resonance imaging. *Bipolar disorders* **11,** 11–18 (2009).

294. George, K. *et al.* Neuroanatomy, thalamocortical radiations (2019).

295. Strakowski, S. M. *et al.* The functional neuroanatomy of bipolar disorder: a consensus model. *Bipolar disorders* **14,** 313–325 (2012).

296. Rajmohan, V. & Mohandas, E. The limbic system. *Indian journal of psychiatry* **49,** 132 (2007).

297. *Queensland Brain Institute: The limbic system* `https://qbi.uq.edu.au/brain/brain-anatomy/limbic-system`. Accessed: 2023-09-25.

298. Wang, F. *et al.* Abnormal anterior cingulum integrity in bipolar disorder determined through diffusion tensor imaging. *The British Journal of Psychiatry* **193,** 126–129 (2008).

299. Minichino, A. *et al.* The role of cerebellum in unipolar and bipolar depression: a review of the main neurobiological findings. *Rivista di psichiatria* **49,** 124–131 (2014).

300. Shinn, A. K. *et al.* Aberrant cerebellar connectivity in bipolar disorder with psychosis. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* **2,** 438–448 (2017).

301. Saleem, A. *et al.* Functional connectivity of the cerebellar vermis in bipolar disorder and associations with mood. *Frontiers in psychiatry* **14,** 755 (2023).

302. Nabulsi, L. *et al.* Normalization of impaired emotion inhibition in bipolar disorder mediated by cholinergic neurotransmission in the cingulate cortex. *Neuropsychopharmacology* **47,** 1643–1651 (2022).

303. Favre, P. *et al.* Widespread white matter microstructural abnormalities in bipolar disorder: evidence from mega-and meta-analyses across 3033 individuals. *Neuropsychopharmacology* **44,** 2285–2293 (2019).

304. Ij, H. Statistics versus machine learning. *Nat Methods* **15,** 233 (2018).

305. Hosseini-Asl, E., Keynto, R. & El-Baz, A. Alzheimer's Disease Diagnostics by Adaptation of 3D Convolutional Network. *2016 IEEE International Conference on Image Processing (ICIP).* arXiv: 1607.00455, 126–130. http://arxiv.org/abs/1607.00455 (2021) (Sept. 2016).

306. Roberts, T. C., Morris, K. V. & Wood, M. J. The role of long non-coding RNAs in neurodevelopment, brain function and neurological disease. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369,** 20130507 (2014).

307. Uddin, M. S. *et al.* Autophagy and Alzheimer's disease: from molecular mechanisms to therapeutic implications. *Frontiers in aging neuroscience* **10,** 04 (2018).

308. Ge, M. *et al.* Role of calcium homeostasis in Alzheimer's Disease. *Neuropsychiatric Disease and Treatment* **18,** 487 (2022).

309. Arendt, H. *The Human Condition* (University of Chicago press, 1958).

310. Smith, G. D. Epidemiology, epigenetics and the 'Gloomy Prospect': embracing randomness in population health research and practice. *International Journal of Epidemiology* **40,** 537–562 (2011).

311. Eddy, S. R. "Antedisciplinary" science. *PLoS computational biology* **1,** e6 (2005).

312. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual* ISBN: 1441412697 (CreateSpace, Scotts Valley, CA, 2009).

313. Team, R. C. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2021). %5Curl%7Bhttps://www.R-project.org/%7D.

314. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585,** 357–362. https://doi.org/10.1038/s41586-020-2649-2 (Sept. 2020).

315. McKinney, W. *Data Structures for Statistical Computing in Python* in *Proceedings of the 9th Python in Science Conference* (eds van der Walt, S. & Millman, J.) (2010), 56–61.

316. Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* **9,** 90–95 (2007).

317. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* ISBN: 978-3-319-24277-4. %5Curl%7Bhttps://ggplot2.tidyverse.org%7D (Springer-Verlag New York, 2016).