INTERPRETABLE MACHINE LEARNING AND DEEP LEARNING FRAMEWORKS FOR PREDICTIVE ANALYTICS AND BIOMARKER DISCOVERY FROM MULTIMODAL IMAGING GENETICS DATA

by
Sayan Ghosal

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
October 2023

Abstract

In healthcare, neuroimaging studies and genetics research are generating torrents of data to understand the hereditary components of neurological and psychiatric disorders. Different neuroimaging techniques like functional Magnetic Resonance Imaging (f-MRI) probes into the neural functioning of the disorder. In parallel, genome sequencing technologies explore the genetic underpinning. Integrating these complementary viewpoints in a single framework improves diagnosis and provides biological insights about the disorders. However, imaging-genetic data lies in a very high-dimensional space with complex interaction and unknown causal factors. Our research aims to integrate multimodal imaging-genetics data to predict neuropsychiatric disorders while identifying biomarkers to provide biological insights.

We propose a novel generative-discriminative framework that integrates imaging and genetics data for simultaneous biomarker identification and disease classification. The generative module extracts representation patterns from the data, while the discriminative module uses the representation vectors for diagnosis. Our experimental analyses show that the discriminative module implicitly guides our framework, leading to improved disease diagnosis and biomarker identification.

Although our model successfully integrates imaging and genetic data, it fails to capture the complex non-linear interaction between the data modalities. To alleviate this problem, we introduce an autoencoder framework coupled with a classifier. The autoencoder learns the subspace shared between the input data modalities, and the classifier ensures that the subspace contains discriminative information. Unlike

traditional encoder-decoder models, our encoder module jointly identifies predictive imaging and genetic biomarkers using Bayesian feature selection.

We extend our Bayesian approach for feature selection to provide a fine-grained interpretation of the genetic variants that causally affect a trait. However, correct identification of the causal variants is challenging due to the correlation structure shared across variants. Our model combines a hierarchical Bayesian model with a deep learning-based inference procedure. We show that this combination provides greater inferential power to handle noise and spurious interactions of the genomic region.

Finally, we focus on solving the problem of encoding whole genome genotype data in an imaging-genetics framework. Traditionally, imaging-genetics models integrate imaging data with a sub-selected set of genetic features to ensure model stability. Our approach, an extension of the autoencoder framework, adheres to the same imaging encoding and Bayesian feature selection strategies. However, it departs from conventional Artificial Neural Networks (ANN) and introduces biologically regularized graph convolution networks to encode the whole genome genotype data. Our approach uses gene ontology to build a hierarchical graph that consolidates the genetic risk through predefined biological processes. We show that this embedding strategy helps us to track the convergence of genetic risk across well-established biological processes while preventing overfitting. Lastly, in an exploratory analysis, we use this model to investigate the underlying biological processes associated with behavioral phenotypes of autism spectrum disorder and schizophrenia.

Thesis Committee

Primary Readers

Dr. Archana Venkataraman (Primary Advisor)
John C. Malone Assistant Professor
Department of Electrical and Computer Engineering
Johns Hopkins University Whiting School of Engineering

Dr. Michael Schatz
Bloomberg Distinguished Professor
Department of Computer Science
Johns Hopkins Whiting School of Engineering

Dissertation Committee Members

Dr. Brian Caffo
Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Dr. Amitabh Basu
Associate Professor
Department of Applied Mathematics and Statistics
Johns Hopkins Whiting School of Engineering

Dr. Vishal Patel
Associate Professor
Department of Electrical and Computer Engineering
Johns Hopkins University Whiting School of Engineering

Acknowledgements

Long journeys are enjoyable if you share them with a set of people, and I have been fortunate to share my Ph.D. journey with some incredible individuals.

Foremost, my deep appreciation goes to my advisor, Archana Venkataraman, for introducing me to the fascinating world of mathematical modeling to solve real-life problems in biology. Her vast knowledge and intuitions about modeling have been critical for the success my projects. I will be forever thankful to her for guiding me in defining my research directions and developing critical thinking. Her unwavering support, patience, and boundless enthusiasm have fueled my dedication to this pursuit.

I also want to thank my thesis committee members, Michael Schatz, Amitabh Basu, Vishal Patel, and Brian Caffo. Each of them, with their expertise spanning optimization, deep learning, fMRI, and genetics, brought unique insights at critical junctures. Brian's online course on fMRI was a stepping stone in functional neuroimaging, which was followed by Amitabh's introduction to optimization. Vishal introduced me to the impressive power of deep learning models. Finally, Mike's insights into genetic data analysis were instrumental in the success of the projects.

This thesis stands on the intellectual foundations laid by Anand Mattay, Danny Chen, Giulio Pergola, Daniel Weinberger, Kevin Pelphrey, and Jack Van Horn. Their knowledge of the data and invaluable feedback about the clinical interpretations of the results were the driving force for developing novel tools to parse complex disorders.

I also want to thank my undergraduate thesis advisor, Ananda Shankar Choudhury,

and my academic internship advisor Nilanjan Ray. The research that I have done in their labs lays the foundation of my journey. They were the first to introduce me to the world of computational modeling and neuroimaging.

I am indebted to my lab members Niharika (I always called her Shimona), Naresh, Jeff, Ravi, Deeksha, and Arun for being an irreplaceable support system. Thank you for the endless banter, playing tennis, research discussions, and proofreading. My grad life is more enjoyable because of you, and I will always cherish our friendships.

I made some lifelong friends in Baltimore. My time in Baltimore wouldn't be as fun without the friendships of Debangan, Soumyodip, Prosenjit, Neha, Anindya, Sayantan, Ayan, Subhra, Pallabi, and Debojyoti. I am grateful for all the time we spend together traveling, playing, hanging out and having spirited conversations. They have definitely made Baltimore a home away from home.

A special shout-out to Arunima, for being my rock and confidante. You are the greatest gift I got during my Ph.D. Thank you for keeping up with my idiosyncrasies and motivating me to stay focused on this journey.

I wouldn't be where I am today without the endless support from my family. I will be forever grateful to my parents, Aditi and Rash Behari, for seeding the thought of pursuing scientific research since my childhood. My mother, a professor herself, nurtured my curious mind from childhood. I learned the value of perseverance and dedication from her. My father is my greatest source of confidence. He always taught me to trust my instincts and stay focused on my path. My brother, Soham, has always been my problem solver. From helping with projects in high school to guiding me in grad school, he always gave me his undying support. Last and most importantly, my family's curiosity and encouragement have always driven me to pursue excellence in my personal and professional life. I will be forever grateful to them for their confidence in me.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Multifaceted Nature of Neuropsychiatric Disorders

Neuropsychiatric disorders such as autism and schizophrenia have two complementary viewpoints. On the one hand, they are associated with behavioral and cognitive dysfunctions [1, 2], which results in altered brain activity [3, 4]. On the other hand, they are highly heritable, which suggests a strong genetic underpinning [5].

Modern-day neuroimaging techniques like Magnetic Resonance Imaging (MRI), and functional Magnetic Resonance Imaging (fMRI) provide a non-invasive way to explore the structural [6] and functional organization [7] of the brain. For instance, MRI gives us measures of morphological changes in brain regions, and fMRI quantifies neuronal activity in response to experimental stimulus. In the past decade, multiple studies have investigated the association of these quantifiable imaging phenotypes with clinical outcomes using both univariate [8–10] and multivariate [11–14] models. These models have successfully pinpointed multiple brain phenotypes but often ignore the heritable component of the disorder, thus lacking in explaining the underlying biology.

On the other front, with the advent of low-cost sequencing technologies [15], multiple consortia (UK Biobank, PGC) were able to generate large volumes of whole

genome genotype data. Genome Wide Association Studies (GWAS) [16, 17] have developed multiple univariate testing tools to identify the level of association of each genetic variant to the outcome. These models can successfully identify the targets [18, 19] on the genome and explain the underlying etiology of the disease using gene expression [20], gene-enrichment [21] or pathway-based analyses [22]. However, multiple gene-gene [23] or gene-environment [24] interactions make it difficult to decipher the genetic contributions without integrating other biological modalities.

Most studies decouple the problems of disentangling the neural mechanisms[25] and pinpointing genetic variations [17], which ultimately provides an incomplete picture of the underlying disorder [26, 27]. Recently, the works of [28, 29] have used brain imaging phenotypes to elucidate the mechanisms through which genetics confer risk. [30] used polygenic risk scores to predict hippocampal activity in schizophrenia. They found that decreased hippocampal-parahippocampal activity is strongly associated with a high polygenic risk score. Similar strategies [31] have seen a significant association between polygenic risk and white matter connectivity in frontal-parietal regions for autism spectrum disorder. On a high level, the imaging modality acts as an intermediate phenotype that guides the imaging genetics models to strategically consolidate the genetic risk associated with the clinical traits. These studies have provided evidence that clinical traits like cognitive dysfunction, and impaired social and communication skills are associated with genetic risks, but they fail to pinpoint the target genetic variants, genes, or molecular and biological functions. This drawback motivates the need to develop robust, and interpretable frameworks for integrating whole brain whole genome data for biomarker discovery and predictive analytics.

1.2 Complex Genetic Heritability of Neuropsychiatric Disorders

Multiple family and twin studies [32, 33] have found that neuropsychiatric disorders like schizophrenia and autism are highly hereditary ($60\% - 80\%$) [32, 33]. However, the heritability estimated by common genetic variants from multiple genome-wide association studies is much lower [16, 17]. The heritability gap [34] is often associated with the gene-gene interaction [23], non-linear genetic effects [35], and gene-environment interactions [24]. In addition, schizophrenia and autism are heterogeneous disorders, often characterized by a broad spectrum of phenotypes [36]. Therefore, multiple genes could potentially affect various biological pathways with downstream effects on the behavioral level. In fact, the recent GWAS on schizophrenia [37] and autism [16] have shown evidence that both these disorders are polygenic [38], which means different genes act in concordance, leading to the disorder.

Parsing the genetic risk associated with these disorders is a two-step process. The first step involves identifying the target [39, 40] risk variants from GWAS results, and the second step involves connecting the risk variants with the downstream biological function using gene enrichment or expression-based analysis [18, 41]. However, identifying the target variant is challenging due to the correlation structure between variants, which arises due to low genomic recombination of nearby DNA regions [42]. Consequently, the strong correlations inflate the effect size of a non-causal genetic variant, thus leading to false positive [43] identifications. The second challenge in parsing the genetic risk is that most of the strongly associated genetic variants are located in the non-coding [44] region of the DNA. The variants affect the regulatory factors [45] of a gene, which results in alterations in the gene expression levels or interactions with other genes. These complexities associated with genetic data necessitate the need to develop new frameworks to parse the genetic risks and identify target variants and

underlying biological processes.

1.3 Challenges With Imaging Genetics Studies

Modeling Interaction Between Imaging and Genetics: The first set of challenges in imaging and genetics study is to model the relationship between imaging and genetics data. Initial approaches [28–30] used a polygenic risk score model to explain the observed phenotype using the genotype data in a linear framework. However, the polygenic risk score is a cumulative score that does not explain the underlying biology. For example, two subjects with similar risk scores can have different risk alleles affecting disorder-relevant biological pathways differently [46]. This issue is addressed in pathway-driven polygenic scoring [28, 29], but these models require *a priori* knowledge of the implicated biological pathways, which is often unavailable.

Despite the shortcomings, the genetic risk score-based studies highlight that multiple genes act in synchrony [38], which ultimately leads to altered neural functioning of different brain regions. This observation led to a group of bi-multivariate imaging genetics studies [11, 47, 48] aiming to identify a cluster of functionally related genetic variants statistically correlated with a network of brain regions. The coherent set of imaging and genetic biomarkers shed light on the complex interactions between the two modalities. However, these approaches are not adaptable to increasing data modalities and often ignore the disease status, resulting in a lack of clinically relevant biomarkers. We address this shortcoming in a generative discriminative framework. The generative model combines multiple data modalities, while the discriminative module guides the framework to find discriminative biomarkers. In addition, we incorporate structural and biological priors as regularization, leading to robust biomarkers and improved risk prediction.

Modeling Multimodalities and Non-linear Interactions: The second challenge in imaging genetics is modeling multiple data modalities in a single framework. As explained in previous sections, neuropsychiatric disorders are characterized by heterogeneous phenotypes resulting from a combined genetic effect along multiple biological processes. As a result, the field is moving towards multimodal imaging acquisitions [49] to capture different snapshots of the brain, all of which may have links to the genotype. However, with the growing emphasis on big datasets comes the challenge of missing data modalities. Traditionally, missing data has been managed [50] by removing subjects from the analysis, which does not use all the information.

The third challenge in imaging genetics is the simplistic assumption of linearity between the two data modalities. However, biology is complex, and the path through which genetics confer risk is often non-linear [51, 52]. This motivates the need to develop adaptable frameworks that extract non-linear feature representations from multiple data modalities while handling missing data.

We solve these problems by introducing robust and adaptable autoencoder frameworks that can successfully combine the non-linear projections of imaging and genetics data. Our autoencoder frameworks are complemented with classifiers and a Bayesian feature selection module. The Bayesian module identifies biomarkers, while the classifier ensures that the biomarkers are clinically relevant.

Parsing the Genetic Information In Data-Driven Models: The fourth set of challenges involves modeling millions of highly correlated genetic variants in a data-driven fashion. Traditional and deep learning-based imaging genetics models use a drastically reduced set of genetic features, often identified by thresholding GWAS p-values. However, GWAS is a univariate approach that does not consider the correlation structure across variants. So, a GWAS-guided identification of genetic variants often does not contain true signals [43]. The works of [39, 53–55] provide a

solution to find target variants using Bayesian and frequentist approaches of variable selections. However, these models also fail to identify robust targets for polygenic traits due to spurious effects from non-target variants.

The GWAS sub-selection step restricts the use of complete genetic information. In terms of scale, a raw dataset of $\sim 300K$ genetic variants is reduced to 1K SNPs. In contrast, neuropsychiatric disorders are polygenetic, meaning that they are influenced by numerous genetic variants interacting across many biological pathways. Pruning out the genetic information effectively removes the information of the downstream genetic effects. Within the genetics realm, there is a vast literature that associates genetic variants and genes to different biological pathways [56]. The works of [57, 58] have used this information to design a sparse artificial neural network that aggregates genetic risk according to these pathways in order to predict a phenotypic variable. While an important first step, their ANN contains just a single hidden layer, which does not account for the hierarchical and interconnected nature of the biological processes.

The above challenges highlight the need for data-driven models that can identify target variants in complex genetic architectures and provide a strategically regularized framework that encodes millions of genetic variants.

1.4 Contributions of This Thesis

This thesis aims to provide data-driven solutions to handle the multifaceted nature of psychiatric disorders while identifying target variants and providing insights about the implicated biological processes. However, integration and parsing of imaging genetics data has many challenges: high data dimensionality, complex interactions, and unknown causal factors. We approach these problems in three steps. First, we build biologically regularized models to extract discriminative patterns from the

data. Second, we combine domain knowledge about the modalities with data-driven learning models. Third, we evaluate these models on multiple studies and quantify their reproducibility and performance. This thesis makes four major contributions in each of these stages to push the boundary of machine learning research in imaging genetics.

- A generative discriminative approach that uses matrix decomposition to model complex interaction between imaging and genetics. The model is regularized and guided by clinical labels, resulting in discriminative biomarkers.
- An autoencoder framework to model the non-linear interaction between imaging and genetics. The autoencoder framework is coupled with a novel Bayesian feature selection layer to identify potential neuroimaging and genetic targets.
- A Deep Bayes variation approach to parse the landscape of genetics and find potential target variants from genome wide association studies.
- A graph-based deep neural network that can encode *millions* of genetics variants using prior biological knowledge of SNP-gene and gene-pathway interactions. We use this model to integrate multimodal imaging and genetics data and extend it to create interpretable and non-linear genetic risk scores.

1.5 Thesis Outline

Chapter 2 provides the background about the data modalities and the existing machine learning and deep learning strategies to analyze the data. We will also introduce the deep learning concepts and architectural designs used in this work. Finally, we will provide the details about our experimental datasets and the preprocessing strategies.

In Chapter 3, we introduce our generative discriminative framework for integrating imaging and genetic data while performing diagnosis. This work builds upon our

SPIE [59], MICCAI [60] and Neuroimage [61] papers where we explore the interactions between brain networks and genetic loci for schizophrenia risk prediction.

Chapter 4 extends the model presented in Chapter 3 to incorporate non-linear interactions between imaging and genetics. This work was originally presented in conference paper form in [62]. This work introduces an autoencoder framework coupled with a classifier module. Additionally, we introduce a novel joint feature selection using a Bayesian approach that provides a posterior feature selection probability map.

Chapter 5 extends our Bayesian approach for feature selection to solve the problem of finemapping. In this work [63], we introduce a deep Bayes variational model to approximate the posterior distribution of the causal variants given the GWAS summary statistics.

Finally, in Chapter 6, we introduce a graph-based convolution model to encode whole genome genotype data. In this work, we introduce a biological knowledge-driven regularization strategy that successfully encodes *millions* of genetics variants in a robust end-to-end framework. We use this encoding strategy to integrate imaging and genetics data, which is published in ICLR [64]. Finally, in our ongoing work, we extend the genetic encoding strategy to create interpretable and non-linear genetic risk scores.

Chapter 2

Background

2.1 Viewing Neuropsychiatric Disorders through altered brain activity

Functional Magnetic Resonance Imaging (fMRI) allows us to assess brain activity in response to a given stimulus or experimental paradigm [3, 4, 65]. It has become a powerful tool for studying brain abnormalities in patients with neuropsychiatric disorders. fMRI measures changes in blood oxygenation within the brain. The brain regions involved in performing a task require more oxygen [66], and to meet this increased demand, blood flow increases to the active areas. Hence, the neural activity induced by an experimental stimulus is highly correlated with changes in blood flow and oxygenation [65], and fMRI helps us to detect those changes. Hemoglobin is diamagnetic when oxygenated but paramagnetic when deoxygenated. This difference in magnetic properties is captured by a $T2^*$-weighted [67] protocol that measures changes in oxygenation over the course of the scan. The primary advantage of fMRI is its ability to provide superior temporal and spatial measures of brain activity in response to an experimental stimulus.

2.1.1 Functional Localization and Brain Networks Identification

Generalized Linear Model (GLM): Statistical analyses of the fMRI are geared towards identifying brain regions that activate or deactivate in response to stimuli. The most popular approach to studying the individual brain is the Generalized Linear Model (GLM) [8], which represents the fMRI data as a linear combination of the stimuli onsets and responses across time. Mathematically, the fMRI time-series signal from a spatial location i in the brain is presented as:

$$\boldsymbol{f}_i = \boldsymbol{A}\boldsymbol{\beta}_i + \epsilon_i \tag{2.1}$$

where $\boldsymbol{f}_i \in \boldsymbol{R}^{T\times 1}$ is the time series signal of T time points from location i, $\boldsymbol{A} \in \boldsymbol{R}^{T\times D}$ is a design matrix whose columns capture the onset of the experimental stimuli, ϵ_i is additive Gaussian noise, and $\boldsymbol{\beta}_i$ is the response of the brain from location i. Mathematically, we can estimate $\boldsymbol{\beta}_i$ by estimating the least-squares solution. [10]:

$$\hat{\boldsymbol{\beta}}_i = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{f}_i \tag{2.2}$$

As shown in Fig. 2-1 the estimated coefficients $\boldsymbol{\beta}$ indicate the response of a region i, thus informing us of the role played by a region i within the brain.

The GLM model evaluates the correlation between the fMRI timeseries of each voxel with the experimental stimuli but fails to capture higher-order relationships [11–13]. In addition, GLM models fail to provide the underlying brain networks [68] that often respond synchronously to a task stimulus. This problem is addressed in Independent Component Analyses, which can find task-related higher-order brain networks from fMRI time-series signals.

Independent Component Analysis (ICA): ICA models assume the fMRI signal is a linear mixture of spatially independent source signals. The goal of ICA is to

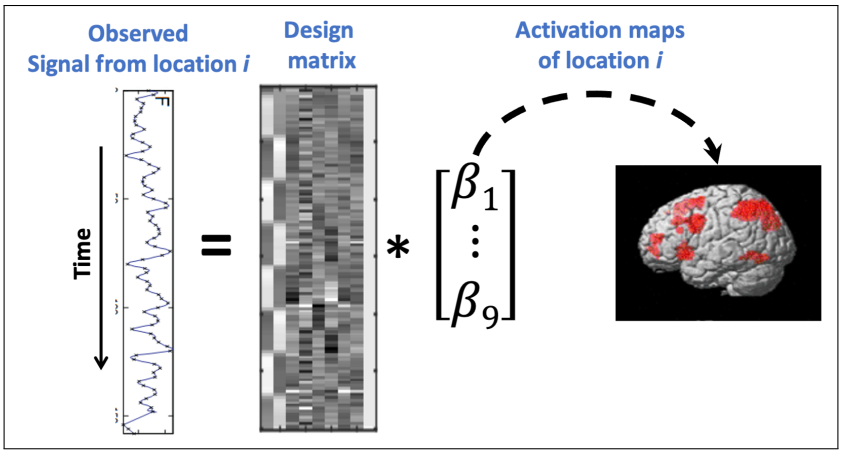


Figure 2-1. A Generalized Linear Model (GLM). The observed signal is the fMRI time-series signal location i. The design matrix contains the input experimental stimulus. The estimated coefficients β_j capture the brain activation in response to the stimuli.

separate the independent source signals using higher-order statistics. Prior works have shown that the independent components [11–13, 69] group all brain regions with synchronized signals, thus identifying temporally coherent functional networks (FNs). Mathematically, the ICA framework can be written as:

$$\boldsymbol{F} = \boldsymbol{M}\boldsymbol{S} \tag{2.3}$$

where we concatenate the fMRI timeseries signals for L locations into a matrix $\boldsymbol{F} = [\boldsymbol{f}_1, \ldots, \boldsymbol{f}_L]$. $\boldsymbol{S} \in \boldsymbol{R}^{S \times L}$ are S independent source signals spanning across L regions and $\boldsymbol{M}$ is the mixing parameter of the source signals. Eq. (2.3) is similar to Eq. (2.2), however unlike GLM here both $\boldsymbol{M}$ and $\boldsymbol{S}$ are estimated by optimizing [70, 71] a higher order statistics in a data-driven fashion. Each row of $\boldsymbol{S}$ could give us a network of brain regions acting synchronously in response to the experimental stimulus. The independent components have proven to identify higher-order brain states [72].

The main drawback of ICA is that it does not naturally generalize to multiple subjects. Traditionally, ICA is done on a single subject and a post-hoc analysis [13] is required to map the independent components across samples. This complexity is in

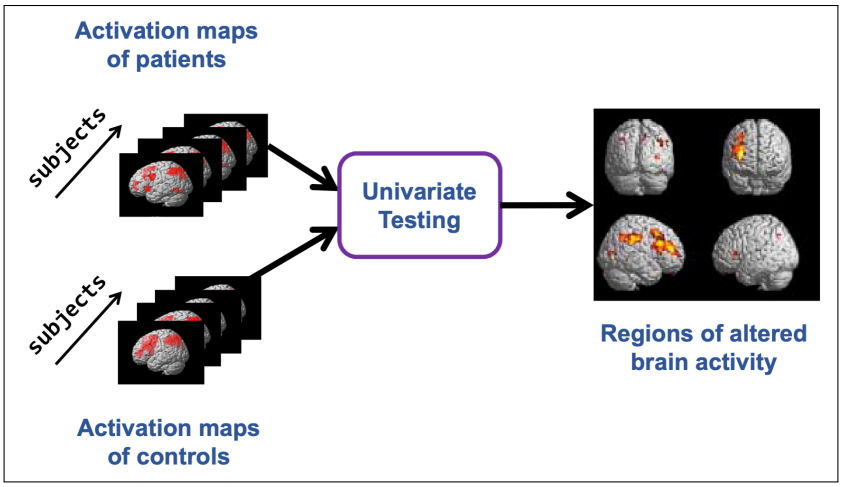


Figure 2-2. The spatial brain activation maps obtained from GLM or ICA framework is passed through univariate tests like t-test to find regions of activity affected by the disorder.

contrast to GLM models where the design matrix is fixed across samples, which allows for comparison of the activation maps (β_i) across samples. This problem is addressed in group-ICA [69] where subjects are concatenated in the temporal directions, and an ICA is performed to find common spatial activation across groups.

2.1.2 Group Analysis

The group-level analysis in fMRI can be divided into two main categories. The first category extracts spatial patterns of brain activation from each subject using the GLM or group-ICA framework and passes it through a mass univariate testing [73, 74] framework for finding group-level differences. Figure 2-2 shows the general strategy for performing mass hypothesis testing across samples. Although these methods are widely used, they fail to capture multivariate interactions [75] across the brain. In addition, the results of the univariate techniques also vary widely across different subsets of data, resulting in low reproducibility.

The second category uses machine learning models to optimize for group separability using the brain activation maps. For example, the Support Vector Machine (SVM) [76,

77], constructs a high-dimensional set of hyper-planes that maximally separates the data into multiple categories. Different ensemble-based methods like random forest (RF) [78], and neural networks [79, 80] have shown their ability to learn multiscale features while enhancing classification accuracy. While these methods can successfully capture multivariate interactions with high generalization accuracy, they treat brain activations as an arbitrary collection of features. As a result, the patterns implicated by these methods can be difficult to interpret.

2.2 Genetic Implication of Neuropsychiatric Disorders

The genetic susceptibility of neuropsychiatric disorders like schizophrenia and autism is complex, resulting from the combined effects of multiple alleles [81, 82]. The allelic changes in the DNA can be measured as Single Nucleotide Polymorphisms (SNP), or Structural Variants (SV). SNPs are germline substitutions of single nucleotides in the genome that are present in a sufficiently large fraction of the population. In comparison, SVs capture changes in wider chromosomal regions and may potentially be responsible for changes in gene structure, resulting in significant phenotypic effects [83]. In this thesis, we will use SNPs as our genetic data modality to investigate their effect and interactions with brain activation in neuropsychiatric disorders.

2.2.1 Single Nucleotide Polymorphisms (SNP)

SNPs are captured as alterations of nucleotides in the DNA [84]. The DNA consists of two polynucleotide chains that coil around each other to form a double helix. At every location of the DNA, we have a pair of the following four nucleotides: Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). This pair of nucleotides can be grouped as homozygous major (AA), heterozygous (AB), and homozygous minor (BB) alleles, where A represents a major allele and B represents a minor allele. The

homozygous major allele consists of two major alleles; the heterozygous allele consists of a major and a minor allele; the homozygous minor allele consists of two minor alleles. Mathematically, SNP data at a locus is coded as $\{0, 1, 2\}$ for AA, AB, and BB, respectively. The allelic configuration of the DNA captured by the SNP data provides us with the complex genetic architectures of a population.

2.2.2 Genome Wide Association Studies (GWAS)

GWAS [85] investigates the effect of each SNP on the outcome. Let's assume $\boldsymbol{y} \in \boldsymbol{R}^{N\times 1}$ is an observed phenotype across N subjects. The genotyped SNP data across M variants are encoded as $\{0, 1, 2\}$ and represented by a matrix $\boldsymbol{G} \in \boldsymbol{R}^{N\times M}$. They are further normalized so that each column has a mean 0 and variance 1, i.e., $\frac{1}{N}\sum_i \boldsymbol{G}_{ij} = 0$ and $\frac{1}{N}\sum_i \boldsymbol{G}_{ij}^2 = 1$. GWAS fits a linear regression between individual SNP and the observed phenotype using the following framework:

$$\boldsymbol{y} = \boldsymbol{g}_i\beta_i + \epsilon \quad \epsilon \sim N\left(0, \frac{1}{\tau}\mathbb{I}\right) \tag{2.4}$$

where $\boldsymbol{g}_i \in \boldsymbol{R}^{N\times 1}$ is the i-th SNP information across N subjects, β_i is the SNP effect, $\boldsymbol{y}$ is the observed phenotype data and ϵ is additive Gaussian noise with mean 0 and variance $\frac{1}{\tau}$. The effect size β_i can be estimated [86] as the following:

$$\hat{\beta}_i = \frac{1}{N}\boldsymbol{g}_i^T\boldsymbol{y} \tag{2.5}$$

In GWAS analyses, the effect sizes are normalized by their standard error to generate a z-score and a p-value. The z-score is generated as:

$$z_i = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \tag{2.6}$$

$$= \sqrt{\frac{\tau}{N}}\boldsymbol{g}_i^T\boldsymbol{y} \tag{2.7}$$

where $SE(\beta_i)$ is the standard error of the effect size $\hat{\beta}_i$. The z-scores across all the SNPs can be compactly represented as $\boldsymbol{z} = \sqrt{\frac{\tau}{N}}\boldsymbol{G}^T\boldsymbol{y}$

2.2.3 Identifying Target SNPs Based On GWAS Results

2.2.3.1 Challenges of GWAS

Genome-wide association Studies (GWAS) provide a natural way to quantify the contribution of each genetic variant to the observed phenotype [85]. However, the univariate nature of GWAS does not consider the correlation structure shared between the genetic variants, which arises due to low genomic recombination of nearby DNA regions [87]. Consequently, strong correlations can inflate the effect size of a non-causal genetic variant, thus leading to false positive identifications [43]. Mathematically, we can show this by assuming that the observed trait can be represented by a linear sum of causal genotypes. Following the same mathematical notation defined in previous sections, let's assume $\boldsymbol{y} \in \mathbb{R}^{N\times 1}$ denote a vector of (scalar) quantitative traits, and $\boldsymbol{G}$ denotes the genotype data. The quantitative trait is generated as follows:

$$\boldsymbol{y} = \boldsymbol{G}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \boldsymbol{\epsilon} \sim N\left(0, \frac{1}{\tau}\mathbb{I}_n\right), \tag{2.8}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{M\times 1}$ is the true effect size with 0 in non-causal locations and non-zero in causal locations, $\boldsymbol{\epsilon} \in \mathbb{R}^{N\times 1}$ is additive white Gaussian noise with variance $\frac{1}{\tau}$, and $\mathbb{I}_N$ is the $N \times N$ identity matrix. Under this generative assumption, the estimated effect sizes given in Eq 2.6 can be written as:

$$\begin{aligned}
\hat{\beta}_j^{\text{non-causal}} &= \frac{1}{N}\boldsymbol{g}_j^T\boldsymbol{y} \\
&= \frac{1}{N}\boldsymbol{g}_j^T\boldsymbol{G}\boldsymbol{\beta} \quad \text{we replace } \boldsymbol{y} \text{ according to Eq. (2.8)} \\
&= \frac{1}{N}\sum_{i\in \text{ CausalSet}} r(i,j)\boldsymbol{\beta}_i
\end{aligned} \tag{2.9}$$

where $\hat{\beta}_j^{\text{non-causal}}$ is the effect size of the j-th non-causal variant, $\boldsymbol{\beta}$ is the true effect sizes of the variants with 0 in non-causal locations, $r(i,j)$ is the correlation between the j-th non-causal variant and i-th causal variant, and $\boldsymbol{\beta}_i$ is the true effect of the i-th causal variant. Eq. (2.9) captures the inflation of the effect sizes of non-causal variants due to the correlation with the causal variants.

2.2.3.2 Finemapping Approaches

Fine-mapping [40, 88] addresses this problem by analyzing the correlation structure of the data to identify small subsets of causal genetic variants [88, 89]. These subsets, known as credible sets, capture the uncertainty of finding the true causal variant within a highly correlated region [90]. Unlike p-values, the corresponding posterior inclusion probabilities (PIPs) computed during fine-mapping can be compared across studies of different sample sizes.

Heuristics Based Finemapping: Traditional fine-mapping methods use a penalized regression model to predict the output phenotype based on the collection of genetic variants [91, 92]. Popular regularizations like LASSO [93] and Elastic Net [92] simultaneously perform effect size estimation while slowly shrinking the smaller effect sizes to zero. The drawback of penalized regression models is that they optimize phenotypic prediction and, due to the correlation structure, do not always identify the true causal variants.

Bayesian Finemapping: The second category relies on Bayesian modeling. Here, the phenotype is modeled as a linear combination of the genetic variants, with sparsity incorporated into the prior distribution for the model weights. Approximate inference techniques, such as Markov Chain Monte Carlo (MCMC) [51] and variational methods [94] have been used to infer the effect sizes, PIPs, and credible sets. While these approaches represent valuable contributions to the field, they require raw genotype and phenotype information, which raises privacy and regulatory concerns, particularly in the cases of publicly shared datasets. MCMC sampling also requires a burn-in period, which adds a substantial (100X) runtime overhead.

Recently, fine-mapping approaches [39, 95, 96] have moved towards using summary statistics, which can be easily shared across sites. These approaches use the GWAS

results to estimate the posterior of the causal configurations. This posterior is estimated using the Bayes rule:

$$P\left(\boldsymbol{c}|\boldsymbol{z},\Sigma_X\right)=\frac{P\left(\boldsymbol{z}|\Sigma_X,\boldsymbol{c}\right)P\left(\boldsymbol{c}\right)}{\sum_l P\left(\boldsymbol{z}|\Sigma_X,\boldsymbol{c}^l\right)P\left(\boldsymbol{c}^l\right)} \tag{2.10}$$

where $\boldsymbol{c}$ is a binary vector denoting the location of causal variants, $\Sigma_X=\frac{1}{N}\boldsymbol{G}^T\boldsymbol{G}$ is the SNP-SNP correlation matrix, $\boldsymbol{z}$ is the z-score estimated from GWAS, and $\sum_l(\cdot)$ is the sum over all possible causal configurations. In the above expression, the likelihood term $P\left(\boldsymbol{z}|\Sigma_X,\boldsymbol{c}\right)$ follows a normal distribution and can be obtained by combining Eq. (2.6) and Eq. (2.8).

$$\boldsymbol{z}=\sqrt{\frac{\tau}{n}}\boldsymbol{G}^T\left(\boldsymbol{G}\boldsymbol{\beta}+\boldsymbol{\epsilon}\right)\ \text{ where } \epsilon\sim N\left(0,\frac{1}{\tau}\mathbb{I}_n\right) \tag{2.11}$$

$$\implies P\left(\boldsymbol{z}|\boldsymbol{\beta}\right)=N\left(\boldsymbol{z};\sqrt{(N\tau)}\boldsymbol{\beta}\boldsymbol{\Sigma}_X,\boldsymbol{\Sigma}_X\right) \tag{2.12}$$

However, estimating Eq. (2.10) is difficult because the denominator is intractable with increasing number of causal variants. The works of [39, 86, 95] solved it by defining a prior over $\boldsymbol{\beta}\sim N\left(0,\frac{1}{\tau}\sigma^2 diag(\boldsymbol{c})\right)$ and using a stochastic or exhaustive search to identify the posterior probabilities of the causal configurations. However, exhaustive search-based methods are restricted by the number of assumed causal variants, leading to an exponential increase in the dimensionality of the approximate posterior distribution. Stochastic search approaches [95] are computationally less expensive, but, by construction, they cannot handle nontrivial effects from spurious non-causal variants. The most recent contribution to fine-mapping is SuSiE [53, 96], which uses a different prior and estimates $\boldsymbol{\beta}$ as the sum of "single effects". These "single effect" vectors contain one non-zero element representing a causal variant and are estimated using a Bayesian step-wise selection approach. SuSiE provides a simple framework to robustly estimate PIPs and credible sets; however, there is limited evidence for its performance, given the presence of spurious genetic effects. Such scenarios can appear due to polygenicity of the trait, trans-interactions of variants, or varying correlation structure of the genomic region.

Our approach in Chapter 5 deviates from the strong assumption that non-casual variants have zero effects. As a result, our approach provides a robust framework that can account for infinitesimal effects from non-causal variants in polygenic traits and spurious effects from non-causal variants due to interaction artifacts.

2.2.4 Polygenic Risk Scores (PRS)

The genotyped SNP data can also predict a phenotype [97]. Polygenic Risk Score (PRS) [98] measures the genetic liability to a disease or trait based on the genotype data of an individual. PRS are estimated as a weighted sum of the genotype data, where the weights are obtained from a GWAS. On a high level, PRS can be assumed to provide an individual-level proxy of genetic liability to a trait. Mathematically, PRS from an individual subject is estimated as the following:

$$s_n = \frac{\sum_j \boldsymbol{G}[n, j] * \beta_j}{2 * M} \tag{2.13}$$

where s_n is the PRS for the n-th subject, $\boldsymbol{G} \in \boldsymbol{R}^{N \times M}$ is the genotyped SNP data, β_j is the estimated effect size obtained from an independent GWAS, and M is the number of SNPs used to calculate the PRS. In addition to phenotype prediction, PRSs are suitable for various applications, such as identifying shared etiology among traits [99], gene-by-environment [100], and gene-by-gene interactions. However, the polygenic risk score collapses all the genetic information to a scalar value, thus ignoring the complex interactions between variants.

Recent approaches parse the genetic risk based on their involvement in different pathways [28, 101, 102]. For example, in schizophrenia, SNPs related to the genes linked with the biological process, glutamatergic signaling, can be used in creating pathway-specific PRSs. In Chapter 6, we take a similar strategy to encode the genotype data. However, instead of handcrafting the pathway-specific scores, we use graph-based models to prioritize pathways and create scores in a data-driven fashion.

2.3 Joint Analyses of Imaging Genetics

In previous sections, we have introduced the prior works in imaging and genetics that try to investigate each modality separately. However, the complex traits inherited in neuropsychiatric disorders encompass interplay between multiple genes [81, 82] that contribute to biological processes like neurogenesis, transcriptional regulation, dopaminergic and glutamatergic signaling. Therefore, disentangling the neural mechanisms [25] and pinpointing genetic variations [17] separately provides an incomplete picture of the underlying disorder [26, 27].

Imaging-genetics is an emerging field that tries to merge these complementary viewpoints [103]. The imaging features are often derived from structural and functional MRI (s/fMRI), and the genetic variants are typically captured by Single Nucleotide Polymorphisms (SNPs). Data-driven imaging-genetics methods can be grouped into three main categories.

2.3.1 Regression Based Approaches:

Initial approaches [104–106] in imaging-genetics evaluated the association between each SNP and brain ROI pair in a linear framework. Such pairwise association analysis results in a large number of hypothesis tests, leading to problems associated with multiple testings [107]. Furthermore, the resulting p-values are not independent because of spatial correlation in the imaging data. This problem is addressed in the work of [47, 108] where multiple SNPs are used to predict the multiple brain regions in a single multivariate linear framework. The general strategy of these models is to minimize the following loss functions:

$$\min_{\boldsymbol{W}} \sum_{i} ||\boldsymbol{i}_i - \boldsymbol{W}^T \boldsymbol{g}_i|| + \lambda \mathcal{R}(\boldsymbol{W}) \tag{2.14}$$

where $\boldsymbol{i}_i \in \boldsymbol{R}^{R\times 1}$ is the observed phenotype from R brain regions, $\boldsymbol{g}_i \in \boldsymbol{R}^{M\times 1}$ is the genotype data from i-th subjects, $\boldsymbol{W} \in \boldsymbol{R}^{M\times R}$ is the multivariate regression coefficients,

and $\mathcal{R}(\cdot)$ is a regularization penalty over the regression coefficients. Additional prior are imposed over $\boldsymbol{W}$ in the form of regularization that control for interaction between genetic variants or brain regions. For example, [47] use a group sparsity penalty that ensures that a sparse set of SNPs localized to genes influence all the brain ROIs similarly. [109] took a different approach and enforces low rank approximation over $\boldsymbol{W}$. Both these approaches are geared towards finding a common pattern that can capture the interplay between a set of genetic variants and brain ROIs.

While the regression-based models are simple to implement and easy to interpret, penalized regression models do not naturally incorporate the effect of a disease. As a result, the identified biomarkers may not be relevant to the underlying disorder. In Chapter 3, we demonstrate how the penalized regression-based models could be extended to model imaging-genetics data while maintaining clinical interpretability.

2.3.2 Bi-multivariate Approaches:

One leading hypothesis in imaging genetics is that both imaging and genetics data share a common latent space. Under this hypothesis, multiple bi-multivariate models [11, 12, 14, 110] try to project the imaging and genetics data to a lower dimensional space and align them by maximizing the correlations. Canonical Correlation Analysis (CCA) and parallel Independent Component Analysis (p-ICA) are the two leading approaches in this space.

Canonical Correlation Analysis (CCA): Canonical Correlation Analysis (CCA) [110–112] finds bivariate associations between the imaging and genetics data. These canonical coefficients are obtained by maximizing the following function:

$$\{\boldsymbol{u}_i^*, \boldsymbol{v}_i^*\} = \max_{\boldsymbol{u}_i, \boldsymbol{v}_i} \; corr(\boldsymbol{I}^T \boldsymbol{u}_i, \boldsymbol{G}^T \boldsymbol{v}_i)$$

where $\{\boldsymbol{u}_i, \boldsymbol{v}_i\}$ are the orthonormal basis vectors, and across N individuals we have concatenated the patient activations maps as $\boldsymbol{I} = [\boldsymbol{i}_1, \ldots, \boldsymbol{i}_N]$, the genetic variants as

$\boldsymbol{G} = [\boldsymbol{g}_1, \ldots, \boldsymbol{g}_N]$. These basis vectors form a low dimensional space where the two data modalities are maximally correlated. In addition, the basis vectors provide us with a cluster of functionally related SNPs that are statistically correlated with brain ROIs. However, the high dimensionality of imaging and genetics data often leads to model overfitting and a lack of robust biomarkers. Sparse CCA (SCCA) [110–112] addresses this problem by incorporating the Least Absolute Shrinkage and Selection Operator (LASSO) constraints on the basis vectors $\{\boldsymbol{u}, \boldsymbol{v}\}$. These sparsity-inducing priors lead to robust and interpretable biomarkers. The s-CCA models are further extended [113, 114] to model diagnosis labels leading to meaningful and clinically relevant biomarkers.

The outputs of CCA-based models can also be used for phenotype predictions [114]. s-CCA identifies a sparse set of imaging and genetic features that can be combined and passed through a classifier for prediction. Additionally, the lower dimensional projections $\left[\boldsymbol{i}_m^T \boldsymbol{U}, \boldsymbol{g}_m^T \boldsymbol{V}\right]$ can also be used a feature vectors in a classifier model.

The main drawback of the CCA-based model is that it is hard to add more data modalities. However, the field is moving towards multimodal imaging acquisitions to capture different snapshots of the brain, all of which may have a link to the genotype. To address this issue, in Chapter 3 and Chapter 4, we use a dictionary learning framework and autoencoder to model multiple data modalities.

Parallel Independent Component Analysis (pICA): Parallel ICA (pICA) is an alternative method that uses statistical independence to identify a set of basis vectors for each modality [115]. Compared to CCA, pICA can extract higher-order dependencies beyond linear correlation [13, 115, 116]. Parallel ICA (p-ICA) decomposes the imaging and genetics data into independent but interrelated networks. This is done by jointly maximizing multiple 'cost functions,' one of which specifies the independence among networks in each of the data sets and another term that maximizes

the correlation among pairs of networks across data sets. Formally, let $\boldsymbol{I} \in \boldsymbol{R}^{R \times N}$ is the imaging data and $\boldsymbol{G} \in \boldsymbol{R}^{M \times N}$ is the genetic data collected from N subjects. Mathematically, the pICA framework can be written as:

$$\boldsymbol{I} = \boldsymbol{S}\boldsymbol{X} \quad \text{and} \quad \boldsymbol{G} = \boldsymbol{W}\boldsymbol{Z}$$

where $\boldsymbol{S}, \boldsymbol{W}$ are independent source matrices and the $\boldsymbol{X}, \boldsymbol{Z}$ are loading matrices whose cross-correlation is maximized. Parallel ICA extracts the correlated pair of components from the two modalities. The importance of each component can be further estimated by computing p-values of the association between the loading scores and the disease label.

Similar to CCA, pICA identifies a robust set of biomarkers that can be used in a classifier framework for disease prediction. The drawbacks of pICA are also similar to CCA. In pICA, the entire model and optimization procedure must be changed to add new modalities and handle missing data.

2.3.3 Deep Learning Methods

The third category for imaging-genetics uses deep learning to link the multiple viewpoints [79, 80, 117, 118]. Unlike traditional methods, deep learning can automatically learn complex representations from data [79, 80, 119]. These techniques have become the state of the art for analyzing fMRI data sets, resulting in performance improvements in diverse fMRI applications. Deep learning is less common in the genetics literature due to the high dimensionality and unstructured nature of the data. However, with the exponentially increasing volume of genomics data, deep learning has proven to be a useful tool for multiple genomic modeling applications [117].

Deep learning models allow us to capture non-linear interactions between imaging and genetics data. Usually, the goal of deep learning is to approximate an optimal function $\boldsymbol{y} = f(\boldsymbol{I}, \boldsymbol{G})$ with a succession of non-linear transformations. The optimal

function $f(\cdot)$ is solved in imaging-genetics using one of the following three approaches.

Artificial Neural Network (ANN): ANN is the simplest form of neural network, which consists of a sequence of fully connected perceptron layers. Each layer consists of a linear transformation followed by a non-linear activation. Formally, let $\boldsymbol{h}^{l-1} \in \boldsymbol{R}^{d_{l-1}}$ is a d_{l-1} dimensional input at layer l. The output of the layer is given by:

$$\boldsymbol{h}^{l} = \zeta\left(\boldsymbol{W}\boldsymbol{h}^{l}\right) \tag{2.15}$$

where $\boldsymbol{h}^{l} \in \boldsymbol{R}^{d_l \times 1}$ is the output of l-th layer, $\boldsymbol{W} \in \boldsymbol{R}^{d_l \times d_{l-1}}$ is the weight matrix, and $\zeta(\cdot)$ is a non-linear activation function. The non-linear activation function is usually modeled as a sigmoid, tanh, ReLU [120], or PReLU [121] function.

ANNs have been widely used to extract complex representation patterns [122] from imaging and genetics data. In the imaging domain, ANNs [123] have been used to model brain activation maps, cross-sectional cortical thickness, or brain volumes to predict the clinical phenotypes of Autism, schizophrenia, and Alzheimer's. However, ANNs are not very popular to model genotype SNP data due to high dimensionality and unstructured nature of the data. With high data dimensionality, ANNs tend to overfit. The works of [57, 58] have tried to solve this issue by developing a sparse artificial neural network that aggregates genetic risk using knowledge of gene-pathway interactions [56, 124]. While an essential first step, their ANN contains just a single hidden layer, which does not account for the hierarchical and interconnected nature of the biological pathways.

Autoencoders: Autoencoders [125] are direct applications of ANN. An autoencoder consists of an encoder branch and a decoder branch. Traditionally, the encoder and decoder branch is made of ANNs. The encoder branch extracts a non-linear lower-dimensional projection of the input data, and the decoder branch reconstructs the input data from the lower-dimensional projections. The decoder branch acts as

a regularizer, which ensures that the encoded representation contains informative information about the input. As a result autoencoders are widely used for denoising and extracting robust representations from the input data. Traditionally, autoencoders are trained by minimizing the following reconstruction loss:

$$\min_{\phi,\alpha} \sum_{n} ||\boldsymbol{i}_n - \mathcal{D}\left(\mathcal{E}\left(\boldsymbol{i}_n; \phi\right); \alpha\right)|| \tag{2.16}$$

where $\boldsymbol{i}_n$ is the input features from subject n, $\mathcal{E}(\cdot;\phi)$ is the encoder branch parametrized by ϕ, and $\mathcal{D}(\cdot;\alpha)$ is the encoder branch parametrized by α. The parameters of the autoencoder are training using backpropagation.

Autoencoders have direct applications in imaging genetics studies. Firstly, the autoencoder provides a natural way to integrate multiple modalities [126, 127] simply by adding new encoder-decoder branches. Mathematically, a new branch will introduce another term to the loss function but does not alter the optimization procedure (e.g., backpropagating gradients). Second, missing data can easily be handled [128] by freezing the affected part of the network and updating the remaining weights. This simplicity is in stark contrast to the classical methods, where the entire model and optimization procedure must be changed for each new modality and missing data configuration. Third, the latent encoding provides a data-driven feature space that can be used for patient/control classification. Again, this is in contrast to classical approaches, which are highly dependent on hand-crafted feature.

One main drawback is that traditional autoencoders lack interpretability. So, their application is limited in imaging genetics. Recently, interpretable AI has provided us with tools to compute feature importance for interpretability. These feature importance maps can be used to identify imaging and genetic biomarkers. However, these models often rely on heuristics based on the gradient of a loss function [129] or the importance of a feature to the downstream task [130]. In addition, these approaches do not provide a probabilistic measure of a feature being causal or not. Instead, they provide an

importance score dependent on the training data and cannot be compared across replication experiments.

In this thesis, we address these issues by developing an interpretable autoencoder that can simultaneously perform feature selection and integrate multiple data modalities. Our feature selection is based on a Bayesian framework, which results in robust probabilistic measures of feature importance maps.

Graph Neural Networks: Graph convolutional networks (GCNs) [131] provide a natural way to leverage the high-dimensional and interconnected relationships in the data. GCNs provide a strategy to combine information based on the neighborhood nodes in a graph. Their effectiveness is widely popular to model network structure in the brain [132]. They are also popular in protein structure prediction [133], drug discovery [134], and gene-gene interactions [135]. In imaging and genetics, a graph provides a natural way to capture the interaction between different components. For example, in imaging, the node in a graph could represent the activation of a brain region, and in genetics, each node could represent a gene or biological pathway.

The graph convolution operation provides an iterative strategy to transfer information between nodes. This message-passing operation is the key component of GCNs. Similar to CNN, multiple layers of message-passing operation provides a way to combine low level features to generate high-level feature representations. Mathematically, the message-passing operation can be shown in the following way:

$$\boldsymbol{h}^{l+1}(i) = \zeta\left(\sum_{j\in\mathcal{N}\rceil(i)} \boldsymbol{E}^{l}(i,j)\boldsymbol{h}^{l}(j)\boldsymbol{W}^{l} + \beta_t \boldsymbol{h}^{l}(i)\boldsymbol{W}_s\right) \tag{2.17}$$

where $\boldsymbol{h}^{l}(i) \in \boldsymbol{R}^{1\times d_l}$ is signal for node i at stage l, $\boldsymbol{W}^{l} \in \boldsymbol{R}^{d_l\times d_{l+1}}$ is the convolutional filter between stages l and $l+1$, β_t is the self-influence for node t, $\boldsymbol{W}_s \in \boldsymbol{R}^{d_l\times d_{l+1}}$ is the convolution filter for self loop, and $\zeta(\cdot)$ is the nonlinearity. The summation in Eq. (2.17) aggregates the influence over all neighborhood nodes of i. In GCNs,

the convolutional weight filters are shared across nodes, which results in restricted parameter space and prevents overfitting.

The traditional graph convolution networks are further extended to incorporate learnable edge weights using graph attention. This strategy allows to track the information flow through the graph [136]. The attention focuses on the most discriminative set of interactions between the nodes, often leading to better generalization on unseen data. Graph attention has been successful in biological applications, like predicting disease-RNA association [137] and essential gene prediction [138]. These models are often coupled graph pooling operations that allows us to aggregate information at each level of the hierarchy [139–141]. During standard graph pooling, the nodes at each level are clustered to form a smaller subgraph. These clusters can be obtained via deterministic algorithms [139] or be learned during training [140, 141].

We utilize graph attention networks coupled with graph convolutions to encode millions of genetic variants using a sparse graph of gene-pathway interaction. The graph contains a hierarchical structure pre-defined by our knowledge of gene ontology [56, 142]. The sparse hierarchy allows us to combine DNA-level information through strategically defined layers of SNP-gene interaction and gene-pathway interactions. We use graph convolution to capture the information flow through the graph and graph attention to learn the salient edges associated with the disorder. This strategy provides a robust framework to explore the biological mechanism associated with a disorder.

2.4 Data: Acquisition and Preprocessing

In this thesis, we investigated two neuropsychiatric disorders, autism and schizophrenia. Our datasets have been obtained from multiple sites, and in this section, we will provide a brief background about data acquisition and preprocessing.

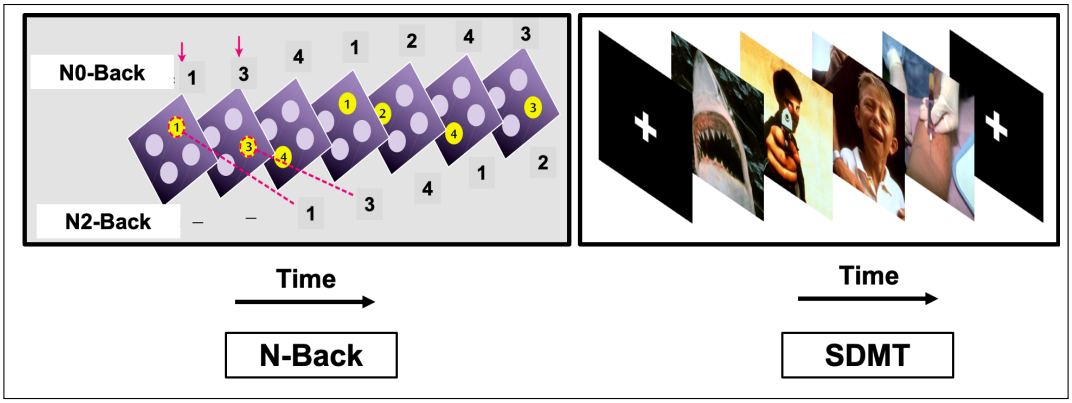


Figure 2-3. Left The experimental paradigm of the N-Back task. The top row shows a sample response for N0-Back and the bottom row shows a sample response for N2-Back. **Right** The experimental setup for the SDMT task.

2.4.1 Case Control Study of Schizophrenia

We validate our frameworks on a task fMRI and genetic data acquired at two different sites from two different study populations. The first dataset was provided by researchers at the Lieber Institute for Brain Development (LIBD) in Baltimore, MD, USA. The second dataset was acquired at the University of Bari Aldo Moro, Italy. The data collection procedures and pre-processing were consistent across sites.

Neuroimaging Data: As shown in Figure 2-3, our datasets include two fMRI paradigms that have been previously used to study schizophrenia [3, 4]. The first paradigm is a block design working memory task (N-Back). During the 0-back blocks, participants were instructed to press a button corresponding to a number displayed on the screen. During the 2-back working memory blocks, participants were instructed to press the button corresponding to the number they had seen two stimuli previously. We use a standard General Linear Model (described in Section 2.1.1) to estimate the activation coefficients from each block separately. The final contrast is the subtraction $\beta_{2-back} - \beta_{0-back}$. Our region-wise inputs are the average of these contrast values across all voxels in each particular region. The second paradigm is a block design declarative memory task (SDMT), which involves incidental encoding of complex

aversive visual scenes. Similar to the N-back analysis, we estimate the coefficients of association from a generalized linear model. The SDMT contrast map is the subtraction $\beta_{aversive} - \beta_{crosshair}$. Our region-wise inputs are the average of these contrast values across all voxels in each parcel of the brain. Further details for generating the contrast maps can be found in [8]. The case-control groups were matched on age, IQ (WRAT score), years of education, and in the case of N-Back, the percent correct response for the 2-Back task.

All fMRI data was acquired on 3-T General Electric Sigma scanners (EPI, TR/TE = 2000/28 msec; flip angle = 90; field of view = 24 cm, res = $3.75 \times 3.75 \times 6mm^3$ for NBack and res = $3.75 \times 3.75 \times 5mm^3$ for SDMT). fMRI preprocessing included slice timing correction, realignment, spatial normalization to an MNI template, smoothing and motion parameter regression. SPM12 was used to generate activation and contrast maps for each paradigm. We use the Brainnetome atlas [143] to define 246 cortical and subcortical regions. The input to our model is the average contrast map over these 246 ROIs. As fMRI data are often subject to noise, we average the activation across voxels in a single region to construct our model input. This averaging mitigates the impact of noise and helps us to find meaningful patterns across groups. In addition, we regress out the effect of age, IQ (WRAT reading score), years of education and percent-correct on the 2-back task for the N-back dataset, and we regress out the effect of age, IQ (WRAT reading score), years of education for the SDMT dataset. Regressing out the Nback performance removes biases in the data that may be due to cognitive performance per se. However, the SDMT contrast used in this work is specific to the encoding phases (aversive scenes vs. crosshair), so we do not regress retrieval performance. The subjects were not informed about the retrieval portion beforehand, so the encoding is incidental [3]. In all cases, we estimate the regression coefficients only from the training set and use them for the test set.

Genetic Data: Genotyping was done using variate Illumina Bead Chips including 510K/ 610K/660K/2.5M. Quality control and imputation were performed using PLINK and IMPUTE2, respectively. The resulting 102K linkage disequilibrium independent SNPs ($r^2 < 0.1$ in 500kb) are used to obtain our genetic data (see [30] for further details). In our works, we use a GWAS to generate gene scores and polygenic risk scores. This study was done on 36,989 schizophrenia patients and 113,075 neurotypical controls by the PGC Consortium. Further details about this study can be found in [17].

2.4.2 Case Control Study of Autism Spectrum Disorder (ASD)

The final part of this thesis explores the genetic foundation of autism. We explore the interactions between genes and pathways to parse the genetic risk associated with autism. Our genetic data for autism are collected from two studies:

Simons Simplex Collections (SSC): DNA of individuals from Simons Simplex Collection (SSC) families were genotyped for a million or more single nucleotide polymorphisms (SNPs) on one of three array versions - Illumina 1Mv1, Illumina 1Mv3 Duo, or Illumina HumanOmni2.5M. Members of each family were analyzed on the same array version. All families were simplex [144], with only the proband being affected by ASD, as assessed by evaluation of family history. Probands, ages 4 to 18, were diagnosed using the ADI-R [145] and ADOS [146] as administered by expert clinicians. Initial preprocessing and imputation are done by the RICOPILI pipeline [16, 147]. The resulting $\sim$ 2591 families are used as inputs to our models.

Autism Centre of Excellence (ACE): Families were recruited by the Autism Centre of Excellence (ACE) Network. Subjects were genotyped on the Illumina Omni-2.5 platform using standard manufacturer protocols (Illumina, San Diego, CA). All DNA samples were hybridized and scanned on the Illumina iScan to minimize batch

effects and variation. All subjects had a genotyping call rate > 95%. Genotyping data were analysed by PLINK v1.0731 using the forward strand and confirmed the reported sex and sibling relationships of all subjects. Autism diagnoses were derived from a combination of assessments on the Autism Diagnostic Interview-Revised (ADI-R) [8] and/or Autism Diagnostic Observation Schedule (ADOS) [146] and clinician's best judgment according to standard protocols at AGRE [148]. Initial preprocessing and imputation are done by the RICOPILI pipeline [147]. After imputation, we subselect 346 subjects belonging to ASD and control groups as input to the model.

Chapter 3

Matrix Decomposition Frameworks Parsing Complex Interactions Between Imaging and Genetics

Initial studies [16, 17, 25] explored the imaging and genetic data separately to pinpoint neural mechanisms and genetic variants. However, imaging and genetics data contain complementary information about the disorder. Imaging biomarkers provide insights into aberrant neural activity. On the other side, genetics data provide the underlying etiology of the disorder. Prior works [11, 14, 110] have combined imaging and genetics data to find a set of correlated biomarkers, but they often ignore patient heterogeneity, leading to clinically irrelevant biomarkers. Other predictive models [114, 149] try to subselect features and pass them through a machine learning classifier for disease diagnosis.

In imaging-genetics, dictionary learning models provide a joint approach to combining multiple data modalities in a single framework. Dictionary learning models use matrix decomposition to represent the observed data as a linear sum of basis vectors. The adaptable framework can simultaneously extract common data representations from multiple modalities [150]. Additionally, such models easily incorporate additional priors over the generative assumptions [151], leading to a robust framework.

In this chapter, we extend the dictionary learning-based models to capture patient

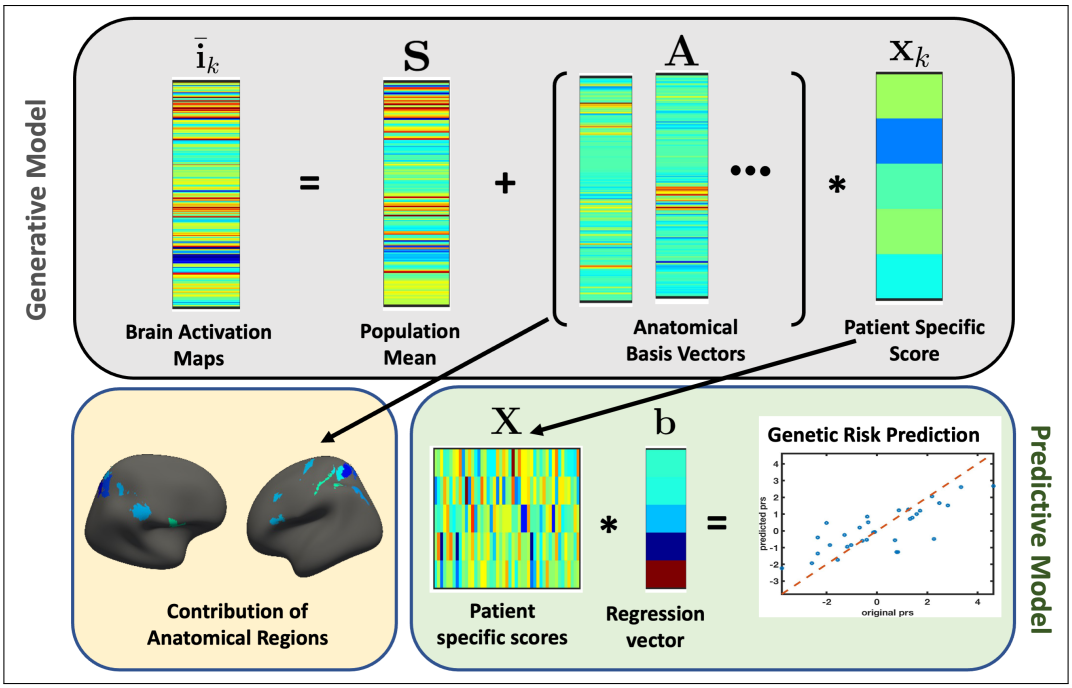


Figure 3-1. The joint modeling framework to capture brain activity and genetic risk. The gray box represents the generative part of the model for a single schizophrenia patient. We captured altered brain activity in the patients as deviations from the population mean. The major contribution of the anatomical regions to overall deviation are shown as surface plots in the yellow box. The green box is the predictive part of the model that track the genetic risk as linear regression.

heterogeneity while finding clinically relevant biomarkers. This chapter will be based on our published works [59–61]. In Section 3.1, we will introduce a generative-predictive framework that captures the differences in regional brain activity between a neurotypical cohort and a clinical population, as guided by polygenic risk scores. One limitation of this work is that we collapse all the SNP information into a single scalar value, which cannot consider the interactions between the SNPs. We further extend this work in Section 3.2, where we combine the raw SNP data with imaging brain activation maps in a generative-discriminative framework.

3.1 Aberrant Neural Activity in Schizophrenia and its Association with Polygenic Risk Scores

Fig. 3-1 represents a overview of our joint modelling approach. The generative part of our model is very closely related to dictionary learning. The modelling of the generative part is based on the assumption that the average functional activity of the clinical group differs from their neurotypical counterparts in certain brain regions, which can be approximated by a set of sparse basis vectors. We rely on the $\ell_{2,1}$ norm for group sparsity; this norm has been previously used for feature selection and for localizing quantitative traits to predict cognitive outcomes [47, 152]. However, its application to identify group-level changes in brain activity while tracking genetic risk has not been explored. The predictive part of our model is a linear regression model, where the feature vectors are constructed as projections of the data onto the subspace spanned by the basis vectors. Here, we assume a linear relationship between the feature vectors and the genetic risk. Our joint optimization enables us to learn a set of regions that capture the group differences in brain activity and a set of projection coefficients, which capture the variability in genetic information across patients.

3.1.1 The Generative Framework

This section provides a formal mathematical description of our generative-predictive framework. Mathematically, let J denote the number of normal controls, and let K be number of clinical patients. We assume that the brain has been parcellated into R ROIs, from which we extract an $R \times 1$ vector $\mathbf{i}$, that quantifies the functional activation across the ROIs. The inputs to our model are the feature vectors $\{\mathbf{i}_j\}_{j=1}^{J}$ for neurotypical controls and $\{\bar{\mathbf{i}}_k\}_{k=1}^{K}$ for clinical patients, along with the patient-specific polygenic risk scores $\{r_m\}_{m=1}^{M}$.

3.1.1.1 Modelling the neurotypical control population

Throughout our analysis we assume that the brain activation of the neurotypical control population is distributed across a population mean, $\boldsymbol{s}$ whereas the neuropsychiatric patient population is distributed across a shifted version of this population mean. So, we model the functional activation of the control group in the following way:

$$\boldsymbol{i_j} = \boldsymbol{s} + \boldsymbol{n} \quad \text{where}\, \boldsymbol{n}(i) \sim \text{iid}, \tag{3.1}$$

where σ^2 is the variance of the noise associated with the activation of each region.

3.1.1.2 Modelling the neuropsychiatric patient population

We hypothesize that the given neurological disorder manifests as coordinated disruptions over a set of brain regions. Accordingly, our model stipulates that the deviation caused by the disorder in different brain regions can be approximated by a set of sparse basis vectors. Unlike the control population the brain activation of the clinical patients are distributed across a shifted mean $\boldsymbol{s} + \boldsymbol{A}\boldsymbol{x}_k$. Hence, we model the activation of region i in patient k as:

$$\bar{\boldsymbol{i}}_k(i) \approx \begin{cases} \boldsymbol{s}(i) & \text{when: } \boldsymbol{A}(i,\cdot) = 0 \\ \boldsymbol{s}(i) + \boldsymbol{A}(i,\cdot)\boldsymbol{x}_k & \text{when: } \boldsymbol{A}(i,\cdot) \neq 0 \end{cases} \tag{3.2}$$

where $\boldsymbol{A} \in \mathbb{R}^{R\times d}$ is the set of sparse canonical basis vectors that captures the contribution of each region to overall activation differences. As seen in Eq. (3.2) if $\boldsymbol{A}(i,\cdot) \approx 0$, the mean activation at region i can be well approximated by the neurotypical mean, $\boldsymbol{s}(i)$, but if $\boldsymbol{A}(i,\cdot) \gg 0$ then the patient population has a substantially different activation contribution at region i than the population mean. Hence, the matrix $\boldsymbol{A}$ captures the set of brain regions that are substantially affected by the neurological disorder. Moreover, the feature vector $\boldsymbol{x}_k$ captures the patient heterogeneity within the cohort. In this model, we assume an additive effect for all the basis vectors, so we introduce the non-negativity constraint $\boldsymbol{x}_k \geq 0$ on the coefficients. At a

high level, our decomposition reduces the data dimensionality while simultaneously capturing patient heterogeneity.

3.1.2 The predictive framework

As shown in Fig. 3-1, we use the patient specific coefficients $\{\boldsymbol{x}_k\}_{k=1}^K$ to predict the genetic risk in a linear regression model. We concatenate the coefficients $\{\boldsymbol{x}_k\}_{k=1}^K$ into a matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K] \in \mathbb{R}^{d \times K}$ and the patient specific risk scores into a vector, $\boldsymbol{r} = [r_1, \ldots, r_K] \in \mathbb{R}^{K \times 1}$. We then fit them into a linear regression model:

$$\boldsymbol{r} \approx \boldsymbol{X}^T \boldsymbol{b}$$

where $\boldsymbol{b} \in \mathbb{R}^{d \times 1}$ is the regression vector. We include an ℓ_2 regularization on $\boldsymbol{b}$.

3.1.2.1 The joint model

We combine both the generative and predictive terms into a joint objective, which can be expressed as a matrix decomposition and regression framework as follows:

$$\mathcal{J}(\boldsymbol{x}_1, ..., \boldsymbol{x}_K, \boldsymbol{s}, \boldsymbol{b}, \boldsymbol{A}) = \sum_{j=1}^{J} ||\boldsymbol{i}_j - \boldsymbol{s}||_2^2 + \sum_{k=1}^{K} ||\bar{\boldsymbol{i}}_k - \boldsymbol{s} - \boldsymbol{A}\boldsymbol{x}_k||_2^2 + \lambda_3 \sum_{k=1}^{K} ||r_k - \boldsymbol{x}_k^T \boldsymbol{b}||_2^2$$
$$\text{Subject to:} \quad \{\boldsymbol{x}_k\}_{k=1}^K > 0$$

where $\sum_{j=1}^{J} ||\boldsymbol{i}_j - \boldsymbol{s}||_2^2 + \sum_{k=1}^{K} ||\bar{\boldsymbol{i}}_k - \boldsymbol{s} - \boldsymbol{A}\boldsymbol{x}_k||_2^2$ represents the cost associated with modelling the fMRI data and $\lambda_3 \sum_{k=1}^{K} ||r_k - \boldsymbol{x}_k^T \boldsymbol{b}||_2^2$ represents the cost associated with predicting genetic risk. The parameter λ_3 denotes the trade-off between the data representation term and the predictive term.

3.1.3 Regularization Penalties

We would like the matrix $\boldsymbol{A}$ to capture a representative set of regions where the brain activations are affected by the disease. We enforce that by putting a sparsity constraint across the rows of $\boldsymbol{A}$. Further we want to model $\boldsymbol{A}$ to implicate a sparse

set of regions. We enforce this by putting a smoothness constrain across the columns of $\boldsymbol{A}$. We combine these two in the form of $\ell_{2,1}$ norm which is $||\boldsymbol{A}||_{2,1} = \sum_i ||\boldsymbol{A}(i,\cdot)||_2$. This norm enforces a smoothness constrain across columns and sparsity constraint across rows. Further, from an optimization standpoint different scaled result of $\{\boldsymbol{X}, \boldsymbol{b}\}$ can lead to the same solution. So, we need to introduce a quadratic penalty over $\boldsymbol{X}$ as $\lambda_1 \sum_k^K ||\boldsymbol{x}_k||_2^2$. Similarly, we also need to introduce a quadratic penalty over $\boldsymbol{b}$ as $\lambda_2||\boldsymbol{b}||_2^2$ which is similar to ridge regression. Gathering these terms, the final regularization cost is:

$$\lambda_0||\boldsymbol{A}||_{2,1} + \lambda_1 \sum_k^K ||\boldsymbol{x}_k||_2^2 + \lambda_2||\boldsymbol{b}||_2^2$$

Now, the complete cost function takes the following form:

$$\begin{aligned}\mathcal{J}(\boldsymbol{x}_1, ..., \boldsymbol{x}_K, \boldsymbol{s}, \boldsymbol{b}, \boldsymbol{A}) = &\sum_{j=1}^{J} ||\boldsymbol{i}_j - \boldsymbol{s}||_2^2 + \sum_{k=1}^{K} ||\bar{\boldsymbol{i}}_k - \boldsymbol{s} - \boldsymbol{A}\boldsymbol{x}_k||_2^2 + \lambda_3 \sum_{k=1}^{K} ||r_k - \boldsymbol{x}_k^T \boldsymbol{b}||_2^2 \\ &+ \lambda_0||\boldsymbol{A}||_{2,1} + \lambda_1 \sum_k^K ||\boldsymbol{x}_k||_2^2 + \lambda_2||\boldsymbol{b}||_2^2 \qquad (3.3) \\ \text{Subject to:} \quad &\{\boldsymbol{x}_k\}_{m=1}^K > 0\end{aligned}$$

3.1.4 Optimization Strategy

We optimize Eq. (3.3) via the alternating minimization procedure illustrated in Fig. 3-2, in which we update each variable independently while holding the others constant. The process is computationally efficient since that the cost function $\mathcal{J}(\cdot)$ is convex over $\boldsymbol{s}$, $\boldsymbol{A}$, $\boldsymbol{x}_k$, and $\boldsymbol{b}$ independently. The variables $\{\boldsymbol{s}, \boldsymbol{b}\}$ have closed form updates, and we use an iterative method to update $\boldsymbol{A}$ and $\boldsymbol{x}_k$. This optimization strategy is further described below.

3.1.4.1 Closed form update for $\boldsymbol{s}$

The global minimizer of $\boldsymbol{s}$ can be found by setting the gradient of $\mathcal{J}(\cdot)$ with respect to $\boldsymbol{s}$ equal to zero. The update for $\boldsymbol{s}$ relies on both the neurotypical and patient imaging

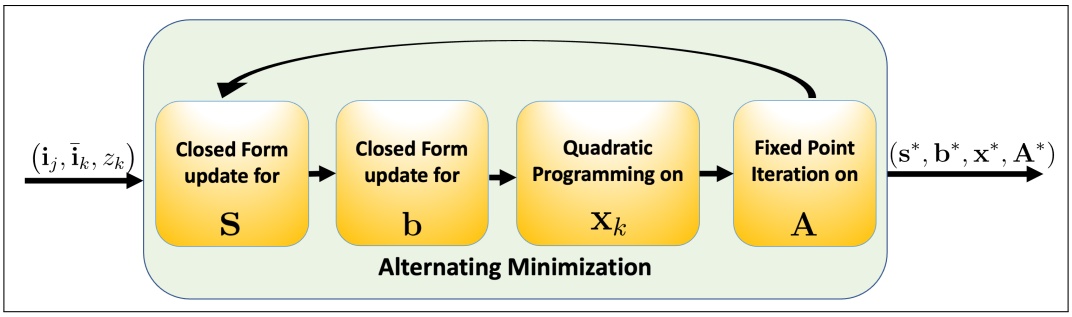


Figure 3-2. The alternating minimization approach to obtain the set of minimizers.

data:

$$\boldsymbol{s}^* = \frac{\sum_j^J \boldsymbol{i}_j + \sum_k^K (\bar{\boldsymbol{i}}_k - \boldsymbol{A}\boldsymbol{x}_k)}{J + M}$$

3.1.4.2 Closed form update for b

The regression coefficient of the prediction term also has a closed-form update. It can also be found by setting the gradient of $\mathcal{J}(\cdot)$ with respect to $\boldsymbol{b}$ to zero. The update is given by:

$$\boldsymbol{b}^* = (\lambda_2 \mathbb{I} + \boldsymbol{X}\boldsymbol{X}^T)^{-1}(\boldsymbol{X}\boldsymbol{r}) \tag{3.4}$$

where we have concatenated the projections $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K]$, and clinical scores $\boldsymbol{r} = [r_1, \ldots, r_K]^T$ for convenience. Notice that Eq. 3.4 parallels the least square regression solution.

3.1.4.3 Optimizing A using fixed point iteration:

The update rule for $\boldsymbol{A}$ does not have a closed form solution due to the $\ell_{2,1}$ regularization term. However, it can be efficiently updated using a fixed point iteration. In this method the ℓ_2 norm of each row of $\boldsymbol{A}$ is kept constant to its value from the previous iteration, leading to the modified objective:

$$\boldsymbol{A}^{t+1} = \underset{\boldsymbol{A}}{\operatorname{argmin}} \sum_{k=1}^{K} ||\bar{\boldsymbol{i}}_k - \boldsymbol{s} - \boldsymbol{A}\boldsymbol{x}_k||_2^2 + \lambda_0 \sum_{i=1}^{R} \frac{||\boldsymbol{A}(i,\cdot)||_2^2}{2||\boldsymbol{A}^t(i,\cdot)||} \tag{3.5}$$

Eq. (3.5) has a closed form update for each row of $\boldsymbol{A}$ according to the following expression:

$$\boldsymbol{A}^{t+1}(i,\cdot) = \bar{\boldsymbol{I}}(i,\cdot)\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{X}^T + \frac{\lambda_0}{2||\boldsymbol{A}^t(i,\cdot)||_2}\mathbb{I}\right)^{-1}$$

where we have concatenated the observed patient activations $\{\bar{\boldsymbol{i}}_m\}_{m=1}^M$ as $\bar{\boldsymbol{I}} = [\bar{\boldsymbol{i}}_1, \ldots, \bar{\boldsymbol{i}}_K] \in \mathbb{R}^{R\times K}$ for convenience. The proof of convergence for this fixed point iteration can be found in Wang et al [47].

3.1.4.4 Optimizing $\boldsymbol{x}_k$ using quadratic programming

The objective function is quadratic in $\boldsymbol{x}_k$ when the other variables are kept constant. Moreover, the patient-specific projection coefficients decouple into K independent quadratic equations, which take the form:

$$\begin{aligned}\boldsymbol{x}_k^* &= \underset{\boldsymbol{x}_k}{\operatorname{argmin}}\, \boldsymbol{x}_k^T\boldsymbol{Q}\boldsymbol{x}_k + \boldsymbol{c}^T\boldsymbol{x}_k \\ \text{Subject to:} \quad & \boldsymbol{B}_k\boldsymbol{x}_k \leq \boldsymbol{d}_k\end{aligned}$$

The cost and the constraints are computed from the original variables in Eq. (3.3):

$$\begin{aligned}\boldsymbol{Q} &= \boldsymbol{A}^T\boldsymbol{A} + \boldsymbol{b}\boldsymbol{b}^T \\ \boldsymbol{c} &= -2(\bar{\mathbf{i}}_m - \boldsymbol{s})^T\boldsymbol{A} \\ \boldsymbol{B} &= -\mathbb{I}_d \\ \boldsymbol{d}_m &= [0, \ldots, 0]^T\end{aligned}$$

The quadratic solvers give us globally optimal solutions for all the patient-specific feature vectors.

3.1.5 Model Evaluation

3.1.5.1 Baseline algorithms

LASSO: We perform a comparison of our proposed model with LASSO regression, which assumes a multivariate linear association between the feature vectors $\{\bar{\boldsymbol{i}}_k\}$ and

the polygenic risk scores $\{r_k\}$. Mathematically:

$$\boldsymbol{r} = \bar{\boldsymbol{I}}^T \boldsymbol{\beta} + \lambda ||\boldsymbol{\beta}||_1 \quad , \tag{3.6}$$

where $\bar{\boldsymbol{I}} = [\bar{\boldsymbol{i}}_1, \ldots, \bar{\boldsymbol{i}}_M] \in \mathbb{R}^{R \times K}$. As seen in Eq. (3.6), we estimate the regression coefficients $\boldsymbol{\beta}$, whose non-zero entries indicate region-wise associations between genetic risk and functional activation. We threshold the regression coefficients to obtain a binary vector $\boldsymbol{I}_{lasso} \in \{0,1\}^{N \times 1}$ with the highest region-wise association:

$$\boldsymbol{I}_{lasso}(i) = 1 \quad \text{if } |\boldsymbol{\beta}(i)| > \sigma \tag{3.7}$$

$$= 0 \quad \text{Otherwise} \tag{3.8}$$

We use this binary vector to evaluate the performance during boostrapping which we discuss in Section 3.1.5.2.

Random Forest: We also compare our model with Random Forest (RF) regression, which estimates a nonlinear association between the features $\bar{\boldsymbol{i}}_k$ and the genetic scores r_k. It is an ensemble learning method that fits multiple regression trees on random subsets of the data. The randomness prevents the trees from overfitting and reduces the error variance while keeping the bias constant. The final prediction is the average over all of these trees. Here, the importance of each feature is quantified by the change in error when values of that predictor are randomly permuted. Features with high importance result in higher change in error. Similar to LASSO, we obtain a binary vector $\boldsymbol{I}_{rf} \in \{0,1\}^{R \times 1}$ by retaining the top 70% of the features according to their importance. We quantified the performance through bootstrapping which we describe in Section 3.1.5.2.

Generative-Predictive Model: The matrix $\boldsymbol{A}$ in our generative-predictive (gp) framework quantifies the region association strength to each basis network. We compute a single region measure via the ℓ_2 norm across the bases. These values are

thresholded to obtain the final $\boldsymbol{I}_{gp} \in \{0,1\}^{N\times 1}$.

$$\boldsymbol{I}(i) = 1 \quad \text{if } ||A(i,\cdot)||_2 > \sigma \tag{3.9}$$

$$= 0 \quad \text{Otherwise} \tag{3.10}$$

where σ is the threshold. Similar to LASSO and RF we will use the binary vectors to quantify the performance of our model across bootstrapped subsets of our data.

3.1.5.2 Performance Metrics

We evaluate the performance of our framework and both the baseline methods in terms of reproducibility, i.e how consistent the inferred associations are across different subsets of data. We quantify this performance using two different metrics; (1) **Jaccard Index**, and (2) **Fractional Occurence**.

We evaluate both metrics using bootstrapping. Bootstrapping is a statistical method that relies on random sampling of data with replacement. The main idea behind bootstrapping is that inferences that are consistent across random subsets of the data are more likely to generalize beyond the experiment. In all our methods, we randomly sampled 90% of our data with replacement for 100 bootstrapping trials. After each trial, $\boldsymbol{t}$ we obtain the binary vectors $\{\boldsymbol{I}^t_{lasso}, \boldsymbol{I}^t_{rf}, \boldsymbol{I}^t_{gp}\}$ as described previously. So, via bootstrapping we get 100 binary vectors for each of the methods, which we use to quantify consistency.

Jaccard Index: This measure quantifies the overlap between two vectors, i.e.,

$$\mathbb{J}(\boldsymbol{I}^s_m, \boldsymbol{I}^t_m) = \frac{\sum_{i=1}^{R} \boldsymbol{I}^s_m(i)\boldsymbol{I}^t_m(i)}{\max\Big(\text{Card}(\boldsymbol{I}^s_m), \text{Card}(\boldsymbol{I}^t_m)\Big)} \tag{3.11}$$

$$\tag{3.12}$$

where m denotes the method under consideration i.e. $\{lasso, rf, gp\}$ and Card($\boldsymbol{I}^t$) denotes the number of non zero elements in the binary vector. Since, we ran 100

bootstrapping iterations, we have 4950 Jaccard indices for each method. We can assess both the average consistency and the variability of the Jaccard index across subsets.

Fractional Occurrence: We introduce another metric to identify the consistency of all the methods. It is defined as the average number of times each region appears in the binary vector across all the bootstrapping trials. The fractional occurrence of region i is computed as:

$$\mathbb{F}_m(i) = \frac{1}{100} \sum_{i=1}^{100} \boldsymbol{I}_m(i) \tag{3.13}$$

where m again denotes the method under consideration. Fractional occurence is closely tied with the Jaccard similarity index. A high fractional occurrence across all the regions will result in a high Jaccard similarity index and vice-versa. However, the Jaccard index is essentially a summary statistic over all the regions, whereas fractional occurrence gives us the individual statistics of each region.

3.1.5.3 Parameter Settings

In our model the hyperparameters $\{\lambda_0, \lambda_1, \lambda_2, \lambda_3\}$ are user specified. We swept over two orders of magnitude for each parameters and over feature dimensions $d = 1, \ldots, 8$. Our final setting was $d = 5$, $\lambda_0 = 6.4$, $\lambda_1 = 0.4$, $\lambda_2 = 0.05$, and $\lambda_3 = 1$ based on optimizing the Jaccard measure. We also swept over different parameter settings for our baseline methods. In LASSO we looked over two orders of magnitude to identify the optimal parameter ($\lambda = 0.01$) that gives the lowest mean square error in Eq. (3.6). In Random Forest we used 1000 randomized regression trees for predicting the genetic risk based on optimizing the Jaccard measure.

3.1.6 Experimental Results

We evaluate the performance of our model on a case-control study of schizophrenia obtained from Lieber Institute for Brain Development (LIBD). The fMRI data is

collected from the N-back working memory task defined in Section 2.4.1. The groups were matched for age, IQ, gender, education, and % correct on the N2-back working memory task. We use the Braintome atlas [143] to define 246 cortical and subcortical regions. The time courses for each region was then fed into a Generalized Linear Model [8] to obtain the activation maps β_0 for 0-back and β_2 for 2-back. The input to our model is the average contrast map $(\beta_2 - \beta_0)$ across the 246 regions. The schizophrenia polygenic risk score for each individual was calculated as the sum of the GWAS imputation probability of reference alleles weighted by the natural log of odds ratio [30]. The genetic risk scores were constructed using SNPs with GWAS association P-value < 0.05.

Fig. 3-3 illustrates the distribution of Jaccard indices for each method based on five number summary, minimum of the data, first quartile, median, third quartile, and maximum. The box plot gives us a good idea of how tightly the data are grouped and if and how the data are skewed. As seen in Fig. 3-3 our generative-predictive model demonstrates superior performance in terms of consistency since the median is significantly higher than the baseline methods. This improvement can be attributed to the structured form of group sparsity, which forces our model to identify the regions of differential contrast but sufficient patient variability to capture genetic risk. In contrast, LASSO and RF identify only the regions of high activity variation in the patient group, which differs across subsets of data. This behavior leads to low reproducibility between bootstrapping trials.

From our discussion in Section 3.1.5.2, we know that Jaccard index is strongly coupled with fractional occurrence. A high Jaccard index should also mean a high fractional occurrence for each of the regions. Fig. 3-4 evaluates the robustness in fractional occurrence for each region across all bootstrapping trials. We have colored each region according to their fractional occurrence and the color bar gives the

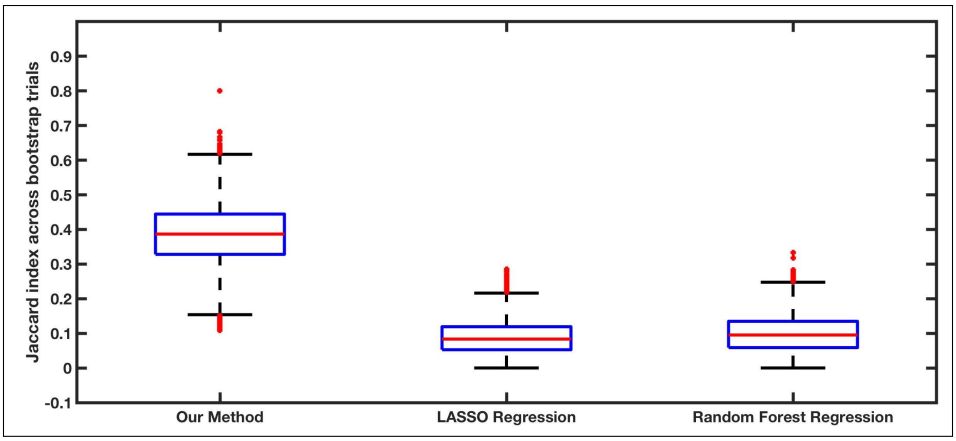


Figure 3-3. The distribution of the Jaccard similarity indices for each of the three methods are shown.

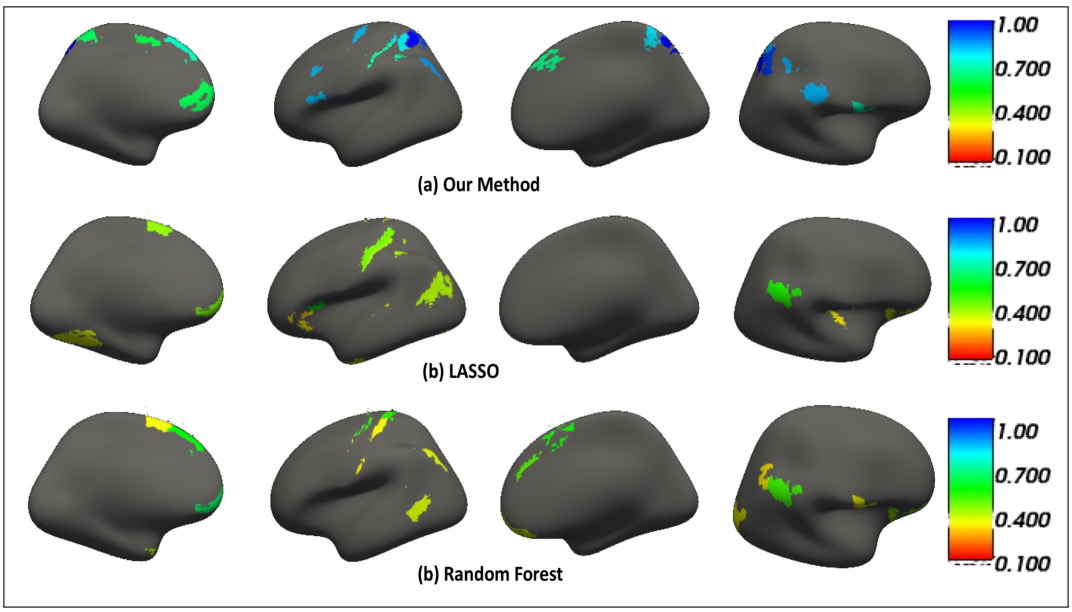


Figure 3-4. **(a)** The fractional occurrence($\mathbb{F}_{gp}$) of the set of regions identified by our generative-predictive model. **(b)** The fractional occurrence($\mathbb{F}_{lasso}$) of the set of regions identified by lasso. **(c)**The fractional occurrence($\mathbb{F}_{rf}$) of the set of regions identified by random forest. For visualization the regions are colored according to their fractional occurence. Blue indicates a high fractional occurence, and red indicates a low fractional occurence. From **Left** to **Right** the images are internal surface of left hemisphere, external surface of left hemisphere, internal surface of right hemisphere, and external surface of right hemisphere.

associated values. As expected our method shows a higher fractional occurrence in the identified set of regions than the two baseline methods.

Table 3-I. The table shows the implicated set of regions identified by our generative-predictive framework, lasso and random forest regression along with the corresponding fractional occurrence.

Methods	Implicated Regions	Fractional Occurrence
Generative-Predictive	Superior Frontal Gyrus (BA - 9)	0.7
	Inferior Frontal Gyrus (BA - 44, 45)	0.83
	Supramarginal Gyrus (BA - 40)	0.78
	Cingulate Gyrus (BA - 24)	0.63
	Precuneus (BA - 7, 5)	0.99
	Angular Gyrus (BA - 39)	0.86
	Superior Parietal Lobule (BA - 7)	0.9
LASSO	Superior Frontal Gyrus (BA - 8)	0.38
	Orbital Gyrus (BA - 11)	0.41
	Precentral Gyrus (BA - 4)	0.47
	Superior Temporal Gyrus (BA - 32)	0.33
	Middle Temporal Gyrus (BA - 37)	0.31
	Inferior Temporal Gyrus (BA - 20)	0.34
	Angular Gyrus (BA - 39)	0.39
	Supramarginal Gyrus (BA - 40)	0.37
	Postcentral Gyrus (BA - 1,2 3)	0.37
RF	Superior Frontal Gyrus (BA - 8, 9)	0.53
	Orbital Gyrus (BA - 11, 12)	0.49
	Postcentral Gyrus (BA - 1,2, 3)	0.38
	Precentral Gyrus (BA - 4)	0.42
	Angular Gyrus (BA - 39)	0.38
	Postcentral Gyrus (BA - 1, 2, 3)	0.38
	Parahippocampal Gyrus (BA - 35, 36)	0.39
	Supramarginal Gyrus (BA - 40)	0.43

Table. 3-I reports the most consistent regions identified by each method. We observe that the set of regions identified by our model include superior frontal gyrus, inferior frontal gyrus, cingulate gyrus, and supramarginal gyrus, all regions well known to subserve executive function including working memory and implicated in the pathophysiology of executive cognition deficits observed in patients with schizophrenia [4]. In contrast, while few regions of LASSO and RF regression are same as in our generative-predictive model, other regions identified by LASSO and RF included the postcentral gyrus, middle temporal gyrus, precentral gyrus, orbital gyrus, and inferior

temporal gyrus. These regions are not strongly associated with the N-back working memory task [4]. Taking this as further evidence, along with the lower Jaccard index and fractional occurrence, we can conclude that both LASSO and RF regression could partly be capturing noise. However, our generative-predictive framework leverages the heterogeneity in genetic risk to compensate for noise. As a result we find differential functional activity in the canonical brain regions underlying cognitive processing required for working memory.

3.1.7 Discussion and Summary

We have introduced a novel matrix decomposition framework that identifies differential regional brain activity that is modulated by genetic risk. Our approach uses group sparsity to select a representative set of features that have a linear association with the patient-specific genetic risk scores. This strategy provides a richer set of features that leverages the information of differential functional activity and genetic variation. Additionally, we leverage genetic patient heterogeneity to identify consistent and robust region assignments across bootstrapping experiments. We demonstrate that our generative-predictive model significantly outperforms two baseline methods that do not leverage patient heterogeneity, in terms of both consistency and robustness. Our generative-predictive model is not tied to any specific paradigm and can be used to draw associations between a variety of neuroimaging phenotypes and variables beyond genetic risk, such as clinical, cognitive, and behavioral scores.

One limitation of our approach is that we identify regions of aberrant neural activity associated with the genetic risk scores. This approach ignores the interaction with individual SNPs. Previous studies [11] have found networks of brain regions interact with multiple SNPs, leading to a fine-grained understanding of the disorder. In addition, we also fail to account for the disease status to guide our biomarker identification.

The next chapter extends this work by investigating the role of single nucleotide polymorphisms (SNPs) in the neurobiology underlying executive cognitive deficits in patients with schizophrenia. We will also explore the efficiency of the model in capturing data variability with multiple imaging data modalities.

3.2 A Generative-Discriminative Framework Exploring Interactions Between Brain Activity and Genetic Variants Guided by the Diagnosis Labels

This work first appeared in [60, 61]. In these papers, we introduce a new optimization framework that uses disease status to regularize the projection of imaging and genetics data onto a shared low-dimensional subspace. This projection is done through a coupled dictionary learning framework. The imaging and genetic bases in this framework provide interpretable biomarkers in each modality, and the patient specific projections into this space are used to classify disease status through a logistic regression model.

In contrast to prior works [11, 13, 14], this framework provides an integrated approach for predicting disease status while finding clinically relevant biomarkers. We show the capabilities of this model of finding biomarkers in a simulation study under multiple noise settings. In addition, we provide statistical validation and a replication study to show the robustness of our approach. Finally, we perform an exploratory pathway analysis on the genetic biomarkers to identify disease-relevant pathways.

3.2.1 Coupled Generative-Discriminative Framework

Fig. 3-5 presents an overview of our imaging-genetic framework. The inputs to the model for each subject n are a vector of region-wise imaging features $\boldsymbol{i}_n$, a vector of genetic SNP variants $\boldsymbol{g}_n$, and patient versus control diagnosis $y_n \in \{0, 1\}$. As seen, our model consists of a generative module and a discriminative module. The

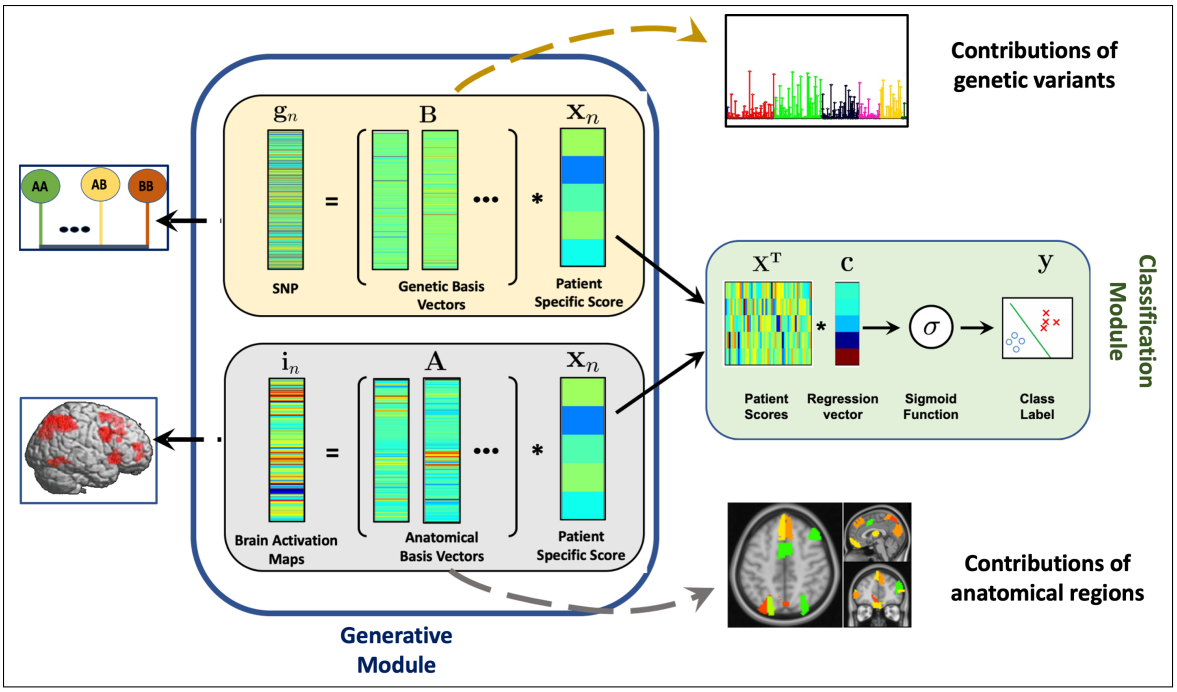


Figure 3-5. Generative-discriminative framework linking imaging ($\boldsymbol{i}_n$), genetics ($\boldsymbol{g}_n$), and diagnosis (y_n). The generative module captures the brain activations and the genetic data in a dictionary learning setup, and the discriminative module tracks the disease status using logistic regression. The classification module also guides the generative process to find a low dimensional space where the patient specific scores $\boldsymbol{x}_n$ are maximally separated. Therefore, the basis vectors $\{\boldsymbol{A}, \boldsymbol{B}\}$ identify biomarkers which capture group level differences between patients and controls. We have shown representative contributions of these basis vectors in the form of a Manhattan plot and a colored brain plot.

generative module is closely related to dictionary learning, where we have coupled the representation of imaging and genetic features by tying them to a common latent space. The discriminative module implements a logistic regression using the patient specific scores, thus ensuring that the latent space captures discriminative facets of the data. Our joint optimization enables us to learn both group level and patient specific information.

3.2.2 Feature Representation using Dictionary Learning

In our model, we assume that the brain has been parcellated into R ROIs, from which we extract an $R \times 1$ vector $\boldsymbol{i}_n$, that quantifies the functional activation across the

ROIs. Our model assumes that $\boldsymbol{i}_n$ can be represented by a low dimensional projection, i.e.,

$$\boldsymbol{i}_n \approx \boldsymbol{A}\boldsymbol{x}_n \quad \text{s.t.} \quad \boldsymbol{A}^T\boldsymbol{A} = \mathbb{I} \tag{3.14}$$

where the columns of $\boldsymbol{A} \in \boldsymbol{R}^{R\times d}$ correspond to the basis vectors and $\boldsymbol{x}_n$ are subject-specific projection weights. The basis vectors capture common patterns across the population, whereas the projection vector describes subject variability. We incorporated an orthogonality constraint over $\boldsymbol{A}$ to remove redundancy from the basis vectors. We also introduce a graph-based Laplacian regularizer on the basis matrix $\boldsymbol{A}$ to enforce that the highly correlated brain regions play a similar role in projection:

$$\boldsymbol{Tr}(\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A}) = \sum_{(i,j)} w_{ij}||\boldsymbol{a}_i - \boldsymbol{a}_j||_2^2 \tag{3.15}$$

where $\boldsymbol{a}_i$ denotes the i^{th} row of $\boldsymbol{A}$, and w_{ij} is the Pearson correlation between the activation map of region i and region j across the training data. To ensure convexity, we threshold these correlations to be positive.

The fMRI data is acquired while the subjects perform a standardized task in the scanner. Hence, most of the data variance will be concentrated in a consistent set of brain regions across subjects. The orthogonality constraint in our model reduces the redundancies in the learned bases vectors while simultaneously ensuring that they capture most of the data variance.

In our model for the genetic data we use a set of LD independent SNPs represented as $\boldsymbol{g}_n$. Let G denote the number of genetic variants under study, so the genetic data has dimensionality $\boldsymbol{g}_n \in \boldsymbol{R}^{G\times 1}$. We represent $\boldsymbol{g}_n$ as a linear combination of basis vectors, i.e.,

$$\boldsymbol{g}_n \approx \boldsymbol{B}\boldsymbol{x}_n \tag{3.16}$$

where $\boldsymbol{B}$ is the basis matrix. Notice that we have coupled the imaging and genetic domains by tying them to the same latent projection $\boldsymbol{x}_n$. We introduce an ℓ_{21} penalty

on the basis matrix as regularization. Mathematically,

$$||\boldsymbol{B}||_{2,1} = \sum_{i=1}^{G} ||\boldsymbol{b}_i^T||_2 \tag{3.17}$$

where $\boldsymbol{b}_i^T$ is the i^{th} row of $\boldsymbol{B}$. Eq. (3.17) selects a sparse set of genetic variants through the ℓ_1 penalty across rows. Simultaneously, ℓ_2 penalty across columns preserves the representational similarity across basis vectors.

We note that even though we use similar representation schemes for the imaging and genetics data, they are different modalities with different biological interpretations. In contrast to fMRI data, the SNP data is more variable across subjects, and tends to be sparse. From a biological standpoint, it is also difficult to decode the downstream functional relationship between each pair of SNPs. Additionally, standard preprocessing for SNP data involves linkage disequillibrium (LD) correction, which removes much of the correlation between pairs of SNPs. Therefore, we have not made additional orthogonality assumptions. Instead, we use an $\ell_{2,1}$ norm to select a sparse set of relevant SNPs across the projections. From an optimization standpoint the SNP data has much higher dimensionality than the imaging data. An orthogonality constraint over the high dimensional SNP data would make the optimization unstable. Since our fMRI activation maps are based on a region parcellation, rather than voxel-wise analysis, we circumvent the issue.

3.2.3 Diagnosis Prediction

We use the subject-specific projection coefficients $\{\boldsymbol{x_n}\}_{n=1}^{N}$ to predict diagnosis. Mathematically, the diagnosis prediction is captured in a logistic regression framework, where we represented the class labels as, $y_n \approx \sigma(\boldsymbol{x}_n^T\boldsymbol{c})$. Here, $\sigma(\cdot)$ is the standard sigmoid function and $\boldsymbol{c} \in \boldsymbol{R}^{d\times 1}$ is the regression vector. We introduce an ℓ_2 penalty on both $\{\boldsymbol{c}, \boldsymbol{X}\}$ to make the optimization bounded and well posed.

Notice that we have coupled both the data modalities by tying the linear projection

coefficients $\boldsymbol{x}_n$ to the same latent space. These coefficients are used as a low-dimensional feature vector to predict diagnosis. This assumption allows us to extract discriminative patterns in $\boldsymbol{A}$ and $\boldsymbol{B}$ that are associated with each other. For example, if the d^{th} basis element is highly discriminative, then the corresponding coefficient of the logistic regression will be large. Thus, our joint formulation enables us to find discriminative patterns that simultaneously capture the data variations while being predictive of the disease. While our framework does not require the imaging and genetics data dimensions R and G to be equal, it assumes that both modalities can be represented by the same number of basis vectors.

3.2.4 Joint Optimization

We combine Eq. (3.14), Eq. (3.16), the logistic regression loss, and the regularization losses in a single joint objective function. This joint learning strategy guides groupwise discrimination informed by the two data modalities. Our joint objective function can be written as

$$\begin{aligned}\mathcal{J}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{X},\boldsymbol{c}) &= ||\boldsymbol{I}-\boldsymbol{A}\boldsymbol{X}||_F^2 + ||\boldsymbol{G}-\boldsymbol{B}\boldsymbol{X}||_F^2 \\ &- \sum_{n=1}^{N}\left(y_n \log\left(\sigma\left(\boldsymbol{x}_n^T\boldsymbol{c}\right)\right) + (1-y_n)\log\left(1-\sigma\left(\boldsymbol{x}_n^T\boldsymbol{c}\right)\right)\right) \\ &+ \frac{\lambda_1}{2}Tr\left(\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A}\right) + \lambda_2||\boldsymbol{B}||_{2,1} + \frac{\lambda_3}{2}||\boldsymbol{X}|_F^2 + \frac{\lambda_4}{2}||\boldsymbol{c}||_2^2 \\ &\text{s.t. } \boldsymbol{A}^T\boldsymbol{A} = \mathbb{I}\end{aligned} \tag{3.18}$$

We have concatenated the patient activations maps as $\boldsymbol{I} = [\boldsymbol{i}_1, \ldots, \boldsymbol{i}_N]$, the genetic variants as $\boldsymbol{G} = [\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n]$, and the projection coefficients as, $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]$. The first two terms in Eq. (3.18) capture the error associated with the imaging and genetic data representations, respectively. We minimize the Frobenius norms, $||\boldsymbol{I}-\boldsymbol{A}\boldsymbol{X}||_F^2$ and $||\boldsymbol{G}-\boldsymbol{B}\boldsymbol{X}||_F^2$ to estimate the unknown variables, $\{\boldsymbol{A},\boldsymbol{B},\boldsymbol{X}\}$. The third term captures the binary cross entropy loss for patient versus control prediction. The hyperparameters $\{\lambda_1,\lambda_2,\lambda_3,\lambda_4\}$ control the influence of the regularization penalties,

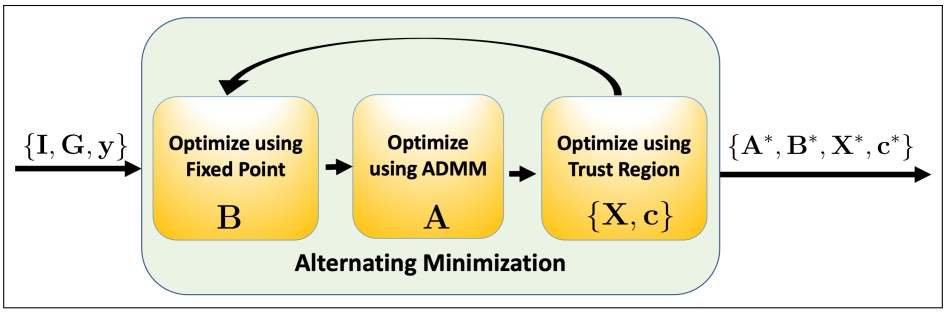


Figure 3-6. The alternating minimization approach to estimate the set of minimizers.

as described in the previous section.

We use an alternating minimization strategy to optimize the unknown variables $\{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{X}, \boldsymbol{c}\}$ in Eq. (3.18) from the data $\{\boldsymbol{i}_n, \boldsymbol{g}_n, y_n\}_{n=1}^N$. This procedure iteratively updates each unknown variable while holding the remaining variables constant. The alternating minimization approach is illustrated in Fig. 3-6.

Optimize $\boldsymbol{A}$ using ADMM: The orthonormality constraint in Eq. (3.18) renders the problem nonconvex with respect to the matrix $\boldsymbol{A}$. We circumvent this problem using Alternating Direction Method of Multipliers (ADMM). At a high level, ADMM introduces auxiliary variables to create a larger problem, such that each subproblem is easy to solve. In this case we introduce the matrices $\boldsymbol{C}$ and $\boldsymbol{D}$ into Eq. (3.18) to obtain the following modified objective for both them and the matrix $\boldsymbol{A}$:

$$\begin{aligned} \{\boldsymbol{A}^*, \boldsymbol{C}^*, \boldsymbol{D}^*\} &= \underset{\boldsymbol{A},\boldsymbol{C},\boldsymbol{D}}{\operatorname{argmin}} ||\boldsymbol{I} - \boldsymbol{C}\boldsymbol{X}||_F^2 + \frac{\lambda_1}{2} Tr(\boldsymbol{D}^T \boldsymbol{L} \boldsymbol{D}) \\ \text{s.t. } & \boldsymbol{A}^T\boldsymbol{A} = \mathbb{I}, \quad \boldsymbol{C} = \boldsymbol{A}, \quad \text{and} \quad \boldsymbol{D} = \boldsymbol{A} \end{aligned} \tag{3.19}$$

We find the closed form solution of $\{\boldsymbol{A}, \boldsymbol{C}, \boldsymbol{D}\}$ for the three subproblems by constructing an augmented Lagrangian to Eq. (3.19) defined as follows:

$$\begin{aligned} \mathcal{L}(\boldsymbol{A}, \boldsymbol{C}, \boldsymbol{D}, \boldsymbol{W}, \boldsymbol{Z}) &= ||\boldsymbol{I} - \boldsymbol{C}\boldsymbol{X}||_F^2 + \frac{\lambda_1}{2} Tr(\boldsymbol{D}^T \boldsymbol{L} \boldsymbol{D}) \\ &+ \frac{1}{\mu}||\boldsymbol{D} - \boldsymbol{A} + \boldsymbol{W}||_F^2 + \frac{1}{\mu}||\boldsymbol{C} - \boldsymbol{A} + \boldsymbol{Z}||_F^2 \\ \text{s.t. } & \boldsymbol{A}^T\boldsymbol{A} = \mathbb{I} \end{aligned} \tag{3.20}$$

where $\{\boldsymbol{W}, \boldsymbol{Z}\}$ are dual variables. We minimize Eq. (3.20) with respect to the primal variables $\{\boldsymbol{A}, \boldsymbol{C}, \boldsymbol{D}\}$ and maximize it with respect to the dual variables $\{\boldsymbol{W}, \boldsymbol{Z}\}$. We solve this problem in a iterative fashion. The pseudo code for our ADMM approach is shown in Algorithm 1. Each step is further detailed below.

(1) Closed form update for $\boldsymbol{A}$: We update $\boldsymbol{A}$ by minimizing corresponding terms of Eq. (3.20).

$$\begin{aligned} \boldsymbol{A}^{i+1} = \underset{\boldsymbol{A}}{\operatorname{argmin}} \ & \frac{1}{\mu}||\boldsymbol{D}^i - \boldsymbol{A} + \boldsymbol{W}^i||_F^2 + \frac{1}{\mu}||\boldsymbol{C}^i - \boldsymbol{A} + \boldsymbol{Z}^i||_F^2 \\ & \text{s.t. } \boldsymbol{A}^T\boldsymbol{A} = \mathbb{I} \end{aligned}$$

Given the other primal and dual variables, the update of $\boldsymbol{A}$ has a closed form analytical solution.

$$\boldsymbol{A} = \boldsymbol{U}\mathbb{I}_{M\times d}\boldsymbol{V}^T$$

where $\mathbb{I}_{M\times d}$ is a matrix of dimension $M \times d$ whose diagonal elements are 1, $\boldsymbol{U} \in \boldsymbol{R}^{M\times M}$, $\boldsymbol{V} \in \boldsymbol{R}^{d\times d}$ are two orthogonal matrices and $\Sigma \in \boldsymbol{R}^{M\times d}$ is a diagonal matrix satisfying the SVD factorization $\boldsymbol{D} + \boldsymbol{C} + \boldsymbol{W} + \boldsymbol{Z} = \boldsymbol{U}\Sigma\boldsymbol{V}^T$. The solution [153] is similar to Procustes problem [154].

(2) Closed form update for $\boldsymbol{D}$ and $\boldsymbol{C}$: The augmented Lagrangian is convex in each of the variables $\{\boldsymbol{C}, \boldsymbol{D}\}$ while keeping the other variables constant. Hence, we can simply set the gradient of the cost function with respect to $\boldsymbol{C}$ and $\boldsymbol{D}$, equal to

Algorithm 1 Iterative procedure for ADMM based on Augmented Lagrangian in Eq.(3.20)

Initialise $\boldsymbol{A}^0, \boldsymbol{C}^0, \boldsymbol{D}^0, \boldsymbol{W}^0, \boldsymbol{Z}^0$
for $i = 0$ to Convergence **do**
$\boldsymbol{A}^{i+1} = \boldsymbol{U}\mathbb{I}_{M\times d}\boldsymbol{V}^T$
$\boldsymbol{D}^{i+1} = \frac{2}{\mu}\left(\lambda_1\boldsymbol{L} + \frac{2}{\mu}\mathbb{I}\right)^{-1}(\boldsymbol{A} - \boldsymbol{W})$
$\boldsymbol{C}^{i+1} = \left(\boldsymbol{I}\boldsymbol{X}^T + \frac{2}{\mu}(\boldsymbol{A} - \boldsymbol{Z})\right)\left(\boldsymbol{X}\boldsymbol{X}^T + \frac{2}{\mu}\mathbb{I}\right)^{-1}$
$\boldsymbol{W}^{i+1} = \boldsymbol{W}^i + \boldsymbol{D}^{i+1} - \boldsymbol{A}^{i+1}$
$\boldsymbol{Z}^{i+1} = \boldsymbol{Z}^i + \boldsymbol{C}^{i+1} - \boldsymbol{A}^{i+1}$
end for

zero.

$$\boldsymbol{D} = \frac{2}{\mu}\left(\lambda_1 \boldsymbol{L} + \frac{2}{\mu}\mathbb{I}\right)^{-1}(\boldsymbol{A} - \boldsymbol{W}) \tag{3.21}$$

$$\boldsymbol{C} = \left(\boldsymbol{I}\boldsymbol{X}^T + \frac{2}{\mu}(\boldsymbol{A} - \boldsymbol{Z})\right)\left(\boldsymbol{X}\boldsymbol{X}^T + \frac{2}{\mu}\mathbb{I}\right)^{-1} \tag{3.22}$$

(3) Update for $\boldsymbol{W}$ and $\boldsymbol{Z}$: We maximize Eq. (3.20) with respect to $\boldsymbol{W}$ and $\boldsymbol{Z}$, by performing gradient ascent:

$$\boldsymbol{W}^{i+1} = \boldsymbol{W}^i + \boldsymbol{D}^{i+1} - \boldsymbol{A}^{i+1} \tag{3.23}$$

$$\boldsymbol{Z}^{i+1} = \boldsymbol{Z}^i + \boldsymbol{C}^{i+1} - \boldsymbol{A}^{i+1} \tag{3.24}$$

Maximizing the Lagrangian with respect to the dual variables ensures that the constraints are satisfied.

Optimize $\boldsymbol{B}$ using fixed point iteration: The matrix, $\boldsymbol{B}$ does not have a closed form solution due to the $\ell_{2,1}$ norm. However, it can be efficiently updated using a fixed point iteration method. In this method the ℓ_2 norm of each row $\boldsymbol{b}_i^T$ is kept fixed to its value $r_i^t = ||\boldsymbol{b}_i^T||_2$ from the previous iteration t. The matrix $\boldsymbol{B}$ is updated by minimizing the modified objective.

$$\mathcal{J}(\boldsymbol{B}) = ||\boldsymbol{G} - \boldsymbol{B}\boldsymbol{X}||_F^2 + \lambda_2 \sum_{i=1}^{G} \frac{||\boldsymbol{b}_i^T||_2^2}{2r_i^t} \tag{3.25}$$

Eq. (3.25) has closed form solution for each row, $\boldsymbol{b}_i^T$.

$$\boldsymbol{b}_i^T = \boldsymbol{g}_i^T \boldsymbol{X}^T \left(\boldsymbol{X}\boldsymbol{X}^T + \frac{\lambda_2}{2r_i^t}\mathbb{I}\right)^{-1}$$

where $\boldsymbol{g}_i^T$ is the i^{th} row of matrix, $\boldsymbol{G}$. Since each iteration has a closed form solution the algorithm converges very quickly. The proof of convergence can be found in [47].

Optimizing $\boldsymbol{X}$ and $\boldsymbol{c}$ using Trust Region Method: The cost function $\mathcal{J}(\cdot)$ in Eq. (3.18) is convex in each of the variables $\{\boldsymbol{X}, \boldsymbol{c}\}$ while keeping the others constant.

However, it does not have a closed form solution due to the logistic function $\sigma(\cdot)$. Therefore, using the unconstrained trust region method, we solve for $\boldsymbol{X}$ and $\boldsymbol{c}$ in an iterative fashion. At each iteration, the optimizer estimates a feasible direction and a step size to update the variable of interest by minimizing the following quadratic program:

$$\begin{aligned} \boldsymbol{s}_k = \text{argmin}_{\boldsymbol{s}} \quad & f(\boldsymbol{u}_k) + \nabla \boldsymbol{f}_k^T \boldsymbol{s} + \frac{1}{2} \boldsymbol{s}^T \boldsymbol{H}_k \boldsymbol{s} \\ & \text{subject to: } ||\boldsymbol{s}|| < \delta \end{aligned} \tag{3.26}$$

where $\nabla \boldsymbol{f}_k$ and $\boldsymbol{H}_k$ are the gradient and Hessian of $f(\boldsymbol{u})$ at $\boldsymbol{u}_k$. The update $\boldsymbol{u} \rightarrow \boldsymbol{u}_k + \boldsymbol{s}_k$ is taken such that $f(\boldsymbol{u}_k + \boldsymbol{s}_k) < f(\boldsymbol{u}_k)$. In our setting $f(\cdot)$ involves the terms of $\mathcal{J}(\cdot)$ that contain the variable under consideration. For example while minimizing over $\boldsymbol{X}$ we consider:

$$\begin{aligned} f(\boldsymbol{X}) = & ||\boldsymbol{I} - \boldsymbol{A}\boldsymbol{X}||_F^2 + ||\boldsymbol{G} - \boldsymbol{B}\boldsymbol{X}||_F^2 \\ & -\lambda_0 \sum_{n=1}^{N} \left(y_n \log \left(\sigma \left(\boldsymbol{x}_n^T \boldsymbol{c} \right) \right) + (1 - y_n) \log \left(1 - \sigma \left(\boldsymbol{x}_n^T \boldsymbol{c} \right) \right) \right) + \frac{\lambda_3}{2} ||\boldsymbol{X}||_F^2 \end{aligned}$$

We can solve for $\boldsymbol{c}$ in a similar fashion.

3.2.5 Prediction on unseen data

We use 10 fold cross-validation to evaluate the performance of our model. In each fold, we optimize the variables $\{\boldsymbol{A}^*, \boldsymbol{B}^*, \boldsymbol{c}^*\}$ over the training set and used them to evaluate the diagnostic classification on the test set. During testing, we remove the cross entropy term and use $\{\boldsymbol{i}_{test}, \boldsymbol{g}_{test}\}$ as input to obtain the projection coefficients, $\boldsymbol{x}_{test}$. We then use the same logistic expression $y_{test} = \sigma(\boldsymbol{x}_{test}^T \boldsymbol{c}^*)$ to predict the class labels.

3.2.6 Baseline Comparisons

We compare the predictive performance of our joint model with five baseline methods. For each case, we use the same 10 fold cross validation described above.

Support Vector Machine Classification: Support Vector Machines (SVM) construct a hyper-plane in a potentially high-dimensional and nonlinear feature space of the input data that maximally separates the two classes [76, 155]. Here, as a baseline we use a linear SVM based on the concatenated imaging and genetic features, $[\boldsymbol{i}_n^T, \boldsymbol{g}_n^T]^T$. Once again the output is the disease status y_n.

Random Forest Classification: Random Forest (RF) uses an ensemble of decision trees [156] to extract predictive features for classification. Each decision tree is constructed using a random subset of the input features. This double randomization provides robustness to overfitting over deterministic models [157]. Once again, the input to the RF will be the concatenated imaging and genetic features, $[\boldsymbol{i}_n^T, \boldsymbol{g}_n^T]^T$, and the output will be a patient versus control prediction, i.e., the label y_n.

Canonical Correlation Analysis + RF Classification Canonical Correlation Analysis (CCA) finds bivariate associations between the imaging and genetics data. These canonical coefficients are obtained by maximizing the following function:

$$\{\boldsymbol{u}_i^*, \boldsymbol{v}_i^*\} = \max_{\boldsymbol{u}_i, \boldsymbol{v}_i} \; corr(\boldsymbol{I}^T\boldsymbol{u}_i, \boldsymbol{G}^T\boldsymbol{v}_i)$$

where $\{\boldsymbol{u}_i, \boldsymbol{v}_i\}$ are the orthonormal basis vectors. These basis vectors form a low dimensional space where the two data modalities are maximally correlated. After obtaining the individual basis vectors, we stack them as matrices $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_R] \in \boldsymbol{R}^{R \times d}$ and $\boldsymbol{V} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_R] \in \boldsymbol{R}^{G \times R}$ to generate the imaging and genetics projection coefficients $\left[\boldsymbol{i}_n^T\boldsymbol{U}, \boldsymbol{g}_n^T\boldsymbol{V}\right]$, which are used as inputs to an RF classifier to predict y_n.

Parallel Independent Component Analysis + RF Classification: Parallel ICA (p-ICA) decomposes the imaging and genetics data into independent but interrelated networks. This is done by jointly maximizing multiple 'cost functions,' one of which specifies the independence among networks in each of the data sets and

another term that maximizes the correlation among pairs of networks across data sets. Formally,

$$\boldsymbol{I} = \boldsymbol{S}\boldsymbol{X} \quad \text{and} \quad \boldsymbol{G} = \boldsymbol{W}\boldsymbol{Z}$$

where $\boldsymbol{S}, \boldsymbol{W}$ are independent source matrices and the $\boldsymbol{X}, \boldsymbol{Z}$ are loading matrices whose cross-correlation is maximized. Since p-ICA is a purely generative model, we concatenate the loading matrices $[X_{test}, Z_{test}]$ and use it as the input feature vector for a random forest classifier.

During training, we apply p-ICA to just the training data to estimate the sources $\{\boldsymbol{S}_{train}, \boldsymbol{W}_{train}\}$. During testing, we use these sources to obtain the loading matrices for the test data via:

$$\boldsymbol{I}_{test} = \boldsymbol{S}_{train}\boldsymbol{X} \quad \text{and} \quad \boldsymbol{G}_{test} = \boldsymbol{W}_{train}\boldsymbol{Z}$$

Imaging Only Variant of Our Framework: We also consider a variant of our method that involves only the imaging terms. This baseline will help us quantify the improvement that we can achieve by incorporating the genetic data. As previously described we optimize the variables, $\{\boldsymbol{A}^*, \boldsymbol{c}^*\}$ on training set and use it for prediction $y_{test} = \sigma(\boldsymbol{x}_{test}^T \boldsymbol{c}^*)$ on test set.

Genetic Only Variant of Our Framework: Finally, we consider a variant of our method that involves only the genetic terms. The setup is similar to the above. Here we optimize the variables, $\{\boldsymbol{B}^*, \boldsymbol{c}^*\}$ on training set and use it for prediction $y_{test} = \sigma(\boldsymbol{x}_{test}^T \boldsymbol{c}^*)$ on test set.

As a sanity check, we verify whether our model can identify the unknown variables when the underlying assumptions of our objective function are met. Notice that our joint framework has an equivalent Bayesian model, as illustrated in Fig. 3-7. Namely, for each patient n, the process starts by sampling a latent projection $\boldsymbol{x}_n$ from a

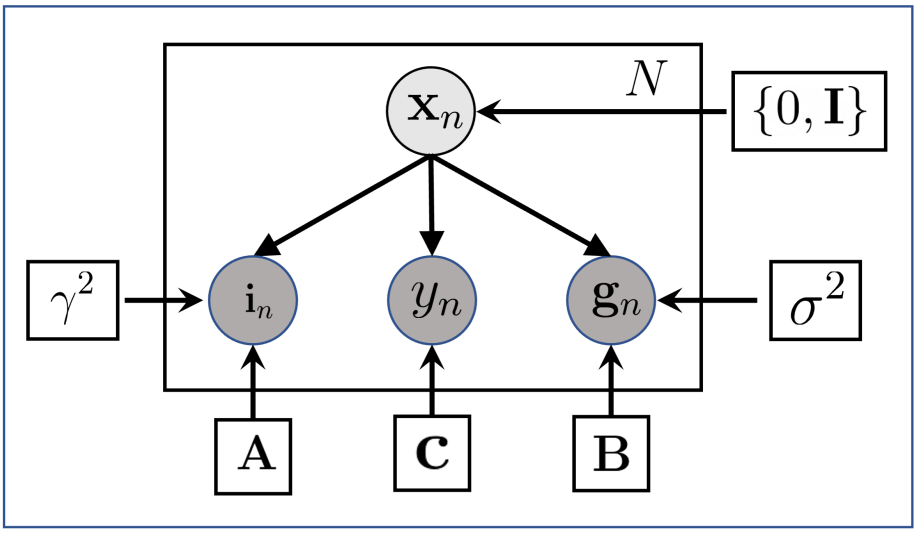


Figure 3-7. The Bayesian framework for our simulation study.

zero-mean Gaussian, corresponding to ℓ_2 regularization in Eq. (3.18). From here, the imaging data $\boldsymbol{i}_n$ is generated as the noisy observation of the linear combination of the orthonormal basis matrix, $\boldsymbol{A}$:

$$\boldsymbol{i}_n = \boldsymbol{A}\boldsymbol{x}_n + \epsilon_n$$

where $\epsilon_n \sim \mathcal{N}(0, \sigma^2\mathbb{I})$ with effective noise level σ. We generate the deterministic orthonormal matrix $\boldsymbol{A}$ as a QR decomposition of random Gaussian matrix, $\tilde{\boldsymbol{A}}$, with each column sampled from $\mathcal{N}(\boldsymbol{\mu_A}, 0.01\mathbb{I})$ with sparse binary mean $\boldsymbol{\mu_A} \in [0,1]^M$. In our analysis we explore the task based fMRI data which has an underlying assumption that a sparse set regions involve in the task show significant activity compared to the rest of the brain. This process approximates the Laplacian constraints enforced on $\boldsymbol{A}$ in Eq. (3.15).

The procedure to generate the genetics vector $\boldsymbol{g}_n$ is similar but based on the projection matrix $\boldsymbol{B}$:

$$\boldsymbol{g}_n = \boldsymbol{B}\boldsymbol{x}_n + \nu_n$$

where $\nu_n \sim \mathcal{N}(0, \gamma^2\boldsymbol{I})$. The columns $\boldsymbol{b}_j$ from the matrix $\boldsymbol{B}$ is sampled as a random multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu_B}, 0.01\boldsymbol{I})$ with a sparse mean vector $\boldsymbol{\mu}_B \in [0,1,2]^G$. This

choice of mean mimics the real-life scenario where SNP values are generally given as $[0, 1, 2]$ based on the variation of the two alleles. Additionally, Gaussian sampling across the columns of $\boldsymbol{B}$ mimic the ℓ_{21} regularization as shown in Eq. (3.17).

Finally, the discriminative term is obtained via

$$y_n = \sigma\left(\boldsymbol{c}^T \boldsymbol{x}_n\right)$$

where $\boldsymbol{c}_n$ is a zero-mean Gaussian. $\boldsymbol{c} \sim \mathcal{N}(0, \boldsymbol{I})$.

We evaluate the performance of our model and optimization for different noise levels on the imaging and genetic representations. The performance metric is the accuracy of our selected features, as quantified by the Jaccard overlap between the non-zero locations of the original bases matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ and the estimated bases matrices $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$, respectively.

In our synthetic experiment the dimensionality of the data is similar to our real data, i.e., $\boldsymbol{i}_n \in \boldsymbol{R}^{246\times 1}, \boldsymbol{g}_n \in \boldsymbol{R}^{1242\times 1}$, and number of subjects $N = 106$. Empirically, this allows us to evaluate whether our generative-predictive framework can identify the set of ground-truth biomarkers in both $\boldsymbol{A}$, and $\boldsymbol{B}$. As our detection strategy we take the absolute sum of the columns of estimated matrices $\{\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}\}$, and identify the top $\{n_g, n_i\}$ regions, where n_i is the number of true non-zero locations in $\boldsymbol{\mu_A}$, and n_g is the number of true non-zero locations in $\boldsymbol{\mu_B}$. Finally, we find the overlap between the estimated locations with the true locations which is shown in Fig. 3-7. A high Jaccard index indicates that our model can correctly find the non-zeros location in $\boldsymbol{\mu_A}$ and $\boldsymbol{\mu_B}$.

Fig. 3-8 shows the performance of our model at varying noise level as governed by σ and γ. As seen, in one case we fix the noise for $\boldsymbol{i}_n$ at $\sigma = 0.2$ and sweep over γ, while in the other case we fix the noise for $\boldsymbol{g}_n$ at $\gamma = 0.4$ and sweep over σ. We allowed a wide range for our noise parameter $\{\sigma^2 \in [0.01, 1], \gamma^2 \in [0.01, 4]\}$ to check the model's robustness against random noise. We observe that $\gamma^2 = 0.16$ and $\sigma^2 = 0.04$

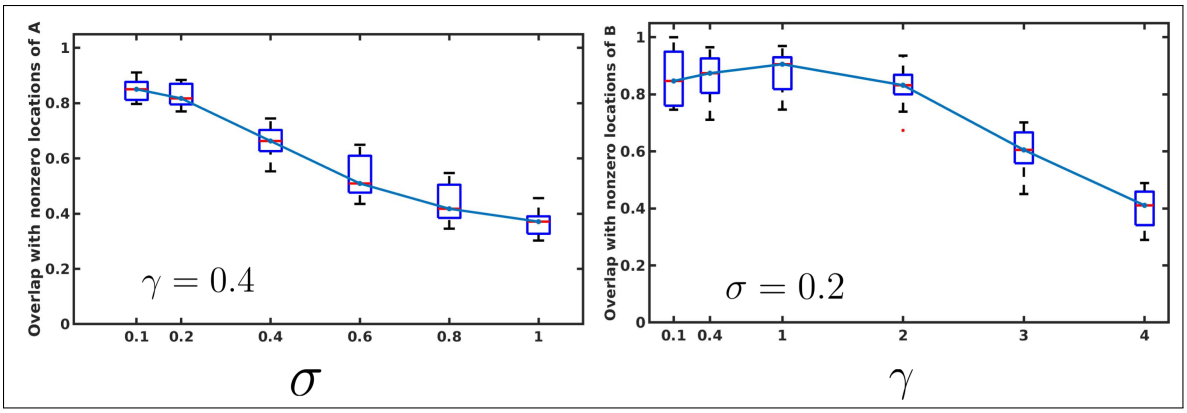


Figure 3-8. The overlap between our estimated bases with the true sparse bases $\boldsymbol{A}$ and $\boldsymbol{B}$ at varying level of noise. Compared to the numerical range of the feature vectors we have sweeped over four standard deviation for the noise.

are the variability in our real-world fMRI and genetic datasets, which lies well within the stable region of our model as shown in Fig. 3-8. With the increase in noise the amount of overlap as quantified by the Jaccard Index decreases. However, the model can extract relevant features with high accuracy over a wide range of input noise. This shows that the optimization strategy is robust and is capable to extract the informative features even when we are outside noise regime of our real-world data.

3.2.7 Experiments

3.2.7.1 Real-World Study of Schizophrenia

We validate our framework on task fMRI and genetic data acquired at two different sites on two different study populations. The first dataset was provided by researchers at the the Lieber Institute for Brain Devel-opment (LIBD) in Baltimore, MD, USA. The second dataset was acquired at the University of Bari Aldo Moro, Italy. The data collection procedures and pre-processing were consistent across sites.

Neuroimaging Data: Our datasets include two fMRI paradigms that have been used to study schizophrenia [3, 4]. The first paradigm is a block design working memory task (N-Back), and the second is a block design declarative memory task (SDMT),

which involves incidental encoding of complex aversive visual scenes. The details of the imaging modalities and the preprocessing can be found in Section 2.4.1 Our inputs to the model are region-wise averages of the contrast values across all voxels in each parcel of the brain. Further details for generating the contrast maps can be found in Section 2.4.1.

Table 3-II reports the subject numbers for each paradigm and site. The groups were matched on age, IQ (WRAT score), years of education and in the case of N-Back, the percent correct response for the 2-Back task. Table 3-III shows the demographic variability of all the subjects used in our analysis. Here we note that the education data for BARI is not available to us and hence is not used in the analysis.

Genetic Data: Genotyping was done using variate Illumina Bead Chips including 510K/ 610K/660K/2.5M. After the initial preprocessing steps described in Section 2.4.1 we obtain 102K linkage disequilibrium independent SNPs. Given the small sample sizes in Table 3-II ($N \approx 100$ for each dataset), we subselect a set of SNPs whose p-value for disease association is $p < 10^{-4}$, as identified by the PGC-Consortium GWAS analysis. In total, this threshold yields 1242 linkage disequilibrium independent SNPs, which balances the representativeness of the genetic data with robustness of our optimization procedure. We use the same reduced set of SNPs for all cross validation folds. This reduced set was obtained from a larger genetics study of 36,989 schizophrenia patients

Table 3-II. The number of subjects present from each experimental paradigms from the two institutions

Institution	fMRI Paradigms			
	N-Back		SDMT	
	Cases	Controls	Cases	Controls
LIBD	53	53	46	47
BARI	43	54	–	–

Table 3-III. The demographic of all the subjects used for our analysis. The education data for BARI is not available and hence is not included in our analysis.

Demographic	LIBD		BARI
	N-back	SDMT	N-back
Sex (M/F)	65/41	57/36	74/23
Age (years)	30 ± 10	33 ± 9	30 ± 9
Education (years)	15 ± 2	15 ± 3	–
IQ	105 ± 10	105 ± 8	107 ± 8

and 113,075 neurotypical controls run by the PGC Consortium. Further details about this study can be found in [17]. Hence, our feature selection procedure does not confound the training and testing data in our analysis.

3.2.7.2 Evaluation Strategy

We quantify the performance of our method and all the baselines in terms of Accuracy (Acc), sensitivity (Sens) and Specificity (Spec). Accuracy is a measure of correct detection of the class labels. Sensitivity is the ratio of the true positives among all predicted positives, whereas specificity is the ratio of the true negatives among all predicted negatives. Formally,

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$
$$\text{Sens} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$
$$\text{Spec} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

where TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative.

3.2.7.3 Hyperparameter Selection

Our generative-discriminative framework contains the following hyperparameters: $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ to control the contributions of the regularization terms in the optimiza-

Table 3-IV. Classification performance of each method. We abbreviated Sensitivity to SENS, Specificity to SPEC, Accuracy to ACC, and Area Under Curve to AUC.

Method	LIBD								BARI			
	N-Back				SDMT				N-Back			
	SPEC	SENS	ACC	AUC	SENS	SPEC	ACC	AUC	SENS	SPEC	ACC	AUC
SVM	0.53 ± 0.03	0.44 ± 0.03	0.49 ± 0.02	0.35 ± 0.04	0.60 ± 0.06	0.57 ± 0.03	0.57 ± 0.02	0.56 ± 0.03	0.73 ± 0.04	0.49 ± 0.05	0.63 ± 0.04	0.70 ± 0.02
RF	0.55 ± 0.05	0.52 ± 0.03	0.53 ± 0.03	0.54 ± 0.03	0.64 ± 0.03	0.57 ± 0.04	0.61 ± 0.02	0.65 ± 0.03	$\mathbf{0.88 \pm 0.01}$	0.49 ± 0.03	0.70 ± 0.01	$\mathbf{0.84 \pm 0.01}$
CCA + RF	0.49 ± 0.10	0.48 ± 0.09	0.49 ± 0.08	0.52 ± 0.08	0.53 ± 0.05	0.48 ± 0.09	0.51 ± 0.05	0.51 ± 0.05	0.75 ± 0.06	0.31 ± 0.05	0.56 ± 0.05	0.56 ± 0.05
p-ICA + RF	0.49 ± 0.09	0.45 ± 0.08	0.47 ± 0.04	0.47 ± 0.05	0.53 ± 0.10	0.41 ± 0.10	0.47 ± 0.08	0.45 ± 0.08	0.75 ± 0.05	0.65 ± 0.05	0.71 ± 0.03	0.76 ± 0.02
Our Method (Imaging Only)	0.55 ± 0.04	$\mathbf{0.62 \pm 0.03}$	0.58 ± 0.02	0.63 ± 0.02	0.63 ± 0.04	0.59 ± 0.03	0.61 ± 0.03	0.67 ± 0.02	0.67 ± 0.04	0.80 ± 0.05	0.73 ± 0.03	0.79 ± 0.02
Our Method (Genetic Only)	0.44 ± 0.03	0.50 ± 0.05	0.47 ± 0.03	0.45 ± 0.02	0.45 ± 0.08	0.45 ± 0.07	0.45 ± 0.04	0.43 ± 0.03	0.65 ± 0.02	0.66 ± 0.02	0.66 ± 0.02	0.69 ± 0.01
Our Method (Imaging + Genetics)	$\mathbf{0.56 \pm 0.04}$	0.60 ± 0.02	$\mathbf{0.58 \pm 0.02}$	$\mathbf{0.63 \pm 0.02}$	$\mathbf{0.64 \pm 0.04}$	$\mathbf{0.61 \pm 0.04}$	$\mathbf{0.63 \pm 0.03}$	$\mathbf{0.69 \pm 0.02}$	0.66 ± 0.04	$\mathbf{0.83 \pm 0.02}$	$\mathbf{0.73 \pm 0.02}$	0.81 ± 0.01

tion, and d specifies the latent space dimensionality. To combat overfitting, our strategy is to optimize these hyperparameters based on the LIBD N-back dataset and use the same values for the LIBD SDMT and Bari N-back analyses. We sweep the regularizers $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ over two orders of magnitude and the latent space dimension from $d = 5, \ldots, 11$. In our analysis we have observed that the hyperparameter λ_3, and λ_4 are stable over a range of $[0.005 - 5]$, so we fix them at $\lambda_3 = 1, \lambda_4 = 1$. The sensitivity plot is shown in Fig. 3-9. Based on our experiments we fix the feature dimension (d), the imaging regularizer (λ_1), the genetic regularizer (λ_2), to $\{d = 7, \lambda_1 = 1, \lambda_2 = 10\}$. We have used the same hyperparameter setting for all the variants of our model for both the SDMT (LIBD), and the N-Back (BARI) datasets. The sensitivity plots, Fig. 3-9 of $\{\lambda_1, \lambda_2\}$ also show stability over a wide range, but they are closely tied with the biomarker detection regime. So, for future applications on a standalone dataset we advise the researcher to fine tune them using some validation techniques, like cross-validation.

As our optimization is non-convex, we use an informed initialization strategy to satisfy the variable constraints while not biasing the solution path. To this end, we initialize the imaging basis matrix $\boldsymbol{A}$ as a QR decomposition of random Gaussian matrix. The QR decomposition satisfies the orthogonality constrain over columns of $\boldsymbol{A}$ in our framework. We initialize $\boldsymbol{B}, \boldsymbol{X}, \boldsymbol{c}$ such that each element is sampled from a uniform distribution between 0 and 1. We note that since our optimization converges to a local optimum, different initialization may produce different final solutions. However,

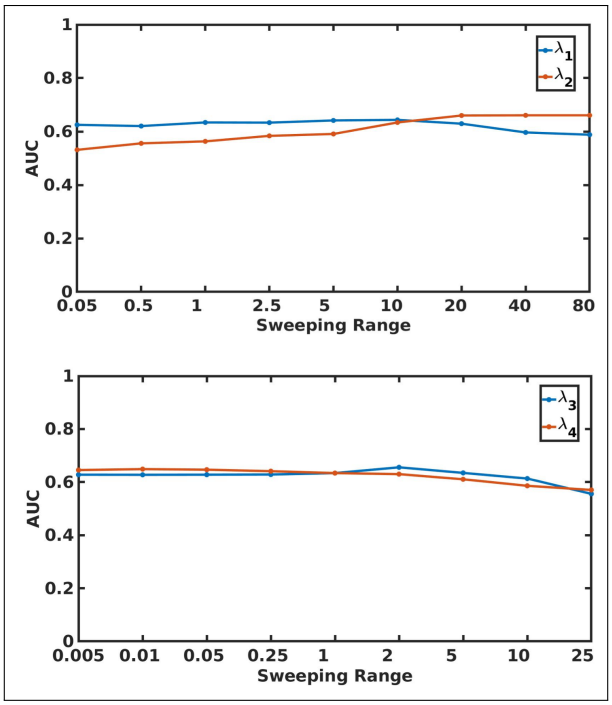


Figure 3-9. The change in AUC for different ranges of the hyperparameters $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$. We sweep one hyperparameter while keeping the others constant at their stable value. This analysis has been done on the N-back dataset.

Table 3-IV suggests that classification performance remains stable across different cross validation folds, each of which has different initialization.

Similar to our method, we optimized the hyperparameters for the baseline methods on the LIBD N-back data and used these settings for the two analyses. For RF classification we swept over the number and depth of the trees. We controlled the depth of the tree depth by setting the minimum number of observations per leaf node. These parameter sweeps were repeated for CCA+RF and pICA+RF. Based on these sweeps, we fixed $\{\text{No. trees} = 2000, MinleafSize = 5\}$ for the standard RF classification $\{\text{No. trees} = 8000, MinleafSize = 10\}$ for CCA+RF and $\{\text{No. trees} = 9000, MinleafSize = 1\}$ for pICA+RF. Additionally, for the implementation of pICA we use the standard hyperparameter setting as explained in the Fusion ICA (FIT) [158] toolbox. The linear SVM includes one hyperparameter, BoxConstraint which controls the outlier penalty. Our final settings was $\{BoxConstraint = 1\}$.

3.2.7.4 Class Prediction

Table 3-IV reports the classification performance of all methods on the three fMRI datasets. We can see that the machine learning baselines perform poorly compared to all the three variants of our model. This result suggests that our coupled generative-discriminative framework is able to extract meaningful features from the data that capture group level differences. Moreover, we observe that our framework achieves the best cross-site performance between the LIBD and Bari cohorts. This performance gain demonstrates that our model is agnostic to the choice of hyperparameters and our optimization procedure is robust enough to handle noises associated with different sample sets. Though all the variants of our model achieve good classification accuracy compared to the baselines, the performance gain obtained by integrating both the imaging and genetic data modalities is apparent across all experiments particularly with regards to accuracy and AUC. This performance gain can also be attributed tot the fact that our method can find patterns from the imaging and genetics data that are highly predictive of the disease.

3.2.7.5 Predictive Biomarkers

In this section, we aim to identify and interpret the underlying biology of potential imaging-genetics biomarkers. We emphasize that our analyses and conclusions are exploratory, and for this reason, we focus on just the LIBD data.

We use the patient specific scores $\boldsymbol{x}_n$ for disease classification and data reduction. They contain information both about the imaging data and the genetic data. These vectors are d dimensional where each dimension can be associated with a column of $\boldsymbol{A}$ and a column of $\boldsymbol{B}$. In order to identify which columns of $\boldsymbol{A}$ and $\boldsymbol{B}$ contain most discriminative patterns we perform a KS test [159] between $\boldsymbol{x}^d_{disease}$($d$-th feature of the disease group) and $\boldsymbol{x}^d_{control}$($d$-th feature of the control group). A low p-value along a specific dimension d would mean that the distribution of that feature is not

Algorithm 2 Subsampling strategy for identifying predictive biomarkers

1: Train the model on the complete dataset.
2: Perform KS test on loading vectors x_n^d (subject n and basis d) between patients and controls.
3: Identify the significant imaging and genetic components $\{(\boldsymbol{a}_1^*, \boldsymbol{b}_1^*), \ldots, (\boldsymbol{a}_c^*, \boldsymbol{b}_c^*)\}$ based on a KS test.
4: **for** $i = 1$ to 50 random subsamples.
5: Randomly sample 90% of the data,
6: Train the model on the sampled dataset.
7: Perform KS test on loading vectors $\hat{x}_n^d$ between patients and controls .
8: Identify the significant imaging and genetic components based on the KS test.
9: Match the estimated basis vectors as identified by the KS test with the reference ectors as shown in Eq. (3.27).
10: Normalize the matched vectors to z scores.
11: **end for**
12: Find the order statistics as shown in Eq. (3.28).
13: **Predictive Biomarkers** $\leftarrow$ Find the locations (rows) of $\mathcal{I}^c$ where $|\mathcal{I}^C(r)| \geq 1.5$.

equal between patients and controls. The KS test gives us d p-values for all the d dimensions of $\boldsymbol{x}_n$. Finally, we select the significant components with FDR corrected $p < 0.01$. Here, we note that this test allows us to prune out regions and SNPs that do not track with diagnosis, the interpretation should be viewed as an exploratory analysis, and further work is required to verify clinical relevance.

We perform a subsampling experiment to quantify the reproducibility of these bases. Namely, we train the model over the complete dataset to identify the reference basis vectors indicated by $\{(\boldsymbol{a}_1^*, \boldsymbol{b}_1^*), \ldots, (\boldsymbol{a}_d^*, \boldsymbol{b}_d^*)\}$. Our subsampling strategy relies on random sampling of data without replacement. At a high level patterns that are consistent with the reference vectors $\{(\boldsymbol{a}_1^*, \boldsymbol{b}_1^*), \ldots, (\boldsymbol{a}_c^*, \boldsymbol{b}_c^*)\}$ across all the trials are more likely to generalize beyond the present experimental setup. The subsampling strategy to identify the biomarkers is shown in Algorithm 2.

Our subsampling procedure uses 90% random sampling without replacement. For each trial, we perform a KS test to identify the significant basis vectors estimated from the sampled data. We then perform a one-to-one mapping between the reference

vectors and the estimated vectors by maximizing the correlation between them. The correlation between the i^{th} reference vector and the j^{th} estimated vector is defined as

$$C_{ij} = \frac{\left(|\boldsymbol{a}_i^*| - \overline{|\boldsymbol{a}_i^*|}\right)^T \left(|\hat{\boldsymbol{a}}_j| - \overline{|\hat{\boldsymbol{a}}_j|}\right)}{\left\|\left(\left(|\boldsymbol{a}_i^*| - \overline{|\boldsymbol{a}_i^*|}\right)\right\|_2^2 \left\|\left(\hat{\boldsymbol{a}}_j - \overline{|\hat{\boldsymbol{a}}_j|}\right)\right\|_2^2\right)^{\frac{1}{2}}} \tag{3.27}$$

where $\boldsymbol{a}_i^*$ is the i^{th} reference basis vector, $\hat{\boldsymbol{a}}_j$ is the j^{th} estimated basis vector and $\bar{(\cdot)}$ denotes the mean of features along the vector. We take an absolute value because our model is invariant to a change of sign of the bases. This correlation analysis allows us to match the set of basis vectors obtained from the sampled data that are strongly correlated with the reference vectors.

Finally, we identify the consistent set of biomarkers across the subsamples via the element-wise median z-score of the basis vectors across the 50 trials.

$$\mathcal{I}_j^a(r) = median\left(\hat{\boldsymbol{a}}_j^1(r), \ldots, \hat{\boldsymbol{a}}_j^{50}(r)\right) \tag{3.28}$$

where $\mathcal{I}_j^a(r)$ quantifies the importance of region r across the subsamples, and $\hat{\boldsymbol{a}}_j^k(r)$ is the estimated basis obtained from the k^{th} subsample. A high value in $\mathcal{I}$ means that the region is consistently selected for diagnosis of a subject during subsampling. We perform a meta analysis on the set of biomarkers thresholded at $|\mathcal{I}_j^c(r)| > 1.5$ to show their relevance in the context of schizophrenia.

As a second stage of our exploration study we perform a correlation analysis between the identified biomarkers and a generalized cognitive score derived from a battery of standard cognitive assessment which were performed on the patients and controls subjects. The generalized cognitive score, or "g" score [160], is composite measure of general cognitive ability based on six broad cognitive domains: verbal memory, n-back, visual memory, processing speed, card sorting and digit span. Here, we consider the imaging components that show significant group level differences between cases and controls, as identified by the KS test. In order to find the association between these components and cognition, we calculate the Pearson's correlation between the

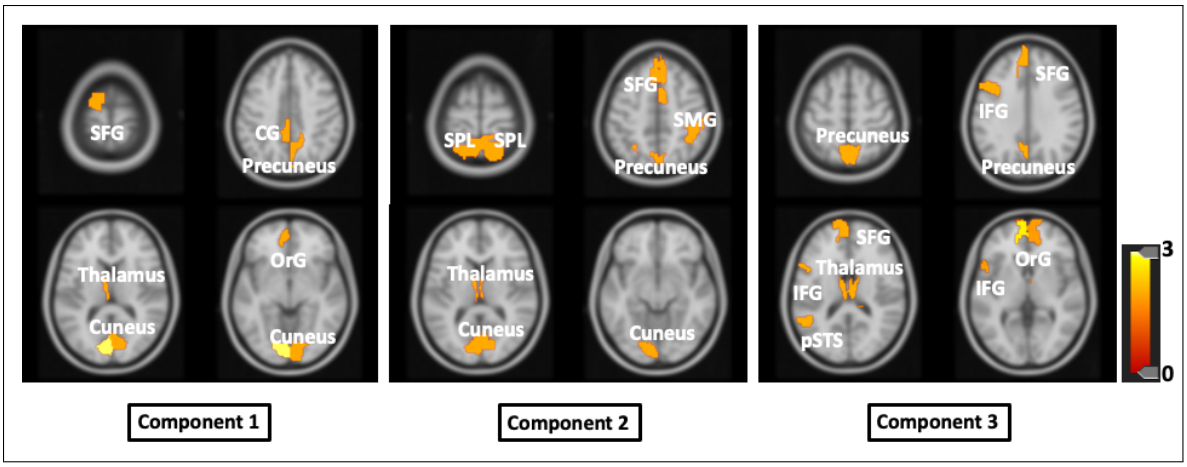


Figure 3-10. A detailed description of all the brain regions identifies by our model for N-Back data.

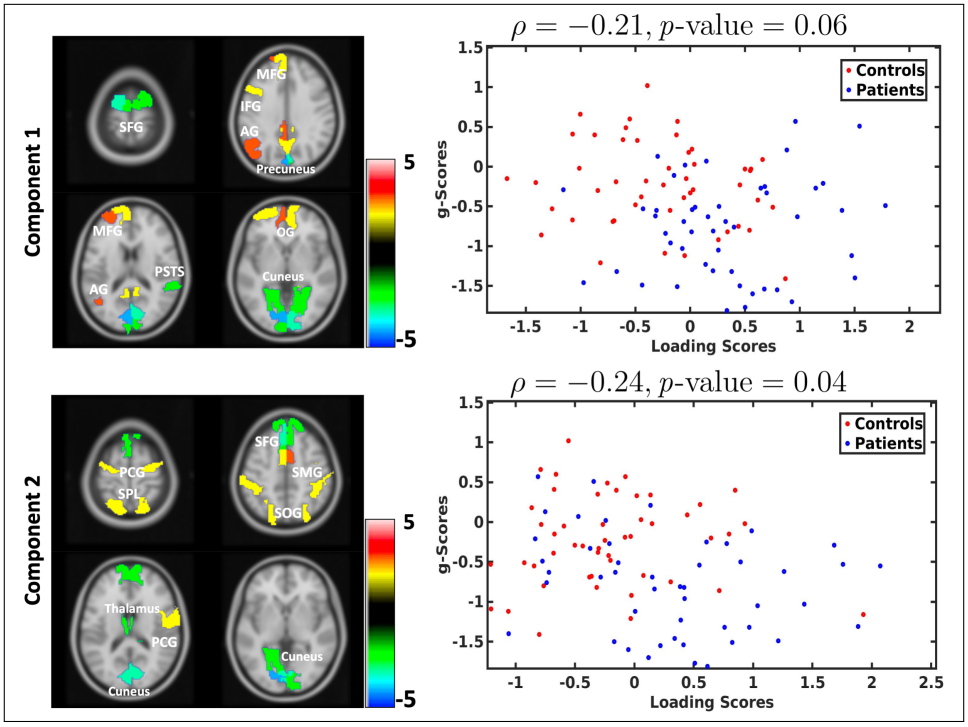


Figure 3-11. Left: The identified set of biomarkers that have shown strong association with the generalized cognitive scores for the Nback dataset. **Right:** The scatter plot between the cognitive scores and the subject specific loading scores for the Nback dataset. The correlation between the loading scores and the "g" scores are identified by ρ, and level of significance is captured by the FDR corrected p-value.

patient specific scores $\{x_n^d\}_{n=1}^N$ and the corresponding patient g-score. Each dimension d of the patient specific scores $\boldsymbol{x}_n^d$ is associated with the basis vector which capture

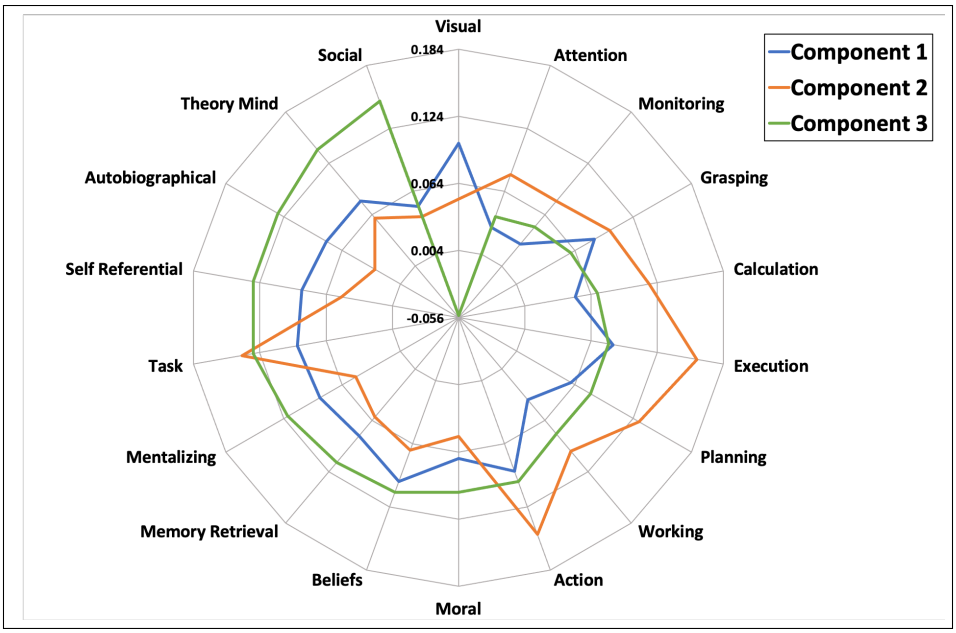


Figure 3-12. The correlation value of each brain component identified in the N-Back dataset with the higher order brain states based on the Neurosynth database.

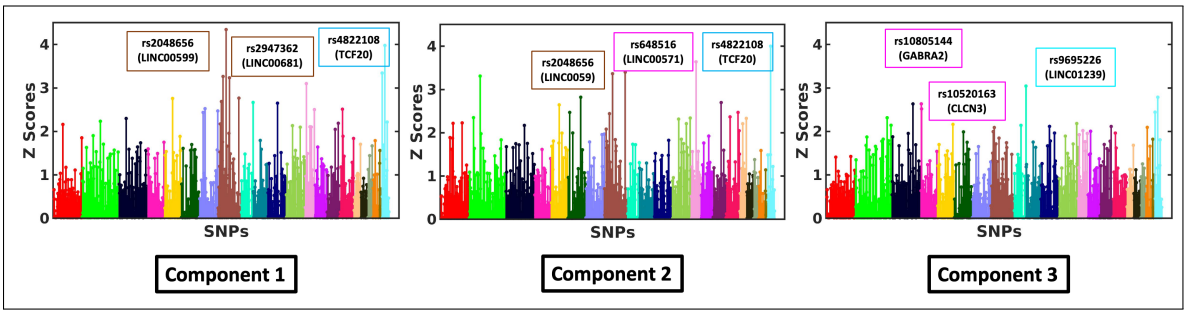


Figure 3-13. The importance map of all the SNP and their overlapping genes across all the subsamples for N-Back data.

group level difference. So, as a next step we plot the basis vectors in the brain. This analysis explores the relationship between the cognitive scores and the identifiesd set of biomarkers.

Analysis of the N-Back Biomarkers: For the N-Back data our initial KS test reveals three components that are significantly different between cases and controls with $p < 0.0021$, $p < 0.0024$, and $p < 0.009$, respectively (FDR corrected). We use these components as reference for our subsampling experiments.

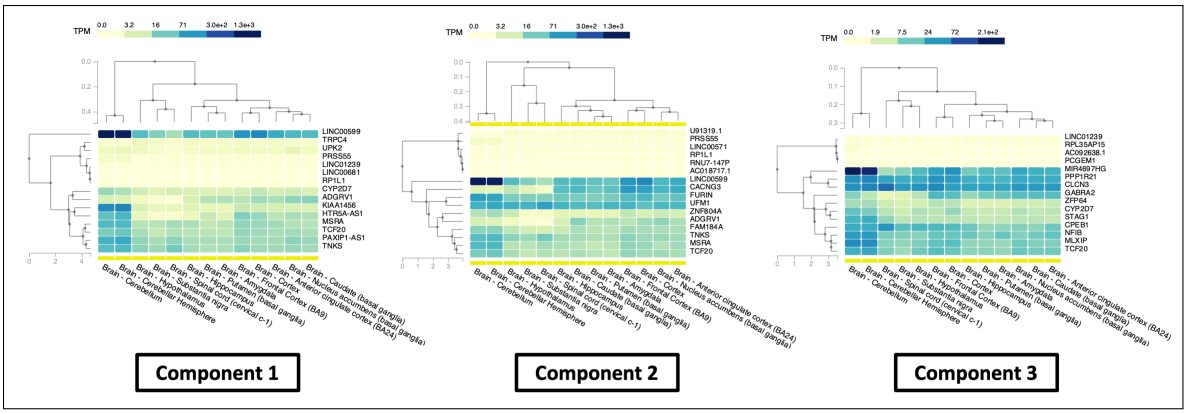


Figure 3-14. The gene expression pattern of the top genes identified from the N-Back task based on the GTEx database.

A detailed diagram of all the brain regions across the three different components along with their corresponding annotations are shown in Fig. 3-10. In **Component 1** and **Component 3** we can see regions that include superior frontal gyrus (SFG), and inferior frontal gyrus (IFG), which are know to subserve executive cognition [4]. Moreover, in **Component 2** we can see regions from the default mode network (DMN) which is also implicated in schizophrenia [161]. We further use Neurosynth [162] to decode the higher order brain states of the the biomarkers aggregated across all the three components. Fig. 3-12 shows the Neurosynth terms that are strongly correlated with our biomarkers. We note that the terms for **Component 2** involve regions used for planning and execution of a task, whereas **Component 1** and **Component 3** involve regions associated with memory retrieval and the default mode. These results show that the model can extract potential imaging biomarkers that contain informative patterns of the data.

Fig. 3-13 illustrates the component-wise SNP contributions, whose z-values are overlapped with a gene. We use the SNPnexus [163] web interface to find the set of overlapping genes or the nearest upstream or downstream gene for each SNP. As parallel to Neurosynth analysis, we perform a gene expression based analysis [164] over the 20 overlapping (or nearest) genes of the top SNPs identified from each of the

three components. This exploratory analysis may help us to understand the *cis*-effects of the SNPs and how they alter the functionalities of genes expressed in different tissues of the brain. Fig. 3-14 shows the gene expression pattern of each gene across different brain tissues. As seen, two of the most expressed genes that appeared in multiple components are *TCF20* and *LINC00599* which are known to be associated with schizophrenia [17] and neuroticism [165].

The scatter plots in Fig. 3-11 show association between each of the Nback components, as selected via the KS test, and the "g" scores. Among the three Nback components the first two components show significant association while the third one was not significantly correlated. Additionally, in Fig. 3-11 we plot the identified set of biomarkers associated with the loading scores as separate brain plots. Both Nback components show that the shared variance between brain regions of the frontoparietal network, such as the inferior frontal gyrus and angular gyrus, is anticorrelated with components of the default mode network such as the cuneus and the medial prefrontal cortex. Positive loading scores were associated with lower g, suggesting that, across individuals, high loading in these two components covaried with greater frontoparietal network activation. At the Nback load we considered, greater frontoparietal activity has been reported in patients with schizophrenia, when performance is equated between groups [4].

Analysis of SDMT Biomarkers: Our KS test on the SDMT data revealed two significant components with $p < 0.0004$ and $p < 0.0004$ (FDR corrected) between patients and controls. Once again, these components served as the reference vectors in our subsamples.

Fig. 3-15 shows the set of brain regions identified by our method along different axial views. The SDMT biomarkers implicate the parahippocampal (P-HIP), superior frontal regions (SFG) along with precuneus, fusifor gyrus and cuneus, all of which are

affected in schizophrenia [166, 167]. We also observe regions from the default mode network that control memory encoding in schizophrenia. Fig. 3-17 reports the results of our Neurosynth meta-analysis. Notice that our biomarkers include regions involve in memory [3] and facial recognition, both of which are impaired in schizophrenia. Taken together, these results highlight the promise of our model for neural biomarker discovery.

Fig. 3-18 shows the SNPs, and their overlapping (or nearest) genes as found from the SNP-nexus web interface. Again, we perform a gene expression phylogeny [164] over the identified set of genes. Fig. 3-19 captures the expression level of the most significant genes implicated by the identified set of SNPs. Here, *LINC00599* shows high expression levels in brain and are also known to be associated with schizophrenia [168] and neuroticism [165].

The association with cognitive scores for the SDMT data has been done by following the same strategy of finding correlation between the patient specific scores x_n^d and the "g" scores. Likewise, Fig. 3-16 shows the identified set of biomarkers associated

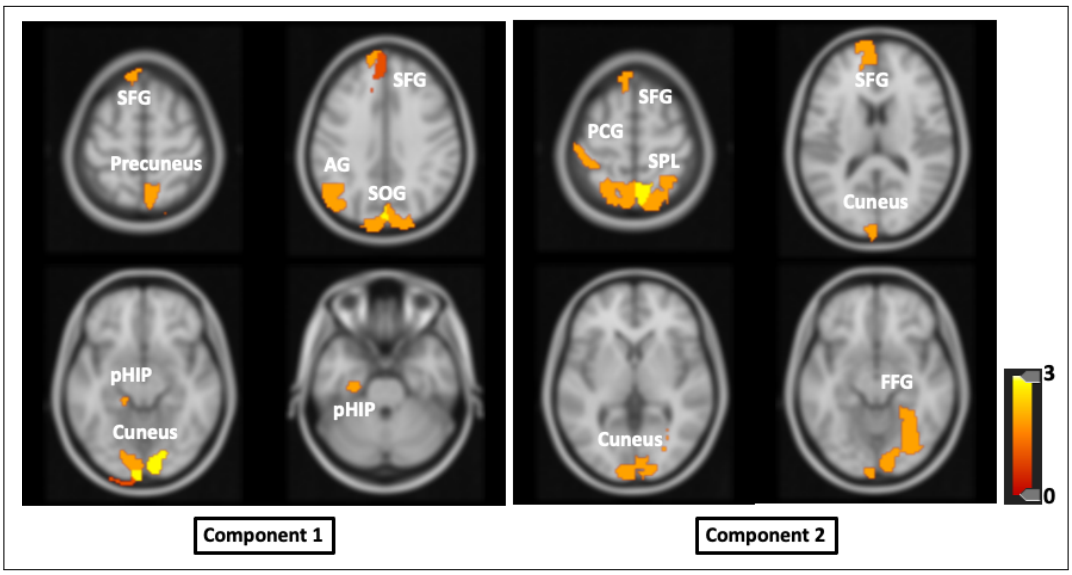


Figure 3-15. A detailed description of all the brain regions identifies by our model for SDMT data.

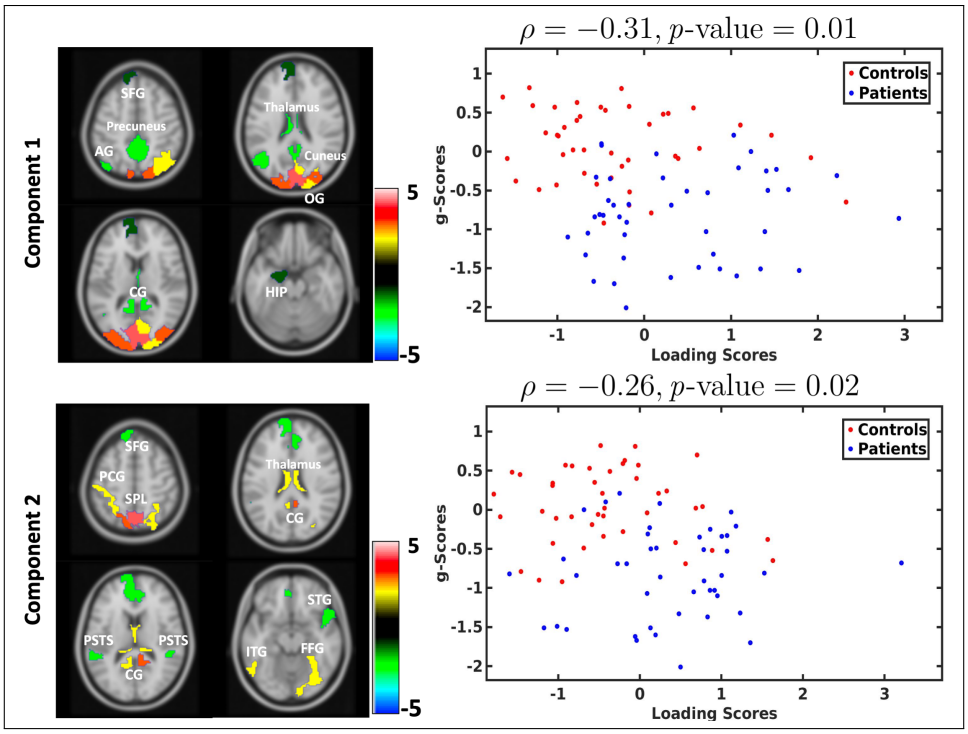


Figure 3-16. Left: The identified set of biomarkers that have shown strong association with the generalized cognitive scores for the SDMT dataset. **Right:** The scatter plot between the cognitive scores and the subject specific loading scores for the SDMT dataset.

with the loading scores along with the scatter plot that shows significant association between the loading scores and the cognitive "g" scores. Both SDMT components tapped into the episodic memory network, including the hippocampus, the medial and dorsolateral prefrontal cortex, posterior cingulate and parietal regions, mostly negatively correlated, with some heterogeneity between components. Considering the correlation with "g", negative loadings suggest that the best cognitive performers showed a greater involvement of the episodic memory network during the task, which is consistent with previous reports on these data [30]. These findings show that the model can be used to explore potential biomarkers and their interactions in a multivariate framework.

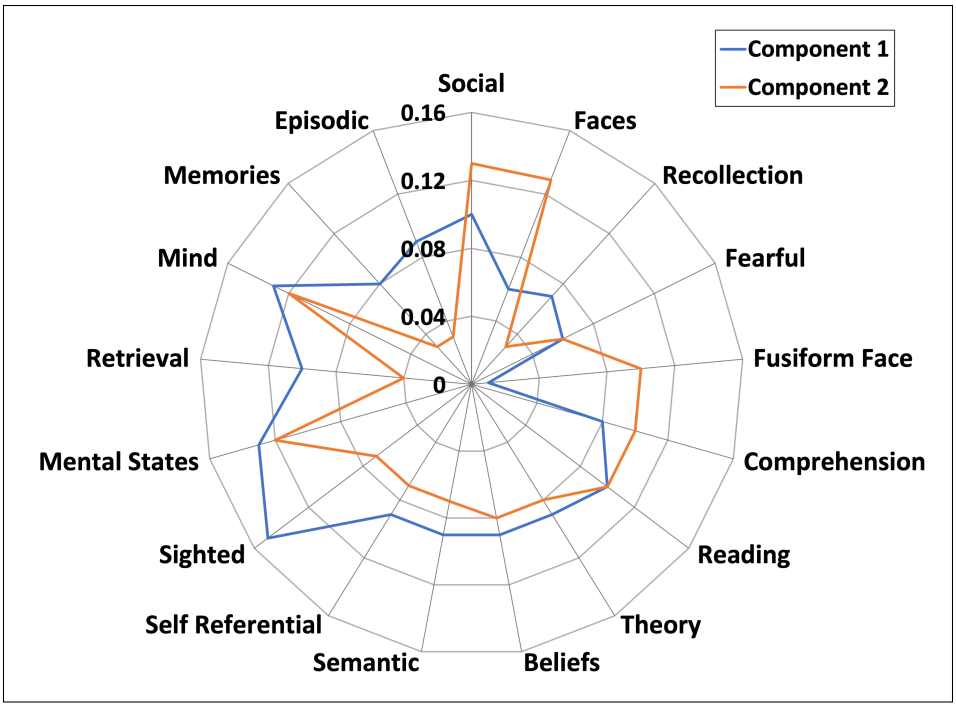


Figure 3-17. A detailed description of all the brain regions identifies by our model for SDMT data. The correlation between the loading scores and the "g" scores are identified by ρ, and level of significance is captured by the FDR corrected p-value.

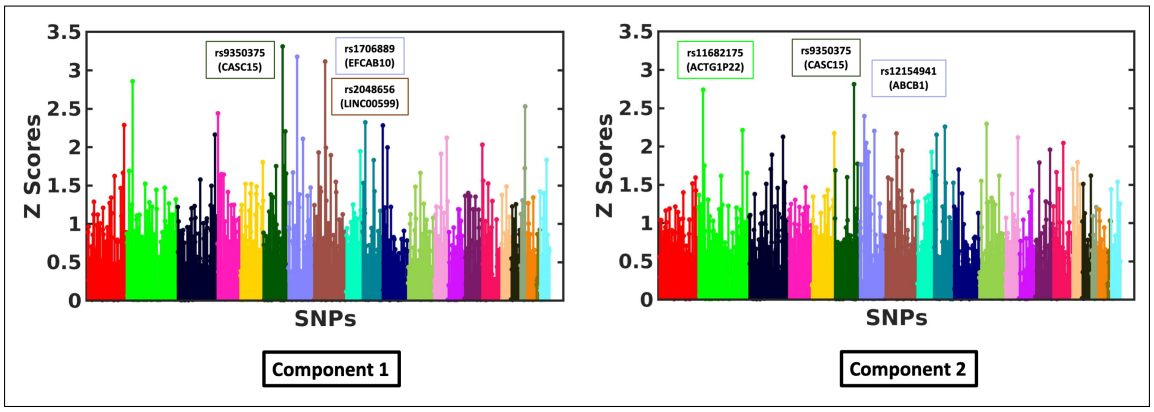


Figure 3-18. The importance map of all the SNP and their overlapping genes across all the subsamples for SDMT data.

3.2.8 Discussion

Our generative-discriminative framework exploits the interconnectedness of two different data modalities. The dictionary learning module extract features from the imaging and genetic data that are strongly connected with each other, while the

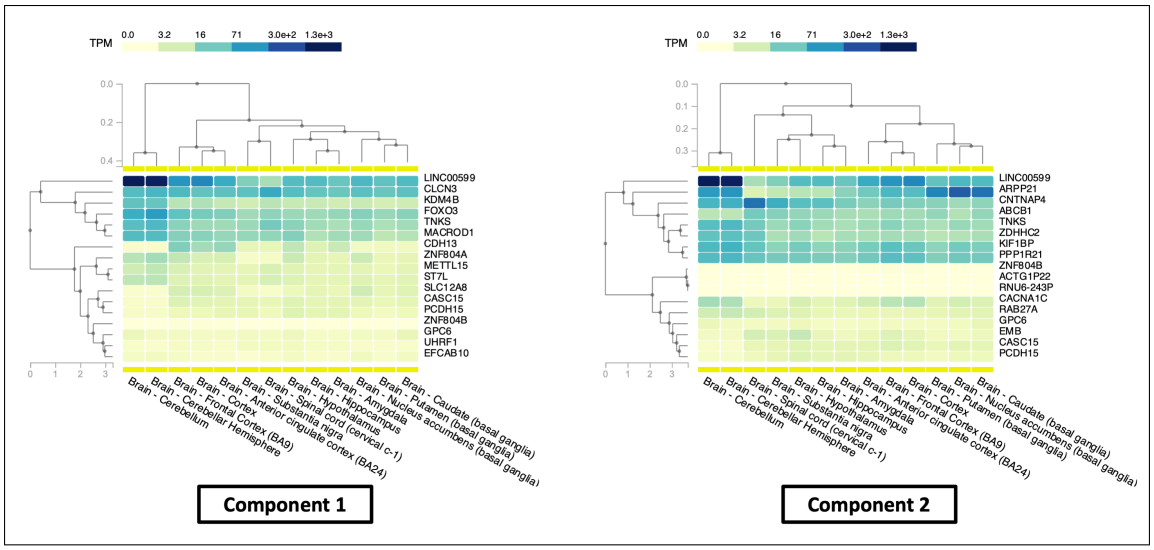


Figure 3-19. The gene expression pattern of the top genes identified from the SDMT task based on the GTEx database.

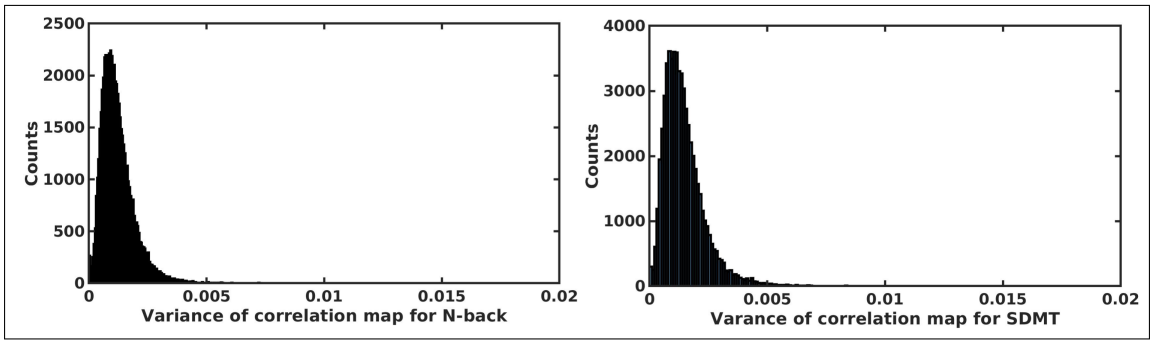


Figure 3-20. The distribution of variance between each pair of brain regions over the 10 cross validation fold.

classification module guides our framework to identify patterns that are representative of the disease. The regularity terms enforce additional structure associated with the data, i.e., the genetic regularizer captures sparse representative patterns from the data and the graph Laplacian penalty captures the grouping effect between different brain regions. Empirically, we find that the Laplacian structure is stable across different cross validation folds. Fig. 3-20 illustrates the histogram of variance in the correlation maps w_{ij} used as the regularizer for $\boldsymbol{A}$ in Eq. 3.18 when computed across the 10 cross validation folds. This stability may be partially attributed to the task fMRI paradigm, which tends to activate similar brain areas across subjects. In the preprocessing stage

of our analysis we parcellate the brain activation maps into 246 regions and use them as input to our model. Given the small dataset $N \sim 100$, this parcellation scheme balances the expressibility of the data while maintaining the stability of our model. Additionally, averaging the brain activation over multiple voxels smooths out the noise and helps us to find meaningful patters across groups. Finally, the consistent labelling of the brain regions across subjects enables us to interpret our results and perform further exploratory analysis.

We use an alternating minimization strategy to optimize our coupled framework. Alternating minimization is popular for large-scale non-convex problems due to the simple implementation and empirically stable performance. With that said, there are few theoretical convergence guarantees. While our objective function is bounded from below, convergence to a local minim depends on how well the objective function decreases after each iteration, which finally depend on the convergence properties of Eq. (3.19), Eq. (3.25), and Eq. (3.26). Our objective function is continuously differentiable and convex with respect to $\{\boldsymbol{B}, \boldsymbol{X}, \boldsymbol{c}\}$. The works of [169, 170] show that under such conditions alternating minimization converges to a stationary point. However, the orthogonality constraint over the imaging basis matrix $\boldsymbol{A}$ makes the problem non-convex. The work of [153] shows the convergence property of the orthogonality constraint using ADMM. Despite the lack of theoretical guarantees, we observe a robust empirical convergence of our alternating minimization procedure to a local minima. Thus, in practice, our optimization strategy is stable across the different datasets and initializations used in our experiments.

In Section 3.2.7.4 we demonstrate that our model achieves better classification accuracy than the baselines across all three datasets. In Section 3.2.7.5 we go a step further and present a strategy to identify a robust set of discriminative biomarkers that are coupled via the latent projections $\boldsymbol{x}$ across the imaging and genetic data. Through the meta-analysis we show that these biomarkers are strongly related with the disease

propagation pathway of schizophrenia. For example, the N-Back biomarkers involve regions from dorsolateral prefrontal cortex, and default mode network, which are known in literature to be affected by schizophrenia. Likewise, the genetic biomarkers are expressed in multiple regions of brain, which shows a probable association between genetic risk and the disease propagation pathway. Similarly, in the SDMT analysis we see association between parahippocampal activity and genes that are associated with multiple behavioral deficits.

In this exploratory analysis we note that the estimated components contain overlapping brain regions. This behavior may be attributed to our optimization strategy. In order to capture the variance of the data, the model may assign more than one basis vector to the same subset of features. The regularizations and the constraints does not prevent our model to identify components with spatial overlap, which facilitates the behavior. As a second stage of our exploration study we further show that these set of biomarkers show strong association with the cognitive "g" scores. Even though performing sub-type analysis is not the target of this model but this post processing strategy helps to identify imaging and genetic interactions which may prove to be significant for identifying novel therapeutic targets.

One disadvantage of our framework is that, it is invariant to changes in sign, so the exact association between a imaging or genetic region with the disease is unknown. Moreover, the identified set of SNPs from our model are most likely tag-SNPs [171], meaning that there is a low probability that they are causal. An added complexity is that the SNPs may not lie in a genetic region, but they still affect a gene by modulating the regulatory factors. Hence, further analysis is required to identify the potential gene targets for therapy.

One limitation of this work is the relatively small sample size. We demonstrate that in this setting our generative-predictive framework can outperform traditional machine learning methods across two task fMRI paradigms and two sites. With that

said, we acknowledge that follow-up studies should be done to validate this framework on a larger cohort.

Finally, our current framework only considers disease classification via the logistic regression term in Eq. (3.18). However, psychiatric research is exploring the utility of a finer-grained characterization of different disorders across multiple cognitive or behavioral axes. In future, we will explore extensions of our generative-predictive framework for patient subtyping via ordinal regression and multivariate linear regression. We will also explore nonlinear relationships between the data modalities. As alluded to above, incorporating more complex relationships may help us to build a bigger picture of the disease under study. Hence, in the future work we will explore pathway specific information for better understanding of the disease propagation.

3.2.9 Summary

We have presented a novel generative-discriminative framework that relies on coupled latent projections to jointly model imaging and genetics data. The projection operations leverage a dictionary learning setup, where the imaging and genetics basis matrices capture representative facets of the data. The projection coefficients are tied across modalities and are input to a logistic regression model to predict class diagnosis. We have demonstrated our framework on a population study of schizophrenia. Our generative-discriminative approach achieves better diagnostic classification accuracy than competing machine learning baselines, and it implicates an interpretable set of biomarkers that underlie the well-documented deficits in schizophrenia. Finally, our model is agnostic to the imaging modality and the clinical population. Hence, it is a powerful tool to study a range of neuropsychiatric disorders.

Chapter 4

A Deep Neural Network Architecture Exploring Non-Linearity In Modeling Multimodal Imaging and Genetic Data

Neuropsychiatric disorders like schizophrenia and autism are multifaceted [172–174]. They are characterized by cognitive dysfunction hallucinations, along with social and behavioral challenges [175–177]. At the same time, family and twin studies [32, 33] have found a strong genetic underpinning associated with these disorders. However, the genetic influence of neuropsychiatric disorders is complex and often guided by additional environmental factors and gene-gene interactions [23, 34]. Traditional imaging genetic studies [11, 14] and our previous works [59–61] focused on exploring biomarkers and disease prediction in a multivariate linear framework. Such models extract a representation from the imaging and genetic data and associate it with disease labels in a linear model. Fitting a linear model is an over-simplistic assumption that does not account for the complex interaction between brain activations and genetics.

Deep learning approaches are widely known for their capability of combining low-level input features and extracting high-level non-linear data representations [79,

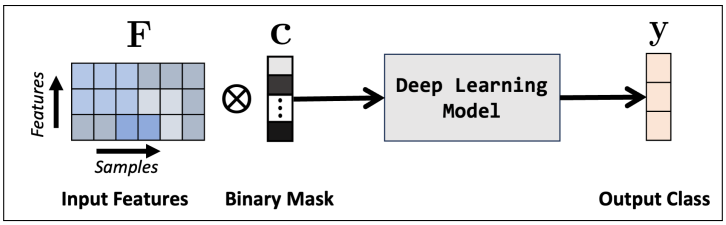


Figure 4-1. A general framework for feature selection in deep learning models. $\boldsymbol{F} \in \boldsymbol{R}^{d \times N}$ is the input data matrix with N samples and d-dimensional features. $\boldsymbol{y}$ is the output class labels. The feature selection mask $\boldsymbol{c}$ is a d dimensional binary vector multiplied elementwise with $\boldsymbol{F}$.

80, 119]. These models provide a strategy to deviate from linear models and explore non-linear interactions between imaging and genetics data. The main drawback of deep learning is the lack of interpretability. However, interpretable AI has recently provided multiple strategies to find biomarkers [129, 130] and track the information flow through the model.

This work [62] introduces a novel autoencoder model to predict neuropsychiatric disorders while finding biomarkers. Our autoencoder is complemented with a Bayesian feature selection module that masks out non-informative features and passes discriminative information through the autoencoder module. On a high level, the Bayesian module subselects biomarkers, while the autoencoder models the non-linear relationship between imaging and genetic data. In addition, we have coupled a classifier with the autoencoder for diagnosis. The autoencoder, coupled with the Bayesian module and the classifier, provides a robust and adaptable framework to predict the underlying disorder while finding susceptible imaging and genetic biomarkers.

4.1 Bayesian Feature Selection Strategy In Deep Learning Models

Fig. 4-1 shows a general framework for using our Bayesian feature selection strategy in deep learning models. Mathematically, let $\boldsymbol{F} \in \boldsymbol{R}^{d \times N}$ be the input data matrix with N samples and d-dimensional features, $\boldsymbol{c}$ is the binary feature selection mask, and $\boldsymbol{y}$

is the class labels. The problem of feature selection can be viewed as masking the irrelevant features before passing them through the model. From a Bayesian viewpoint, the importance of each feature can be estimated by inferring the posterior probability distribution $P(\boldsymbol{c}|\boldsymbol{F}, \boldsymbol{y})$ given the paired dataset: $\mathcal{D} = \{\boldsymbol{F}, \boldsymbol{y}\}$. However, the desired posterior distribution $p(\boldsymbol{c}|\boldsymbol{F}, \boldsymbol{y})$ is intractable due to an exponentially large number of possible binary configurations of $\boldsymbol{c}$.

One strategy is to minimize the KL divergence between an approximate distribution $q(\cdot)$ and the true posterior distribution $KL(q(\boldsymbol{c})||p(\boldsymbol{c}|\boldsymbol{F}, \boldsymbol{y}))$. Mathematically, this optimization can be written as:

$$\underset{q(\cdot)}{\operatorname{argmin}} \quad -E_q\left[\log\left(p\left(\boldsymbol{y}|\boldsymbol{F}, \boldsymbol{c}\right)\right)\right] + KL\left(q(\boldsymbol{c})||p(\boldsymbol{c})\right), \tag{4.1}$$

where $p(\boldsymbol{c})$ is a prior over the binary masks. In our approach, we use Binary concrete distribution [178, 179] as $q(\cdot)$. Under this distribution $\boldsymbol{c}$ can be viewed as a continuous relaxation of Bernoulli random variable. Mathematically, each element of $\boldsymbol{c}$ can be written as:

$$\boldsymbol{c}(i) = \sigma\left(\frac{\log(\boldsymbol{p}(i)) - \log(1 - \boldsymbol{p}(i)) + \log(U) - \log(1 - U)}{t}\right) \tag{4.2}$$

where U is sampled from $Uniform(0, 1)$, the parameter t controls the relaxation from the $\{0, 1\}$ Bernoulli, and $\boldsymbol{p}$ are the parameters of the proposal distribution. A unique property of binary concrete vectors is that $\lim_{t\to 0} P(\boldsymbol{c}(i) == 1) = \boldsymbol{p}(i)$. This property shows that, like Bernoulli, $\boldsymbol{p}$ captures the relative importance of each feature.

Eq. (4.1) does not have a closed-form solution. However, it can be optimized via Monte Carlo integration by sampling the vectors $\boldsymbol{c}$ according to Eq. 4.2. The continuous relationship between $\boldsymbol{c}$ and $\boldsymbol{p}$ allows us to optimize $\boldsymbol{p}$ using stochastic gradient descent [180, 181]. Finally, the first term of Eq. 4.1 can be viewed as a likelihood loss for the deep learning model. During disease prediction, this loss can be generated as the binary cross-entropy loss, where the input features $\boldsymbol{F}$ are

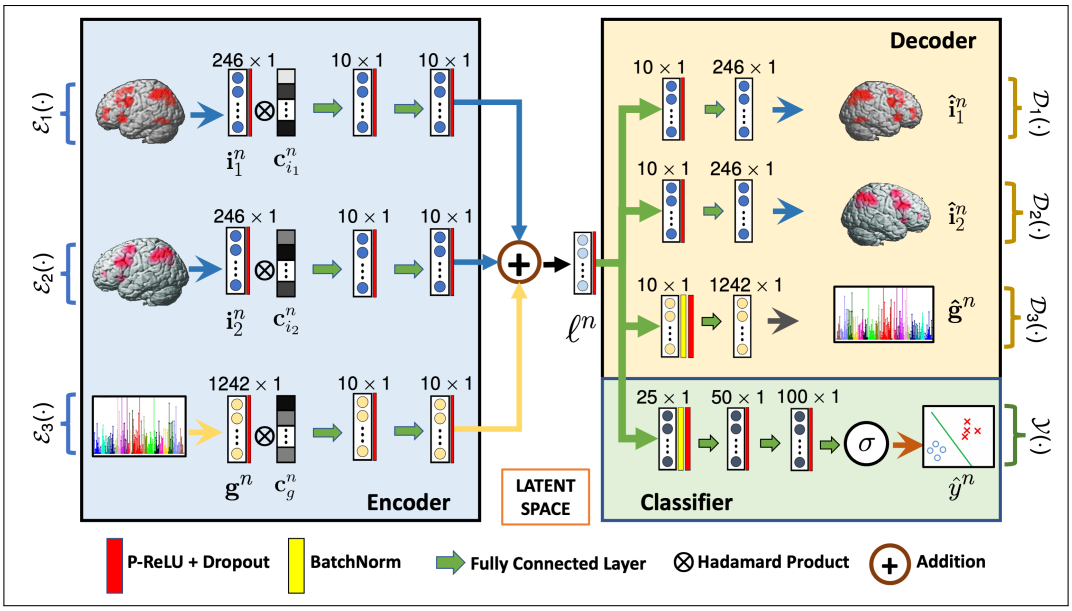


Figure 4-2. G-MIND architecture. The inputs $\{\boldsymbol{i}_1, \boldsymbol{i}_2\}$ and $\{\boldsymbol{g}\}$ corresponds to the two imaging modalities and genetic data, respectively. $\mathcal{E}_i(\cdot)$ and $\mathcal{D}_i(\cdot)$ captures the encoding and decoding operations, and $\mathcal{Y}(\cdot)$ captures the classification operation. $\boldsymbol{c}_i$ is the Bayesian feature selection mask, and $\boldsymbol{\ell}^n$ is the low dimensional latent space.

masked according to $\boldsymbol{c}$. Lastly, we want to note that the feature selection strategy is independent of the deep learning model so that it can be used across many models.

4.2 GMIND: The Multi-modal Encoder-Decoder Framework

In this work, we introduce an autoencoder framework that can combine multiple imaging data modalities with genetic data while finding discriminative biomarkers. Fig. 4-2 illustrates our full model. The inputs $\boldsymbol{i}_1^n$ and $\boldsymbol{i}_2^n$ denotes the input imaging modalities for subject n. In our case $\boldsymbol{i}_1^n$ and $\boldsymbol{i}_2^n$ are activation maps from two different fMRI paradigms. The input $\boldsymbol{g}^n$ represents the SNP genotype, and y^n is a binary class label (patient or control). Let N_1, N_2, and N_g denote the number of subjects from whom we have the corresponding imaging or genetic modality. Let R be the total number of ROIs in the brain, and G be the total number of SNPs. The imaging data has the dimensionality $\boldsymbol{i}_1^n, \boldsymbol{i}_2^n \in \boldsymbol{R}^{R\times 1}$, and the genetic data has the dimensionality

$\boldsymbol{g}^n \in \boldsymbol{R}^{G\times 1}$. We jointly model the imaging and genetic modalities using an auto-encoder framework. The first layer of the encoder incorporates the Bayesian feature selection layer, parameterized by $\boldsymbol{p}_m$ for each modality m. We use the resulting low dimensional representation ℓ^n for subject classification.

4.2.1 Feature Importance using Learnable Dropout

We followed the general framework of the Bayesian feature selection strategy explained in Section 4.1. We incorporate the Bayesian feature selection as a learnable dropout layer. The standard Bernoulli dropout independently drops nodes using a fixed probability defined by the user. Here, we wish to learn these values, so we reparameterize the Bernoulli dropout mask as defined in Eq. 4.2. This continuous relaxation [179, 180, 182] of the Bernoulli random variable enables us to update the dropout probabilities while training the network. During each forward pass through the network we sample random variables $\boldsymbol{c}^n_{i_1}, \boldsymbol{c}^n_{i_2} \in \boldsymbol{R}^{R\times 1}$, and $\boldsymbol{c}_g \in \boldsymbol{R}^{G\times 1}$ for imaging and genetic data, respectively, from a binary concrete distribution and use it as a dropout mask for patient n:

$$\boldsymbol{c}^n_{i_1} = \sigma\left(\frac{\log(\boldsymbol{p}_{i_1}) - \log(1-\boldsymbol{c}_{i_1}) + \log(\boldsymbol{u}^n_{i_1}) - \log(1-\boldsymbol{u}^n_{i_1})}{t}\right) \tag{4.3}$$

where $\boldsymbol{u}^n_{i_1}$ is a random vector sampled from $Uniform(0,1)$, the parameter t (temperature) controls the extent of relaxation from the Bernoulli distribution and $\boldsymbol{p}_m$ captures the probabilities with which the features of modality m are selected. As seen in Eq. (4.3) when the probability $\boldsymbol{p}_m$ is close to 1 that feature will be selected most of the time, as compared to a feature whose probability is close to 0. We further incorporate a sparsity penalty over the probabilities $\boldsymbol{p}_m$ via the KL divergence $KL\left(Ber(\boldsymbol{p}_0)||Ber(\boldsymbol{p}_m)\right)$ where $\boldsymbol{p}_0$ is a hyperparameter fixed to 0.001. Effectively, this term encourages sparsity in the elements of $\boldsymbol{p}_m$.

4.2.2 Multimodal Latent Encoding

The encoder learns a nonlinear latent space that is shared between all the data modalities. As shown in Fig. 4-2 we encode the data following the dropout using a cascade of fully connected layers followed by a PRelu activation [121]. Unlike standard autoencoder-based networks, we couple the low-dimensional representations of each data modality to leverage the common structures shared between them. The latent embedding ℓ^n is computed as

$$\boldsymbol{\ell}^n = \frac{1}{M_n}\left(\mathcal{E}_1(\boldsymbol{i}_1^n, \boldsymbol{c}_{i_1}^n) + \mathcal{E}_2(\boldsymbol{i}_2^n, \boldsymbol{c}_{i_2}^n) + \mathcal{E}_g(\boldsymbol{g}^n, \boldsymbol{c}_g^n)\right) \tag{4.4}$$

Here $\mathcal{E}_i(\cdot)$ represents the encoding operation for modality m, and M_n is the number of modalities present for subject n. As seen in Eq.(4.4), our latent representation is the sum of the individual projections, scaled by the amount of available data M_n. This fusion strategy encourages the latent encoding for an individual patient to have a consistent scale, even when constructed using a subset of the modalities.

4.2.3 Data Reconstruction

The decoder reconstructs the data from the latent representation to ensure that the encoder is preserving sufficient information about the inputs. We use fully connected layers along with PRelu, dropouts, and batchnorm for decoding. Mathematically, the autoencoder loss is the l_2 norm between the input and reconstruction:

$$\sum_{n=1}^{n_1} ||\boldsymbol{i}_1^N - \mathcal{D}_1(\ell^n)||_2^2 + \sum_{n=1}^{N_2} ||\boldsymbol{i}_2^n - \mathcal{D}_2(\ell^n)||_2^2 + \sum_{n=1}^{N_g} ||\boldsymbol{g}^n - \mathcal{D}_3(\ell^n)||_2^2$$

where $\mathcal{D}_m(\cdot)$ is the decoding operation for modality m.

4.2.4 Disease Classification

The final piece of our network is a classifier for disease prediction, which will encourage the dropout mask and latent embeddings to select discriminative features from the

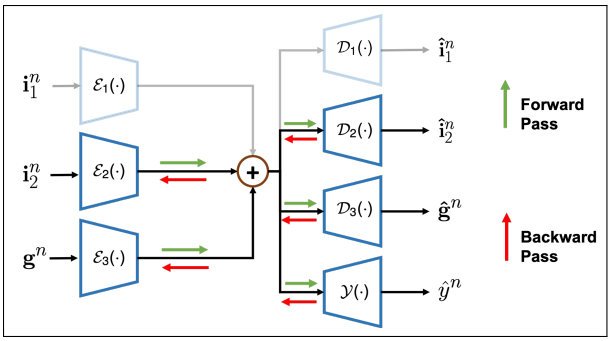


Figure 4-3. Information flow during the forward pass (green) and backward pass (red) when $\boldsymbol{i}_1^n$ is absent.

Institution	Modalities		
	NBack	SDMT	SNP
LIBD	160	110	210
BARI	97	–	97

Table 4-I. The number of subjects present for each modality from the two institutions. Note that the SDMT task was not acquired for BARI.

data. We employ fully connected layers, and a cross entropy loss for classification: $-\sum_{n=1}^{N}\left(y^n \log(\hat{y}^n) + (1-y^n)\log(1-\hat{y}^n)\right)$, where y is the original class label and $\hat{y}^n$ is the predicted class label.

Our combined G-MIND objective function can be written as follows:

$$\begin{aligned}\mathcal{L}(\boldsymbol{i}_1,\boldsymbol{i}_2,\boldsymbol{g}) = \lambda_1 \sum_{n=1}^{N_1} ||\boldsymbol{i}_1^n - \mathcal{D}_1(\ell^n)||_2^2 + \lambda_2 \sum_{n=1}^{N_2} ||\boldsymbol{i}_2^n - \mathcal{D}_2(\ell^n)||_2^2 \\ + \lambda_3 \sum_{n=1}^{N_g} ||\boldsymbol{g}^n - \mathcal{D}_3(\ell^n)||_2^2 - \lambda_4 \sum_{n=1}^{N} \left(y^n \log(\hat{y}^n) + (1-y^n)\log(1-\hat{y}^n)\right) \\ + \lambda_5 \sum_{m=1}^{3} \sum_{k} KL(Ber(\boldsymbol{p}_0(k)) || Ber(\boldsymbol{p}_m(k))\end{aligned} \tag{4.5}$$

where N is the total number of subjects. The parameters $\{\lambda_1, \lambda_2, \lambda_3\}$ control the contributions of the data reconstruction error, λ_4 controls the contribution of classification error, and λ_5 regularizes the sparsity on $\boldsymbol{p}_m$.

The summation in Eq. (4.5) enables G-MIND to handle missing data. For example, if $\boldsymbol{i}_1^n$ is not available for subject n, then the gradients with respect to encoder $\mathcal{E}_1(\cdot)$ and decoder $\mathcal{D}_1(\cdot)$ will be zero. As illustrated in Fig. (4-3), information will flow into and out of the latent space through the other network branches and will only be used to update those parameters.

4.2.5 Prediction on New Data

During training, we learn the encoder, decoder, and classifier weights, along with the probabilistic masks $\boldsymbol{p}_m$ by minimizing Eq. (4.5). We then threshold the probabilistic mask $\hat{\boldsymbol{p}}_m = (\boldsymbol{p}_m > \tau_m)$ to select the most important features for reconstruction and classification. When testing on a new subject data, we premultiply the available modalities by the thresholded dropout mask, i.e., $\hat{\boldsymbol{i}}_1^n = \boldsymbol{i}_1^n \otimes \hat{\boldsymbol{p}}_{i_1}$. The masked input $\hat{\boldsymbol{i}}_1^n$ is sent through encoder and the classifier for diagnosis. We do not use the learned dropout procedure during testing, since different samples of $\boldsymbol{c}_m^n$ may lead to a different diagnosis, whereas our goal to obtain a deterministic label for each subject.

4.2.6 Implementation Details

We set the regularization parameters of our model $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\}$ as $10^{-\beta_i}$ where β_i is selected such that λ_i multiplied by the appropriate loss term lies within the same order of magnitude (1—10). This criterion is intuitive (i.e., equal importance is given to both the imaging and genetic data), and it is not performance driven (i.e., we do not cherry-pick the values to optimize prediction accuracy). The corresponding values for all the experiments are: $\lambda_1 = 0.1, \lambda_2 = 0.1, \lambda_3 = 0.01, \lambda_4 = 0.1$, and $\lambda_5 = 0.01$. We fix the Bernoulli probability, to $q = 0.001$ and the temperature variable to $t = 0.1$. Based on 10-fold cross validation results we fix all detection threshold values to $\tau_i = 0.1$. The architecture of our model (layer sizes and nonlinearities) is shown in Fig. 4-2.

4.2.7 Baseline Comparison Methods

We compare G-MIND to classical machine learning techniques and architectural variants that omit key features.

- **Multimodal Support Vector Machine (SVM):** We construct a linear SVM classifier after concatenating all the data modalities $[\boldsymbol{i}_1^T, \boldsymbol{i}_2^T, \boldsymbol{g}^T]^T$. Notice that this

model cannot handle missing data. Therefore, we fit a multivariate regression to impute missing imaging modalities based on the available one for each subject. For example if $\boldsymbol{i}_1^n$ is absent, we impute it as: $\boldsymbol{i}_1^n = \boldsymbol{\beta}^* \boldsymbol{i}_2^n$, where $\boldsymbol{\beta}$ is the regression coefficient matrix obtained from training data. We use a grid search method to find the best set of hyper-parameters. Notice that this tuning provides an *added advantage* for SVM over G-MIND.

- **Multimodal CCA + RF:** Canonical correlation analysis (CCA) identifies bi-multivariate associations between imaging and genetics data. This approach is similar to our coupled latent projection, but the traditional CCA does not accommodate more than two data modalities. In order to overcome this we concatenate the imaging features obtained from two experimental paradigms and perform CCA with the genetics data. We then construct a random forest classifier based on the latent projections. We use the same approach for data imputation and to find the best set of hyperparameters.

- **Encoder Only:** We compare our model to an ANN architecture based on the encoder and the classifier of G-MIND. This comparison will show us importance of using the decoder and the learnable dropout layer.

- **Encoder+Dropout:** We compare our model to another ANN architecture where we only used the encoder, the classifier, and the learnable dropout layer. This experiment will show us the performance improvement from including a decoder. Based on our 10-fold cross validation we fix the learned dropout threshold values to $\{\tau_{i_1} = 0.05, \tau_{i_2} = 0.05, \tau_g = 0.1\}$.

Method \ Perf	Sens	Spec	Acc	Auc
SVM	0.66	0.47	0.58	0.55
CCA+RF	0.15	**0.92**	0.51	0.56
Encoder Only	0.57	0.57	0.57	0.59
Encoder + Dropout	0.61	0.56	0.59	0.62
G-MIND	**0.75**	0.58	**0.67**	**0.68**

Table 4-II. Testing performance of each method on LIBD during 10 fold cross validation.

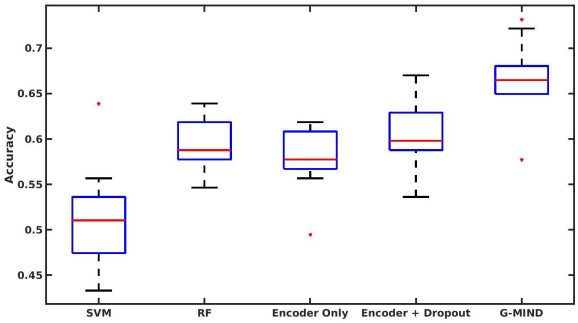


Figure 4-4. Distribution of accuracies by the models trained in all 10 CV folds, when directly evaluated on BARI.

4.3 Experimental Results

4.3.1 Data and Preprocessing

Our first dataset includes two task fMRI paradigms and SNP data provided by Lieber Institute for Brain Development (LIBD) in Baltimore, MD, USA. The first fMRI paradigm is a Nback working memory task and the second fMRI paradigm is an event-based simple declarative memory task (SDMT). Our replication dataset includes just Nback and SNP data acquired at the University of Bari Aldo Moro, Italy (BARI). The distribution of the subjects is shown in Table 4-I. The fMRI acquisition and the preprocessing are described in Section 2.4.1. We use the Brainnetome atlas [143] to define 246 cortical and subcortical regions. The input to our model is the contrast map over these ROIs.

In parallel, genotyping was done using variate Illumina Bead Chips including 510K/ 610K/660K/2.5M. After quality control (Section 2.4.1), the 102K linkage disequilibrium independent SNPs are further subselected based on a GWAS p-value threshold of $P < 10^{-4}$. The resulting 1242 linkage disequilibrium independent SNPs are used as inputs to the model. As a preprocessing step, we remove the effect of age, IQ, and education from the imaging modalities, and we have mean centered all the data modalities.

4.3.2 Model Performance

Table 4-II quantifies the 10-fold testing performance of all the methods on multimodal data obtained from LIBD. We can clearly see that G-MIND achieves the best overall accuracy. Even in the presence of missing data our multi modal approach can successfully extract meaningful information from all the data modalities that are essential for diagnosis prediction. Our results also show the importance of the decoder and the dropout layer.

In order to show the generalizability of our method, we trained our model on LIBD data and tested it without fine-tuning on a cross-site dataset from BARI. This experiment captures the transference property of our model. We note that the SDMT task was not acquired at BARI, so the corresponding branch of G-MIND is not used. We evaluate the 10 best models obtained from the 10 different folds to run this experiment. Fig. 4-4 shows the distribution of accuracies of all the models in the form of a boxplot. Here we can see that our method shows the best transference property compared to all the baselines. This is an interesting result as it shows the robustness of our model against data acquisition noise and population-specific noise. This performance gain further suggests that the learnable dropout mask can identify a robust set of features most predictive of the disease.

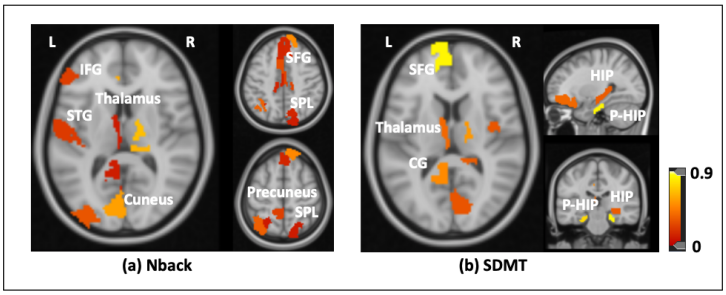


Figure 4-5. The representative set of brain regions as captured by the dropout probabilities $\{\boldsymbol{p}_1, \boldsymbol{p}_2\}$. The color bar denotes the median value across 10 folds.

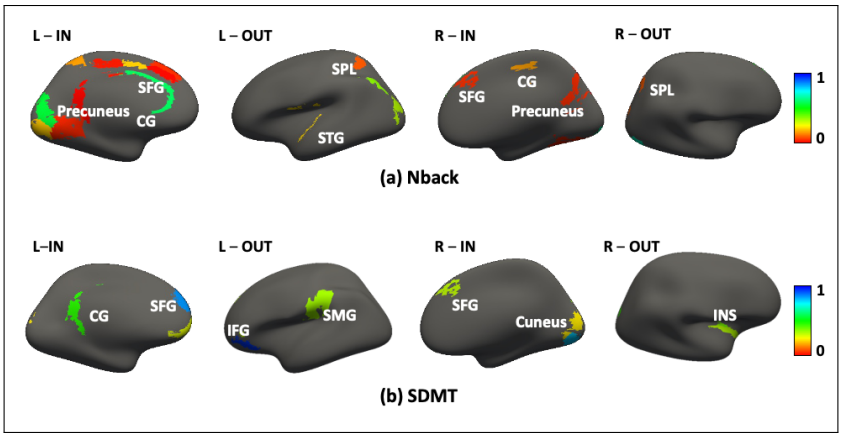


Figure 4-6. The surface plot of the brain regions as captured by the dropout probabilities $\{\boldsymbol{p}_1, \boldsymbol{p}_2\}$. The color bar denotes the median value across 10 folds. From **Left** to **Right** the images are internal surface of left hemisphere (**L–IN**), external surface of left hemisphere (**L–OUT**), internal surface of right hemisphere (**R–IN**), and external surface of right hemisphere (**R–OUT**).

4.3.3 Analysis of Imaging Biomarkers

Fig. 4-5 illustrates the most important set of brain regions as identified by the median concrete dropout probability maps $\{\boldsymbol{p}_{i_1}, \boldsymbol{p}_{i_2}\}$ across the 10 validation folds. We further show a more global picture of the high importance brain regions as a surface plot in Fig. 4-6. Both from Fig. 4-5 and Fig. 4-6 for the Nback task, we can see regions that include superior frontal gyrus (SFG), and inferior frontal gyrus (IFG), which

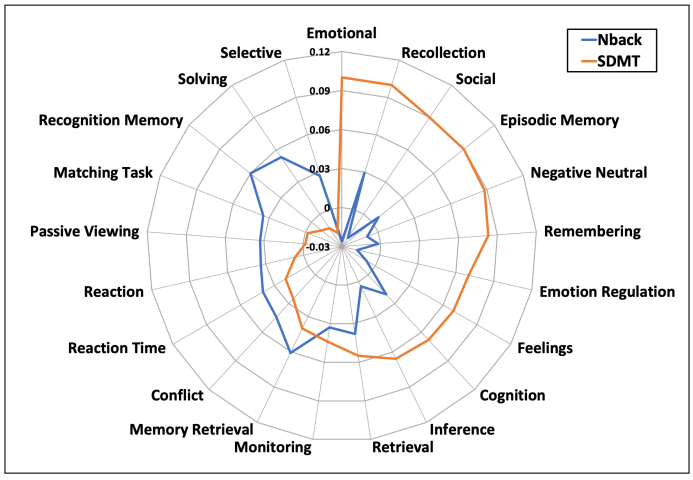


Figure 4-7. The level of association with different cognitive states of all the brain regions identified by our model as found in the Neurosynth database.

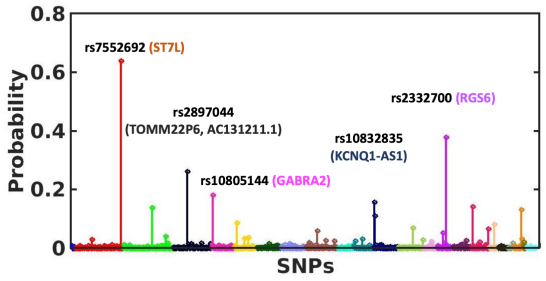


Figure 4-8. The median importance map of all the SNP across and their overlapping genes across the 10 folds.

Biological Processes	FDR
Central nervous system development	0.005
→ Nervous system development	0.001
→ System development.	0.005
Generation of neurons	0.005
→ Neurogenesis	0.004
Regulation of calcium ion transport into cytosol	0.04
→ Regulation of sequestering of calcium ion	0.008

Table 4-III. The enriched biological processes and their level of significance obtained via GO enrichment analysis.

are known to sub-serve executive cognition [4]. Moreover, we can see regions (SFG, IFG) from the dorsolateral prefrontal cortex [4] and regions (SPL, STG) from the posterior parietal cortex that overlaps with the fronto-parietal network, which is known to be altered in schizophrenia. Further clusters incorporate components of the default mode network also implicated in schizophrenia [161]. The SDMT biomarkers implicate the hippocampal, parahippocampal, and superior frontal regions along with the anteromedial thalamus, which are also affected in schizophrenia [3]. These regions control executive cognition and memory encoding and that are also known to be associated with the disorder.

We further use Neurosynth [162] to decode the higher-order brain states of the biomarkers associated with Nback and SDMT tasks. This analysis allows us to quantitatively compare the selected brain regions with previously published results and gives us a level of association with different brain states as identified by other studies. Fig. 4-7 shows the Neurosynth terms that are strongly correlated with our biomarkers. We note that the terms associated with the Nback task correspond to recognition and solving, while the brain states for SDMT are associated with emotions and memory encoding. These results provide further evidence that G-MIND can extract potential imaging biomarkers that are highly relevant to the task and the disorder under study.

4.3.4 Analysis of Genetic Biomarkers

Fig. 4-8 shows the importance map across the 1242 SNPs as computed by the median $\boldsymbol{p}_g$ across the 10 folds. We annotated each SNP based on its overlapping or nearest gene as found from the SNP-nexus web interface [163]. In addition, we ran a gene ontology enrichment analysis of the overlapping genes of the top 300 SNPs to identify the enriched biological processes [142]. This enrichment analysis allows us to identify the set of over-represented genes in a biological pathway that may be associated with the disease phenotype. Table 4-III captures the most significant biological processes implicated by the set of SNPs, which include the *nervous system development* [183], and *calcium ion regulation* [184] which are known to be strongly associated with schizophrenia. As parallel to Neurosynth analysis, we perform a gene expression based analysis [164] over the 10 overlapping (or nearest gene if there is no overlap) genes of the top SNPs identified from our analysis. Here we use the GTEx database to identify the set of brain tissues where these genes show high levels of expression. This exploratory analysis may help us to understand the *cis*-effects of the SNPs and

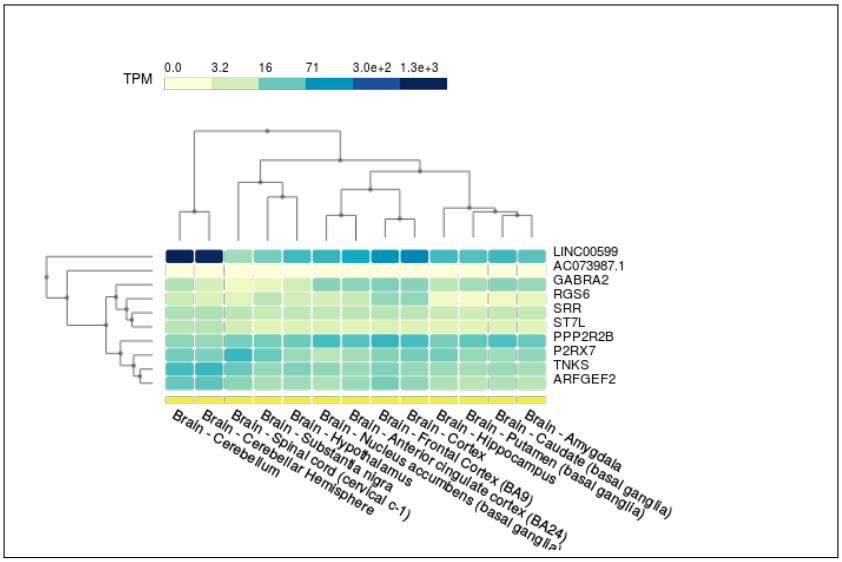


Figure 4-9. The gene expression pattern of the selected set of genes in different brain tissues based on the GTEx database. Higher level of a gene expression in a brain tissue imply that alteration in that gene may have a stronger effect on those specific brain regions.

how they alter the functionalities of genes expressed in different tissues of the brain. Fig. 4-9 shows the gene expression pattern of each gene across different brain tissues. Here, *LINC00599* shows high expression levels in brain and are also known to be associated with schizophrenia [168] and neuroticism [165]. These findings show that the model can be used to explore potential genetic biomarkers and their interactions in a multivariate framework.

4.4 Discussion

We introduce a novel autoencoder that combines interpretability and multimodal data integration in a deep learning framework. The first key contribution of GMIND is the Bayesian feature selection strategy that allows us to jointly learn the biomarkers in an end-to-end data driven fashion. The feature selection layer is added to the encoder via a learnable dropout layer. Unlike traditional dropout, the continuous relaxation between the underlying probability map and the dropout mask allows us to train the model using straightforward gradient descent-based approaches. Adding a dropout layer requires minimum changes to the deep learning model, which makes this approach adaptable to other neural network architectures. Additionally, the Bayesian framework provides a probabilistic measure of the importance of each feature. The probabilistic interpretation allows us to compare feature importance across different populations and experiments. In contrast, standard heuristic-based feature selection methods provide a relative score that cannot be compared across experiments due to a lack of probabilistic intuitions.

The second key contribution of our approach is the autoencoder framework. The autoencoder architecture provides a robust and adaptable framework to integrate new data modalities [185] simply by adding new encoder-decoder branches. Mathematically, a new branch will introduce another term to the loss function but does not alter the optimization procedure (e.g., backpropagating gradients)[186]. In addition,

missing data can easily be handled by freezing the affected part of the network [128] and updating the remaining weights. This simplicity starkly contrasts the classical methods [11, 48, 61], where the entire model and optimization procedure must be changed for each new modality and missing data configuration.

One limitation of this approach is that we heavily rely on the subselection of genetic variants to make our model stable and prevent overfitting. However, neuropsychiatric disorders are polygenic [16, 17, 38], which means the genetic risk is spread across the whole genome. The subselection step ignores the majority of the variants and fails to account for the complexity of the genetic architecture associated with neuropsychiatric disorders. In the following two chapters, Chapter 5,6, we will provide strategies and tools to parse the complexity of the genetic risk by identifying target loci and pinpointing discriminatory pathways.

4.5 Summary

We have presented G-MIND, a novel deep network to integrate multimodal imaging and genetic data for targeted biomarker discovery and class prediction. Our unique use of learnable dropout with a classification module helps us to identify discriminative biomarkers of the disease. Our unique loss function enables us to handle missing modalities while mining all the available information in the dataset. We demonstrate our framework on fMRI and SNP data of schizophrenia patients and controls from two different sites. The improved performance of G-MIND across all the experiments shows the capability of this model to build a comprehensive view of the disorder based on incomplete information obtained from different modalities. We note that our framework can easily be applied to other imaging modalities, such as structural and diffusion MRI simply by adding autoencoder branches. In future work, we will develop a hybrid extension of G-MIND in which we incorporate pathway-specific information into the deep learning architecture for a better understanding of disease propagation.

Chapter 5

Identifying Genetic Biomarkers from GWAS Summary Statistics

The genetic risks associated with neuropsychiatric disorders like schizophrenia and autism are highly complex and polygenic. Although the genetic architectures of schizophrenia and autism are still elusive, evidence suggests they involve many genes [16, 37] and are distributed across pathways [187]. The common approach to identifying the risk loci is using Genome Wide Association Studies (GWAS). The recent GWAS in schizophrenia has identified 287 risk loci [37], where each variant only explains a tiny proportion of risk for schizophrenia. GWAS provides regions of high importance in the DNA but fails to identify the target variants. Pinpointing the causal variants is essential to discovering their downstream regulatory effect on gene expression profiles and biological processes. In addition, as described in Section 2.2.3.1, the univariate nature of GWAS leads to inflation of effect sizes due to the correlation structure present between the variants [43]. The inflated effects often result in many false positives.

This drawback is addressed in finemapping approaches [40, 88]. Fine-mapping provides a way to uncover genetic variants that causally affect some trait of interest while considering the correlation structure of the data [88, 89]. In the space of finemapping, Bayesian finemapping approaches [39, 86, 95, 96] provide the posterior

probability of a variant being causal given the GWAS summary statistics data. In addition, they also provide small subsets of causal genetic variants [88, 89]. These subsets, known as credible sets, capture the uncertainty of finding the true causal variant within a highly correlated region [90]. Unlike p-values, the corresponding posterior inclusion probabilities (PIPs) computed during fine-mapping can be compared across studies of different sample sizes.

Current Bayesian fine-mapping approaches take into account the correlation structure between the genetic variants, but they are often computationally intensive to run and cannot handle spurious effects from non-causal variants. In this paper, we introduce a novel framework for Bayesian fine-mapping from GWAS summary data. We extend the idea of Binary concrete vectors described in Section 4.2.1. In our approach, we impose the binary concrete prior over the causal configurations that can handle spurious non-causal effects and infer the posterior probabilities of causal configuration. In a simulation study, we demonstrate that our model achieves comparable or better performance to the current fine-mapping methods across increasing numbers of causal variants and increasing noise, as determined by the polygenecity of the trait.

5.1 BEATRICE: Bayesian Fine-mapping from Summary Data using Deep Variational Inference

5.1.1 Generative Assumptions of Fine-mapping

BEATRICE is based on a generative additive effect model. Formally, let $\boldsymbol{y} \in \mathbb{R}^{n \times 1}$ denote a vector of (scalar) quantitative traits across n subjects. The corresponding genotype data $\boldsymbol{G} \in \mathbb{R}^{n \times m}$ is a matrix, where m represents the number of genetic variants in the analysis. Without loss of generality, we assume that the columns of $\boldsymbol{G}$ have been normalized to have mean 0 and variance 1, i.e., $\frac{1}{n}\sum_i \boldsymbol{G}_{ij} = 0$ and

$\frac{1}{n}\sum_i \boldsymbol{G}_{ij}^2 = 1$ for $j = 1, \ldots, m$. The quantitative trait is generated as follows:

$$\boldsymbol{y} = \boldsymbol{G}\beta + \eta \qquad \eta \sim N\left(0, \frac{1}{\tau}\mathbb{I}_n\right), \tag{5.1}$$

where $\beta \in \mathbb{R}^{m\times 1}$ is the effect size, $\eta \in \mathbb{R}^{n\times 1}$ is additive white Gaussian noise with variance $\frac{1}{\tau}$, and $\mathbb{I}_n$ is the $n \times n$ identity matrix.

5.1.2 Genome Wide Association Studies (GWAS)

GWAS uses a collection of element-wise linear regression models to estimate the effect of each genetic variant. Mathematically, the GWAS effect sizes are computed as $\hat{\beta} = \frac{1}{n}\boldsymbol{G}^T\boldsymbol{y}$, with the corresponding vector of normalized z-scores equal to $\boldsymbol{z} = \sqrt{\frac{\tau}{n}}\boldsymbol{G}^T\boldsymbol{y}$ [85, 86]. The derivations are provided in Section 2.2.2. The main drawback of GWAS is that non-causal genetic variants can have large effect sizes due to polygenicity of the quantitative trait [188], varying degrees of linkage disequilibrium (LD) with causal variants [43], and/or interactions of the variant with enriched genes [188]. One popular strategy to mitigate this drawback is to impose a sparse prior over β given the set of causal variants:

$$\beta \sim N(0, \frac{1}{\tau}\sigma^2\boldsymbol{\Sigma}_C) \tag{5.2}$$

$$\boldsymbol{\Sigma}_C(i,j) = \begin{cases} 0, & i \neq j \\ 1, & i = j \text{ and } i \text{ is causal} \\ \epsilon, & i = j \text{ and } i \text{ is non-causal with non-zero effect} \\ 0, & \text{otherwise} \end{cases} \tag{5.3}$$

Notice from Eq. (5.3) that the variance of $\beta(i)$ for a causal variant is $\frac{\sigma^2}{\tau}$ and the variance of $\beta(i)$ for a non-causal variant with non-zero effect is $\epsilon\frac{\sigma^2}{\tau}$, where ϵ is assumed to be small. This formulation handles residual influences from the non-causal variants, which are often observed in real-world data. Under this assumed prior, we can show [86, 189] that the normalized GWAS effect sizes $\boldsymbol{z}$ are distributed as:

$$p\left(\boldsymbol{z}|\boldsymbol{\Sigma}_X, \boldsymbol{\Sigma}_C\right) = N\left(\boldsymbol{z}; 0, \boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_X\left(n\,\sigma^2\,\boldsymbol{\Sigma}_C\right)\boldsymbol{\Sigma}_X\right) \tag{5.4}$$

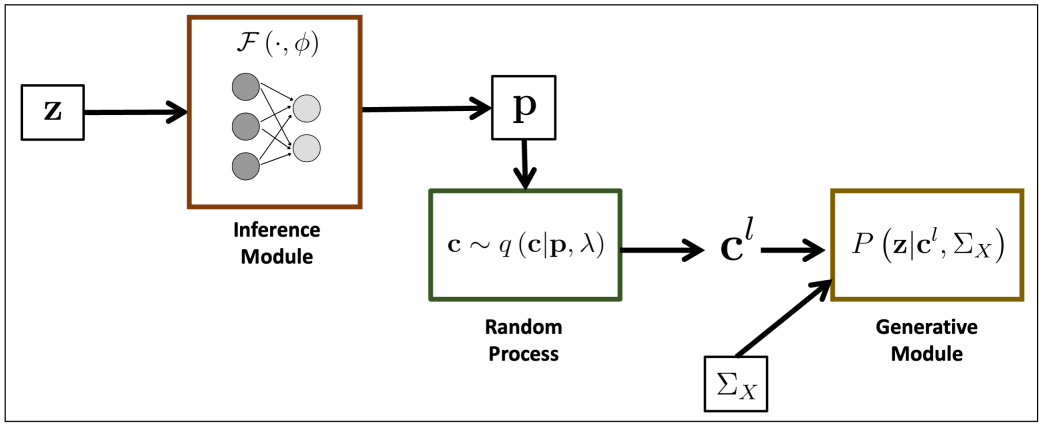


Figure 5-1. Overview of BEATRICE . The inputs to our framework are the LD matrix $\boldsymbol{\Sigma}_X$ and the summary statistics $\boldsymbol{z}$. The inference module uses a neural network to estimate the underlying probability map $\boldsymbol{p}$. The random process generates random samples $\boldsymbol{c}^l$ for the Monte Carlo integration in Eq. (5.12). Finally, the generative module calculates the likelihood of the summary statistics from the sample causal vectors $\boldsymbol{c}^l$.

where $\boldsymbol{\Sigma}_X = \frac{1}{n}\boldsymbol{G}^T\boldsymbol{G}$ is the empirical correlation matrix of the genotype data, also known as the LD matrix. Broadly, the goal of fine-mapping is to identify the diagonal elements of Σ_C that corresponds to 1 given the effect sizes $\boldsymbol{z}$ and the LD matrix $\boldsymbol{\Sigma}_X$. The derivation is provided in Section 2.2.3.2.

5.1.3 The Deep Bayesian Variational Model

BEATRICE uses a variational inference framework for fine-mapping. For convenience, we represent the diagonal elements of $\boldsymbol{\Sigma}_C$ by the vector $\boldsymbol{c} \in \mathbb{R}^{m\times 1}$, and by construction, $\boldsymbol{c}$ encodes the causal variant locations. Fig. 5-1 provides an overview of BEATRICE . Our framework consists of three main components: an inference module, a random sampler, and a generative module. The inputs to BEATRICE are the summary statistics $\boldsymbol{z}$ and the LD matrix $\boldsymbol{\Sigma}_X$. The inference module estimates the parameters $\boldsymbol{p}$ of our proposal distribution $q(\cdot; \boldsymbol{p}, \lambda)$ using a neural network. The random process sampler uses the parameters $\boldsymbol{p}$ to randomly sample potential causal vectors $\boldsymbol{c}$ according to the given proposal distribution. Finally, the generative module calculates the likelihood of the observed summary statistics $\boldsymbol{z}$ according to Eq. (5.4).

5.1.3.1 Proposal Distribution

The goal of fine-mapping is to infer the posterior distribution $p(\boldsymbol{c}|\{\mathbf{z}, \boldsymbol{\Sigma}_{\mathbf{X}}\})$, where $\boldsymbol{c}$ corresponds to the diagonal elements of $\boldsymbol{\Sigma}_C$. Due to the prior formulation in Eqs. (5.2-5.3), solving for the true posterior distribution is computationally intractable, as it requires a combinatorial search over the possible causal configurations. Thus, we approximate the posterior distribution $p(\boldsymbol{c}|\{\mathbf{z}, \boldsymbol{\Sigma}_{\mathbf{X}}\})$ with a binary concrete distribution $q(\boldsymbol{c}; \boldsymbol{p}, \lambda)$ [178], where the parameters $\boldsymbol{p}$ of the distribution are functions of the inputs $\{\mathbf{z}, \boldsymbol{\Sigma}_{\mathbf{X}}\}$. Samples $\boldsymbol{c}$ generated under a binary concrete distribution can be viewed as continuous relaxations of independent Bernoulli random variables. This reparametrization [179] allows us to learn $\boldsymbol{p}$ from the data using standard gradient descent.

Formally, let $\boldsymbol{c}_i$ and $\boldsymbol{p}_i$ denote the ith element of the vectors $\boldsymbol{c}$ and $\boldsymbol{p}$, respectively. Each entry of $\boldsymbol{c}$ is independent and is distributed as follows:

$$q(\boldsymbol{c}_i; \boldsymbol{p}_i, \lambda) = \frac{\lambda \boldsymbol{p}_i \boldsymbol{c}_i^{-\lambda-1}(1-\boldsymbol{p}_i)(1-\boldsymbol{c}_i)^{-\lambda-1}}{\left(\boldsymbol{p}_i \boldsymbol{c}_i^{-\lambda} + (1-\boldsymbol{p}_i)(1-\boldsymbol{c})^{-\lambda}\right)^2}, \tag{5.5}$$

where the parameter λ controls the extent of relaxation from a Bernoulli distribution.

We can easily sample from the binary concrete distribution in Eq. (5.5) via

$$\boldsymbol{c}_i = \xi\left(\frac{\log\left(\frac{U}{1-U}\right) + \log\left(\frac{\boldsymbol{p}_i}{1-\boldsymbol{p}_i}\right)}{\lambda}\right), \tag{5.6}$$

where $\xi(\cdot)$ is the sigmoid function, and the random variable U is sampled from a uniform distribution over the interval $[0, 1]$. As seen, $\boldsymbol{p}_i$ specifies the underlying probability map and U provides stochasticity for the sampling procedure in Eq. (5.6). We note that the gradient of Eq. (5.6) with respect to $\boldsymbol{p}_i$ tends to have a low variance in practice, which helps to stabilize the optimization. Fig. 5-2 shows the change in binary concrete values ($\boldsymbol{c}_i$) with varying degree of U, $\boldsymbol{p}_i$ and λ. As shown in Fig. 5-2(a) smaller values of λ lead to progressively discretized $\boldsymbol{c}_i$, while larger values provide a smoother mapping. In Fig. 5-2(b) we show how the joint relationship between $\boldsymbol{p}_i$

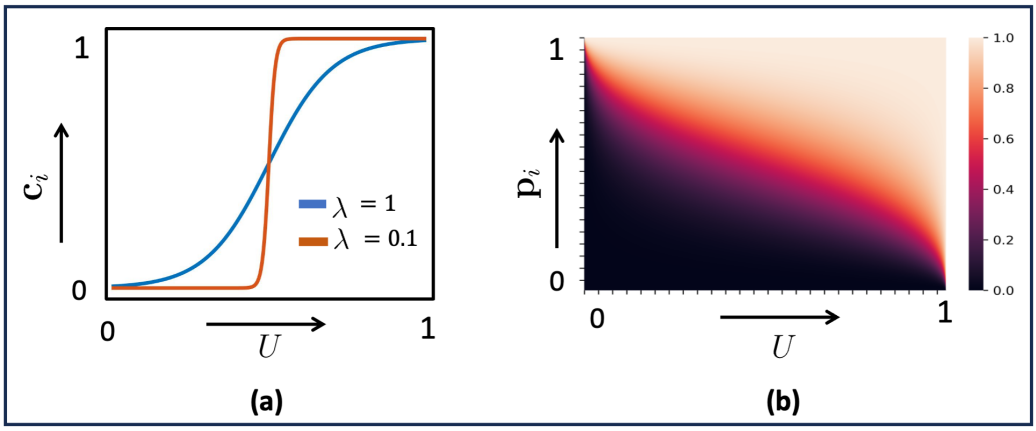


Figure 5-2. Properties of the binary concrete distribution. (a) Relationship between $\boldsymbol{c}_i$ and U for different values of λ. (b) The change in $\boldsymbol{c}_i$ for varying probability map value $\boldsymbol{p}_i$ and uniform noise U. The darker and brighter colors represents $\boldsymbol{c}_i$ close to 0 and 1, respectively.

and the sampled uniform random variable U generates the binary concrete random variable $\boldsymbol{c}_i$. As seen, higher values of $\boldsymbol{p}_i$ push $\boldsymbol{c}_i$ closer to 1, irrespective of the uniform random variable.

Intuitively, every element of the binary concrete random vector $\boldsymbol{c}$ can be regarded as a continuous relaxation from a Bernoulli random (Fig. 5-2(a)). Specifically, the parameter $\boldsymbol{p}$ captures the underlying probability map, analogous to the selection probability of a Bernoulli distribution. The parameter λ controls the extent of relaxation from the 0/1 Bernoulli distribution, such that increasing λ results in a smoother transition between the extremal values $\{0, 1\}$. This continuous representation allows us to model the infinitesimal effects of the non-causal variants. Additionally, the underlying probability map $\boldsymbol{p}$ captures the relative importance of a variant containing a causal signal. The two unique properties of the probability maps are $P(\boldsymbol{c}_i > \frac{1}{2}) = \boldsymbol{p}_i$ and $\lim_{\lambda \to 0} P(\boldsymbol{c}_i = 1) = \boldsymbol{p}_i$. The first property indicates that $\boldsymbol{p}_i$ controls the degree to which $\boldsymbol{c}_i$ assumes low values close to 0 and high values close to 1. This property also give BEATRICE flexibility to handle genetic variants with different levels of association, thus aligning with our generative process that assumes some non-causal variants may have small, non-zero effects. The second property implies that a high

probability $\boldsymbol{p}_i$ (Fig. 5-2(b)) at location i is highly indicative of a causal variant. Taken together, the binary concrete distribution has an easily-optimized parameterization with desirable properties.

5.1.3.2 Variational Inference

We select the variational parameters $\{\boldsymbol{p}, \lambda\}$ to minimize the Kullback–Leibler (KL) divergence between the proposal distribution and the posterior distribution of the causal vector $\boldsymbol{c}$ given the input data $\{\mathbf{z}, \boldsymbol{\Sigma}_{\mathbf{X}}\}$, that is

$$\{\boldsymbol{p}^*, \lambda^*\} = \arg\min_{\{\boldsymbol{p},\lambda\}} KL\left(q(\boldsymbol{c}; \boldsymbol{p}, \lambda) \,||\, p(\boldsymbol{c}|\{\mathbf{z}, \boldsymbol{\Sigma}_{\mathbf{X}}\})\right) \tag{5.7}$$

Using Bayes' Rule, we can show that the optimization in Eq. (5.7) can be rewritten

$$\{\boldsymbol{p}^*, \lambda^*\} = \arg\min_{\{\boldsymbol{p},\lambda\}} KL\left(q(\boldsymbol{c}; \boldsymbol{p}, \lambda) \,||\, p(\boldsymbol{c}; \boldsymbol{p}_0, \lambda_0)\right)) - E_{q(\cdot;\boldsymbol{p},\lambda)}\left[\log\left(p(\boldsymbol{z}|\boldsymbol{\Sigma_X}, \boldsymbol{c})\right)\right], \tag{5.8}$$

where we have assumed an element-wise binary concrete prior $p(\boldsymbol{c}; \boldsymbol{p}_0, \lambda_0)$ over the vector $\boldsymbol{c}$. We fix the relaxation parameter to be small ($\lambda = 0.01$) and the probability map to be uniform $\boldsymbol{p}_0 = \left[\frac{1}{m}, \ldots, \frac{1}{m}\right]^{\mathrm{T}}$. Thus, the first term of Eq. (5.8) can be viewed as a regularizer that encourages sparsity in causal vectors $\boldsymbol{c}$. The second term of Eq. (5.8) can be interpreted as the likelihood of the observed test statistics. The works of [190, 191] have demonstrated that under certain assumptions, the likelihood term of the summary statistics is the same as the original data likelihood $p\left(\boldsymbol{y}|\boldsymbol{G}, \boldsymbol{c}\right)$ derived from Eq. (5.1).

During optimization, the relaxation parameter λ is annealed [178, 179] to a small non-zero value (0.01) with a fixed constant rate, and the underlying probability map $\boldsymbol{p}$ is optimized using gradient descent. Specifically, we use a neural network to generate the vector $\boldsymbol{p} = \mathcal{F}(\boldsymbol{z}; \phi)$. The details of the neural network architecture are provided Fig. 5-3. The neural network estimates the parameters $\boldsymbol{p}$ of our proposal distribution $q\left(\cdot; \boldsymbol{p}, \lambda\right)$ using backpropagation during optimization. Practically speaking, the neural network ties the input data $\{\boldsymbol{z}, \Sigma_X\}$ to the parameter space of the proposal distribution

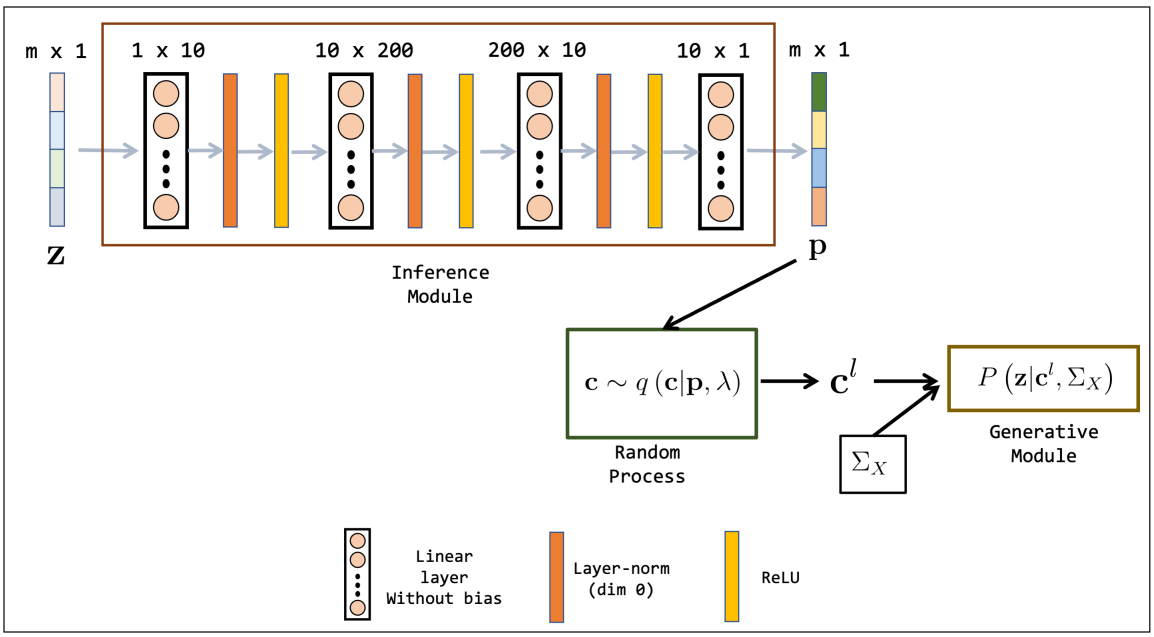


Figure 5-3. Neural network architecture for the inference module used in BEATRICE . The neural network uses a sequence of linear layers, layer normalization, and activation layers. The dimensions of the linear layers are shown on top of each layer. The input to the inference module is the normalized z-scores obtained from GWAS. The output of the inference module is the estimated parameters of our binary concrete distribution.

in a data-driven fashion. Empirically, we find that generating $\boldsymbol{p}$ as a function of the input data regularizes the model and leads to a stable optimization.

Optimizing $\boldsymbol{p}^*$ now amounts to learning the parameters of the neural network ϕ. Given a fixed value of λ, the neural network loss function follows from Eq. (5.8) according to

$$\mathcal{L}(\phi) = KL\left(q(\boldsymbol{c}; \boldsymbol{p}(\phi), \lambda) \,||\, p(\boldsymbol{c}; \boldsymbol{p}_0, \lambda_0))\right) - E_{q(\cdot;\boldsymbol{p}(\phi),\lambda)}\left[\log\left(p(\boldsymbol{z}|\boldsymbol{\Sigma_X}, \boldsymbol{c})\right)\right], \tag{5.9}$$

where we have defined $\boldsymbol{p}(\phi) \triangleq \mathcal{F}(\boldsymbol{z}; \phi)$ for notational convenience.

5.1.3.3 Optimization Strategy

The expectations in Eq. (5.9) do not have closed-form expressions. Therefore, we use Monte Carlo integration to accurately approximate $\mathcal{L}(\phi)$ in the regime of small λ, i.e., when the binary concrete distribution behaves similar to a Bernoulli distribution.

Let $\boldsymbol{c}^1(\phi), \ldots, \boldsymbol{c}^L(\phi)$ be a collection of causal vectors sampled independently from $q(\cdot|\boldsymbol{p}(\phi), \lambda)$ according to Eq. (5.6). The likelihood term of Eq. (5.9) is computed as

$$E_{q(\cdot;\boldsymbol{p}(\phi),\lambda)}\left[\log\left(p(\boldsymbol{z}|\boldsymbol{\Sigma_X}, \boldsymbol{c})\right)\right] = \frac{1}{L}\sum_{l=1}^{L}\log\left(p\left(\boldsymbol{z}|\boldsymbol{\Sigma_X}, \boldsymbol{c}^l(\phi)\right)\right), \tag{5.10}$$

where the right-hand side probability is computed according to Eq. (5.4) by substituting $\boldsymbol{c}^l(\phi)$ for the diagonal entries of $\boldsymbol{\Sigma}_C$ in each term of the summation. Once again, the continuous relaxation used to generate $\boldsymbol{c}^l(\phi)$ in Eq. (5.6) allows us to directly optimize ϕ.

We approximate the first term of Eq. (5.9) under the assumption of small $\{\lambda, \lambda_0\}$ on the order of 0.01. In this case, the binary concrete distribution behaves like a $\{0, 1\}$ Bernoulli distribution. Under these conditions, we can write the first term of Eq. (5.9) as

$$\begin{aligned} &KL\left(q(\boldsymbol{c}; \boldsymbol{p}(\phi), \lambda) \,||\, p(\boldsymbol{c}; \boldsymbol{p}_0, \lambda_0))\right) \\ &\quad\approx \sum_{i=1}^{m}\left[\boldsymbol{p}_i(\phi)\log\left(\frac{\boldsymbol{p}_i(\phi)}{p_0}\right) + (1-\boldsymbol{p}_i(\phi))\log\left(\frac{1-\boldsymbol{p}_i(\phi)}{1-p_0}\right)\right], \end{aligned} \tag{5.11}$$

where p_0 is a fixed scalar parameter used to construct the (constant) prior vector $\boldsymbol{p}_0$. We note that the criteria $\{\lambda \to 0.01, \lambda_0 = 0.01\}$ is satisfied in practice, as λ is annealed during the optimization to progressively smaller values and λ_0 is fixed *a priori*.

The above approximations allow us to rewrite the neural network loss as

$$\begin{aligned} \mathcal{L}(\phi) \approx &-\frac{1}{L}\sum_{l=1}^{L}\log N\left(\boldsymbol{z}; 0, \boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_X\left(n\,\sigma^2\,\boldsymbol{\Sigma}_C^l(\phi)\right)\boldsymbol{\Sigma}_X\right) \\ &+ \sum_{i=1}^{m}\boldsymbol{p}_i(\phi)\log\left(\frac{\boldsymbol{p}_i(\phi)}{p_0}\right) + (1-\boldsymbol{p}_i(\phi))\log\left(\frac{1-\boldsymbol{p}_i(\phi)}{1-p_0}\right), \end{aligned} \tag{5.12}$$

where $\boldsymbol{\Sigma}_C^l(\phi)$ corresponds to the diagonal matrix using the vector $\boldsymbol{c}^l(\phi)$ as the diagonal entries. We use a stochastic gradient descent optimizer [192] to minimize the loss $\mathcal{L}(\phi)$ with respect to the neural network weights ϕ. This process is detailed in Algorithm 3.

Algorithm 3 Optimization scheme to minimize Eq. (5.12)

$\mathcal{B}^R = \{\}$
Initialize ϕ_0
for $t = [1 \dots T]$ **do**
  Generate $\boldsymbol{p}(\phi_t) = \mathcal{F}(\boldsymbol{z}; \phi_t)$
  Randomly sample $\boldsymbol{c}_t^l$ according to Eq. (5.6)
  Binarize $\boldsymbol{c}_t^l$ to $\boldsymbol{b}_t^l$ and add to $\mathcal{B}^R$
  $\mathcal{S}_t^l = \{i\} \;\; s.t. \;\; \boldsymbol{c}_t^l(i) > 0.01$
  Prune set $\mathcal{S}_t^l$ such that it consists of 50 indices.
  $\boldsymbol{c}_t^l(i) = 0$ if $i \notin \mathcal{S}_t^l$
  Generate $\mathcal{L}(\phi_t)$ according to Eq. (5.12)
  $\phi_{t+1} = \phi_t - StepSize \nabla \mathcal{L}(\phi^t)$

5.1.3.4 Computational Complexity

Each iteration of stochastic gradient descent requires us to compute the data log-likelihood term $\left[\log N\left(\boldsymbol{z}; 0, \boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_X\left(n\,\sigma^2\,\boldsymbol{\Sigma}_C^l(\phi)\right)\boldsymbol{\Sigma}_X\right)\right]$. This computation is expensive due to the covariance matrix inversion, whose run-time is on the order of $O(m^3)$, where m is the total number of variants. To mitigate this issue, the works of [193] show that if $\boldsymbol{\Sigma}_C^l(\phi)$ is sparse, then the matrix inversion can be done with order $O(k^3) + O(mk^2)$ run-time, where k is the number of non-zero diagonal elements of $\boldsymbol{\Sigma}_C^l(\phi)$. We leverage this result in the optimization by thresholding the elements of $\boldsymbol{c}^l(\phi)$ to set small values exactly to zero. In every iteration, we sparsify $\boldsymbol{c}_t^l$ by considering the top 50 non-zero locations of $\boldsymbol{c}_t^l$ with values $\boldsymbol{c}_t^l(i) > 0.01$. This strategy provides a way to optimize the parameters of our models in $O(50^3) + O(m50^2)$ run-time for all scenarios. We also regularize $\boldsymbol{\Sigma}_X$ with a small diagonal load to ensure invertibility of the covarance matrix at each iteration. Finally, we run stochastic gradient descent with a batch size of one to further speed up BEATRICE. Effectively, this means that we sample a single $\boldsymbol{c}^l(\phi)$ at each epoch rather than perform a true Monte Carlo integration. The authors of [192] have previously shown that a single random sample ($L = 1$) is sufficient to guarantee convergence to a local minimum of Eq. (5.12). Algorithm 3 provides a detailed description of these optimization steps.

5.1.4 Verification and Comparison

5.1.4.1 Causal Configurations and Posterior Inclusion Probabilities

The desired outputs of each fine-mapping method are Posterior Inclusion Probabilities (PIPs) and credible sets. PIPs estimate how likely each variant is causal as a measure of its importance. Credible sets identify the subset of variants that are likely to contain a causal variant, which captures the uncertainty of finding the true variant.

The main challenge to estimating the posterior probability of a given causal configuration (i.e., set of causal variant locations) is the exponentially large search space. Let $\boldsymbol{b}$ denote a binary vector with a value of 1 at causal locations and a value of 0 at non-causal locations. At a high level, $\boldsymbol{b}$ can be viewed as a binarized version of the causal vector $\boldsymbol{c}$ in the previous sections. Using Bayes' Rule, the posterior probability of $\boldsymbol{b}$ given the input data $\{\mathbf{z}, \mathbf{\Sigma_X}\}$ can be written as follows:

$$p\left(\boldsymbol{b}|\boldsymbol{z}, \mathbf{\Sigma}_X\right) = \frac{p\left(\boldsymbol{z}|\mathbf{\Sigma}_X, \boldsymbol{b}\right) p\left(\boldsymbol{b}\right)}{\sum_{\boldsymbol{b}' \in \mathcal{B}} p\left(\boldsymbol{z}|\mathbf{\Sigma}_X, \boldsymbol{b}'\right) p\left(\boldsymbol{b}'\right)} \tag{5.13}$$

where $\mathcal{B}$ is the set of all 2^m possible causal configurations. Once again, $\boldsymbol{z}$ captures the summary statistics and $\mathbf{\Sigma}_X$ is the LD matrix. Even though $\mathcal{B}$ is exponentially large, it has been argued [194] that the majority of these configurations have negligible probability and do not contribute to the denominator of Eq. (5.13).

Our stochastic optimization provides a natural means to track causal configurations with non-negligible probability to compute $p\left(\boldsymbol{b}|\boldsymbol{z}, \mathbf{\Sigma}_X\right)$. Namely, at each iteration of stochastic gradient descent, we randomly generate a sample causal vector $\boldsymbol{c}^l$ to minimize Eq. (5.12). In parallel, we binarize the vector $\boldsymbol{c}^l$ via

$$\boldsymbol{b}_i^l = \begin{cases} 1, & \boldsymbol{c}_i^l > \gamma, \\ 0, & otherwise \end{cases}$$

and add the resulting vector $\boldsymbol{b}^l$ to a reduced set of causal configurations $\mathcal{B}^R$. The variational objective ensures that our proposal distribution converges to the true

posterior distribution of the causal vectors. Thus, the samples $\boldsymbol{c}^l$ lie near modes of the posterior distribution which is the neighborhood of non-negligible probability.

In this work, we use a threshold $\gamma = 0.1$ to binarize the vectors $\boldsymbol{c}^l$. Thresholding at $\gamma = 0.1$ only considers variants whose estimated effect size variance is $> 0.1\sigma^2$. This operation prunes out spurious non-causal configurations generated by the non-causal variants. A higher threshold is beneficial in presence of high interaction effects from non-causal variants and a lower threshold could be useful when the causal variants are weakly associated with the outcome. Empirically, we find that the threshold value of 0.1 preserves the main interactions between variants. However, the user of BEATRICE can adjust this threshold as needed.

After obtaining the sampled vectors, we replace the exhaustive set $\boldsymbol{B}$ in Eq. (5.13) with the reduced set $\mathcal{B}^R$ for tractable computation of $p\left(\boldsymbol{b}|\boldsymbol{z}, \boldsymbol{\Sigma}_X\right)$. We then compute the posterior inclusion probability (PIP) of each variant by summing the probabilities over the subset of $\mathcal{B}^R$ with a value of 1 at that variant location. Mathematically,

$$P(\boldsymbol{b}_i = 1|\boldsymbol{z}, \boldsymbol{\Sigma}_X) \approx \sum_{\boldsymbol{b}\in\mathcal{S}} p\left(\boldsymbol{b}|\boldsymbol{z}, \boldsymbol{\Sigma}_X\right) \tag{5.14}$$

$$\text{s.t. } \mathcal{S} \subset \mathcal{B}^R \text{ and } \mathcal{S} = \{\boldsymbol{b} \,|\, \boldsymbol{b}_i = 1\} \tag{5.15}$$

where $\mathcal{S}$ is a subset of $\mathcal{B}^R$ that contains binary configurations with 1 at location i.

5.1.4.2 Identification of Credible Sets for BEATRICE

One of the notable features of BEATRICE is its ability to identify a comprehensive set of causal configurations with non-negligible posterior probability within the exponentially large search space. As described in the previous section, the reduced search space $\mathcal{B}^R$ is comprised of vectors that BEATRICE randomly samples at each iteration of the optimization. We identify credible sets from $\mathcal{B}^R$ in two steps.

First, in a sequential fashion, we identify the "key" variant with the highest conditional probability given the previously selected variants. Formally, let $\mathcal{K}$ be the

Algorithm 4 Algorithm to find credible sets

$\mathcal{K} = \{\}$
$\mathcal{CS} = \{\}$
Estimate posterior probabilities accoding to Eq. 5.16.
while $max\left[P(\boldsymbol{b}_i|\mathcal{K}, \boldsymbol{z}, \boldsymbol{\Sigma}_X) \mid \forall i \notin \mathcal{K}\right] < \gamma_{key}$ **do**
  $\mathcal{K} = \mathcal{K} \cup argmax\left[P(\boldsymbol{b}_i|\mathcal{K}, \boldsymbol{z}, \boldsymbol{\Sigma})_X \mid \forall i \notin \mathcal{K}\right]$
  Estimate posterior probabilities accoding to Eq. 5.16.
for $k \in \mathcal{K}$ **do**
  $\mathcal{S} = \{\}$
  $cov = 0$
  Generate $\mathcal{K}'$ by removing k for "key" set.
  for $i = [1, \ldots, m]$ and $i \notin \mathcal{K}'$ **do**
    Estimate posterior probability according to Eq. 5.17
    Stack the probabilty scores in a vector $\mathcal{P}$.
  while $max\left[\mathcal{P}_i \mid \forall i \notin \mathcal{K}'\right] > \gamma_{selection}$ and $cov < \gamma_{coverage}$ **do**
    $\mathcal{S} = \mathcal{S} \cup argmax\left[\mathcal{P}_i \mid \forall i \notin \mathcal{K}'\right]$
    $cov = cov + max\left[\mathcal{P}_i \mid \forall i \notin \mathcal{K}'\right]$
  Add $\mathcal{S}$ to $\mathcal{CS}$ as credible set of k.

indices of previously identified "key" variants. The conditional probability for variant i given $\mathcal{K}$ in each iteration can be calculated as follows:

$$P\left(\boldsymbol{b}_i = 1|\mathcal{K}, \boldsymbol{z}, \boldsymbol{\Sigma}_X\right) = \frac{\sum_{\boldsymbol{b}\in\mathcal{C}} P\left(\boldsymbol{b}|\boldsymbol{z}, \boldsymbol{\Sigma}_X\right)}{\sum_{\boldsymbol{b}'\in\mathcal{D}} P\left(\boldsymbol{b}'|\boldsymbol{z}, \boldsymbol{\Sigma}_X\right)} \tag{5.16}$$

$$\text{s.t. } \mathcal{D} \subset \mathcal{B}^R \text{ and } \mathcal{D} = \{\boldsymbol{b}|\boldsymbol{b}_j = 1 \ \forall \ j \in \mathcal{K}\}$$

$$\mathcal{C} \subset \mathcal{B}^R \text{ and } \mathcal{C} = \{\boldsymbol{b}|\boldsymbol{b}_j = 1 \ \forall \ j \in \{i\} \cup \mathcal{K}\}$$

where, $\mathcal{D}$ is the subset of $\mathcal{B}^R$ that includes all of "key" variants and $\mathcal{C}$ is the subset of $\mathcal{B}^R$ that includes both variant i and the "key" variants.

We perform this sequential variant selection until the maximum posterior probability reduces below a threshold, which we define as the "key" threshold γ_{key} and fix at $\gamma_{key} = 0.2$ for all experiments. We note that this threshold can be controlled by the user. The selected "key" variants act as proxy for highly plausible causal variants.

In the second step, we identify the set of variants that can replace the "key" variant in the causal configurations while maintaining a high posterior probability. This set of variants act as a credible set for that particular "key" variant. To do this, we first

remove one of the key variants from $\mathcal{K}$ and estimate the posterior probability of other variants given the remaining "key" variants. For example, let variant k_1 be a " key" variant. We estimate the posterior probabilities as follows:

$$P\left(\boldsymbol{b}_i = 1 | \mathcal{K}', \boldsymbol{z}, \boldsymbol{\Sigma}_X\right) = \frac{\sum_{\boldsymbol{b} \in \mathcal{G}} P\left(\boldsymbol{b} | \boldsymbol{z}, \boldsymbol{\Sigma}_X\right)}{\sum_{\boldsymbol{b}' \in \mathcal{H}} P\left(\boldsymbol{b}' | \boldsymbol{z}, \boldsymbol{\Sigma}_X\right)} \tag{5.17}$$

$$\text{s.t. } \mathcal{K} = \mathcal{K}' \cup \{k_1\}$$

$$\mathcal{G} = \{\boldsymbol{b} | \boldsymbol{b}_j = 1 \ \forall \ j \in \{i\} \cup \mathcal{K}'\}$$

$$\mathcal{H} \subset \mathcal{B}^R \text{ and } \mathcal{H} = \{\boldsymbol{b} | \boldsymbol{b}_j = 1 \ \forall \ j \in \mathcal{K}'\}$$

where $\mathcal{K}'$ is the set of "key" variants without k_1, $\mathcal{G}$ is the set of configurations that include both variant i and the remaining "key" variants, and and $\mathcal{H}$ is the set of causal configurations that include all "key" variants except k_1.

Once computed, we sort these posterior probabilities in descending order and add the variants to the credible set until the cumulative sum reaches the coverage threshold $\gamma_{coverage}$. We fix the coverage threshold at $\gamma_{key} = 0.95$ in this work, but it too can be set by the user. Finally, we prune uncorrelated variants by thresholding the posterior probability according to the selection threshold $\gamma_{selection} = 0.05$, again a tunable parameter for users. Algorithm 4 provides a detailed description of these steps.

5.1.4.3 Baselines

We compare our approach with the state-of-the-art methods, FINEMAP-v1.4.1 and SuSiE-v0.12.27.

FINEMAP: This approach uses a stochastic shotgun search to identify causal configurations with non-negligible posterior probability. FINEMAP defines the neighborhood of a configuration at every step by deleting, changing or adding a causal variant from the current configuration. The next iteration samples from this neighbor-

hood, thus reducing the exponential search space to a smaller high-probability region. The identified causal configurations are used to determine the posterior inclusion probabilities for each variant. In addition, FINEMAP outputs a collection of credible sets under the assumption of multiple causal variants $d = 1, \ldots, D$. Similar to the approach used in [96], we sub-select the credible sets from this collection with the highest posterior probability. From here, we pruned the sets with minimum absolute purity greater that 0.5. As defined in [96], purity is the pairwise correlation coefficient between the variants, obtained from the LD matrix. The computationally efficient shotgun approach makes FINEMAP a viable tool for finemapping from multiple GWAS summary data in [37, 195]. We implement FINEMAP using the stochastic shotgun approach. During implementation, we fix the number of causal variants to 20 and the rest of the hyperparameters are fixed to default values, as described in http://christianbenner.com/

SuSiE: The recent works of [53, 96] introduced an iterative Bayesian selection approach for fine-mapping that represents the variant effect sizes as a sum of "single-effect" vectors. Each vector contains only one non-zero element, which represents the causal signal. In addition to finding causal variants, SuSiE provides a way to quantify the uncertainty of the causal variants locations via credible sets. SuSiE has also been used widely to find putative causal variants GWAS summary statistics [196, 197].

During the implementation of SuSiE, we provide the un-normalized effect sizes (β), the Standard Error (SE) of the effect sizes, the LD matrix, the phenotype variance, and the number of samples. Additionally, we fix the number of causal variants to 20 and we estimate the residual variance. The rest of the hyperparameters are fixed to default values, as described in stephenslab.github.io.

5.1.4.4 Evaluation Strategy

We evaluate several metrics of performance in our simulation study.

Area Under Precision Recall Curve (AUPRC): We have compared the quality of the PIPs via the AUPRC metric. AUPRC (area under the precision-recall curve) is computed by sweeping a threshold on the PIPs and computing precision and recall against the true configuration of causal and non-causal variants. High precision indicates a low false positive rate in the estimated causal variants. High recall indicates that the model correctly identifies more of the causal variants. Thus, the AUPRC, can be viewed as a holistic measure of performance across both classes. AUPRC is also robust to severe class imbalance [198], which is the case in fine-mapping, as the number of causal variants is small.

Coverage, Power and Size of the Credible Sets: We follow the strategy of [53, 96] to define a credible set. A credible set is defined as a collection of variants that contain a single causal variant with a probability equal to the coverage. Given that the number of causal variants can be arbitrary, we use two metrics to assess the quality of the credible sets. Specifically, coverage is the percentage of credible sets that contain a causal variant, and power is the percentage of causal variants identified by *all* the credible sets. Higher coverage indicates that the method is confident about its prediction of *each* causal variant, whereas higher power indicates the method can accurately identify all the causal variants.

One caveat is that a method can generally achieve both higher coverage and higher power simply by adding variants to the credible sets. To counter this trend, we report the average size of the credible sets identified by each method. Ideally, we would like the credible sets to be as small as possible while retaining high coverage and high power.

5.1.5 Applications

5.1.5.1 Experimental Setup

Genotype Simulations: We use the method of [199] to simulate genotypes $\boldsymbol{G}$ based on data from the 1000 Genomes Project. We select an arbitrary sub-region ($39.9Mb - 40.9Mb$) from Chromosome 2 as the base. After filtering for rare variants (MAF < 0.02), the remaining 3.5K variants are used to simulate pairs of haplotypes to generate 10,000 unrelated individuals. We chose a MAF threshold of 0.02, as it lies in the middle of the range $0.01 - 0.05$ commonly used in GWAS studies [200]. In each experiment below, we randomly select $m = 1000$ variants and $n = 5000$ individuals to generate the phenotype data.

Phenotype Generation: We generate the phenotype $\boldsymbol{y}$ from a standard mixed linear model [190], where the influences of the causal variants are modeled as fixed effects, and the influences of other non-causal variants are modeled as random effects. In this case, the genetic risk for a trait is spread over the entire dataset, with each variant having small individual effects, as per the polygenicity assumption of a complex trait. We randomly select the causal variants in our simulations. Thus, some simulations will have causal variants in LD, while others will select causal variants with low correlation.

Given a set of d causal variants C, let $\boldsymbol{G}_C \in \boldsymbol{R}^{n \times d}$ denote the corresponding subset of the genotype data and $\boldsymbol{G}_{NC} \in \boldsymbol{R}^{n \times (m-d)}$ denote the remaining non-causal variants.

From here, we generate the phenotype data $\boldsymbol{y}$ as follows:

$$\begin{aligned}
\boldsymbol{y} &= \boldsymbol{G}_C\,\beta + \boldsymbol{g}_{NC} + \epsilon \triangleq \boldsymbol{g}_C + \boldsymbol{g}_{NC} + \epsilon \\
\boldsymbol{g}_{NC} &\sim N\left(0, \frac{1}{m-d}\,\boldsymbol{G}_{NC}\,\boldsymbol{G}_{NC}^T\right) \\
\beta &\sim N\left(0, \mathbb{I}_d\right) \\
\epsilon &\sim N\left(0, \alpha^2\mathbb{I}_n\right)
\end{aligned}$$

where β is the d-dimensional effect sizes sampled from a Gaussian, and ϵ is an zero-mean Gaussian noise with variance α^2. The random variable $\boldsymbol{g}_{NC}$ models the effect of the non-causal variants as a multivariate Gaussian vector with mean 0 and covariance $\frac{1}{m-d}\boldsymbol{G}_{NC}\boldsymbol{G}_{NC}^T$. Likewise, $\boldsymbol{g}_C = \boldsymbol{G}_C\,\beta$ captures the effect of the causal variants.

In our experiments, we define ω^2 as the total phenotypic variance attributed to the genotype (e.g., both $\boldsymbol{g}_C$ and $\boldsymbol{g}_{NC}$) and p as the proportion of this variance associated with the causal variants in $\boldsymbol{g}_C$. Using the strategy described in [201], we enforce these conditions by normalizing the phenotype $\boldsymbol{y}$ as follows:

$$\begin{aligned}
\tilde{\boldsymbol{y}} &= \sqrt{\frac{p\,\omega^2}{var(\boldsymbol{g}_C)}}\,\boldsymbol{g}_C + \sqrt{\frac{(1-p)\,\omega^2}{var(\boldsymbol{g}_{NC})}}\,\boldsymbol{g}_{NC} + \tilde{\epsilon} \\
\tilde{\epsilon} &\sim N(0, (1-\omega^2)\mathbf{1}_n)
\end{aligned} \tag{5.18}$$

where $var(\boldsymbol{g}_C)$ and $var(\boldsymbol{g}_{NC})$ are the empirical variances of $\boldsymbol{g}_C$ and $\boldsymbol{g}_{NC}$, respectively.

After generating the genotype $\boldsymbol{G}$ and the normalized phenotype $\tilde{\boldsymbol{y}}$, we run a GWAS to estimate the effect size $\hat{\beta}_i$ of each variant i. From here, we convert the estimated effect sizes to z-scores via $\boldsymbol{z}_i = \frac{\hat{\beta}_i}{se\left(\hat{\beta}_i\right)}$, where $se(\cdot)$ denotes the standard error. The LD matrix is computed from the genotype data as $\boldsymbol{\Sigma}_X = \frac{1}{n}\boldsymbol{G}^T\,\boldsymbol{G}$. The z-scores and LD matrix are input to each of the fine-mapping methods above.

Noise Configurations: We evaluate the performance of each method while varying the number of causal variants d, the total genotype variance ω^2, and the proportion

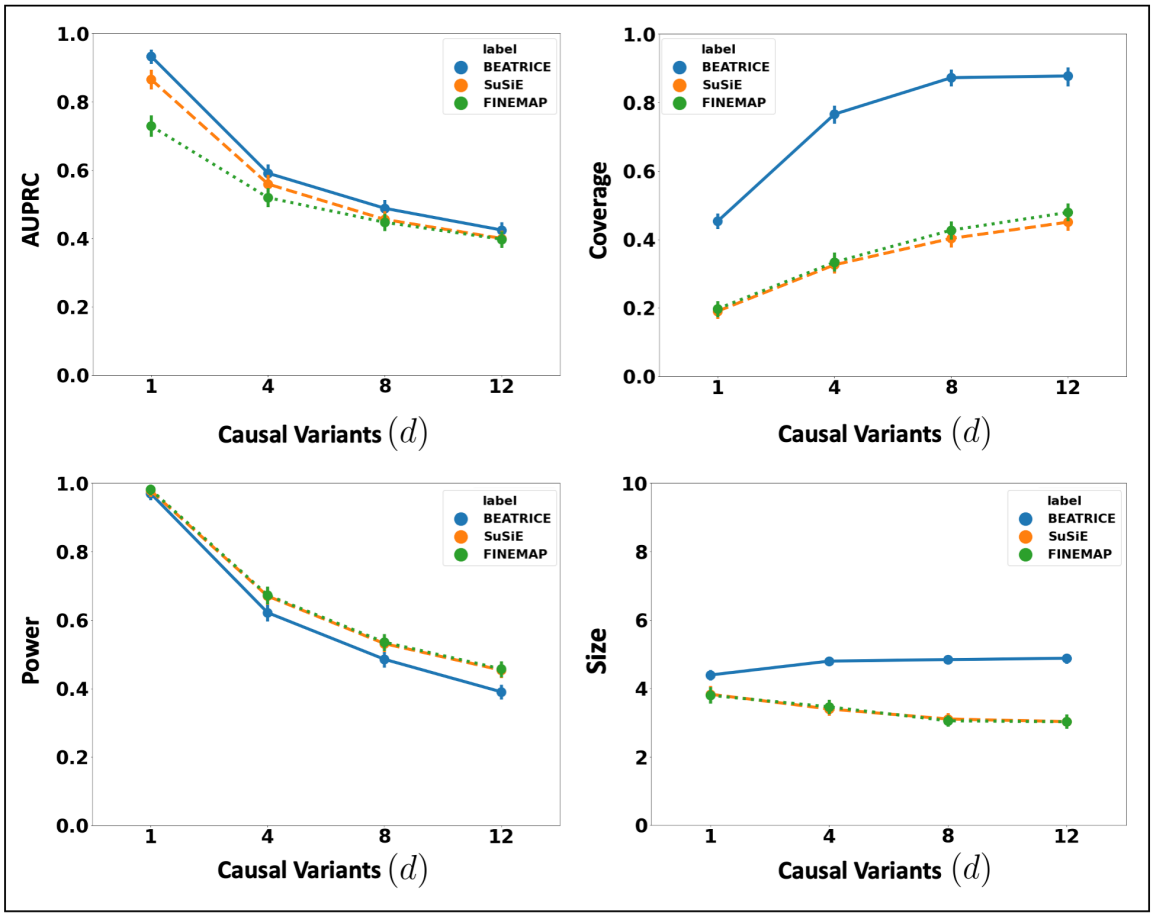


Figure 5-4. The performance metrics for the three methods across varying numbers of causal variants. Along the x-axis, we plot the number of causal variants, and across the y-axis, we plot the mean and confidence interval (95%) of each metric. We calculate the mean by fixing d to a specific value $d = d^*$ and sweep over all the noise settings where $d = d^*$.

of this variance associated with the causal variants p. Formally, we sweep over one order of magnitude for $d = [1, 4, 8, 12]$, $\omega^2 = [0.1, 0.2, 0.4, 0.5, 0.7, 0.8]$, and $p = [0.1, 0.3, 0.5, 0.7, 0.9]$. For each noise setting, we randomly generate 20 datasets by independently re-sampling the causal variant locations, the effect sizes $\{\beta_i\}$, the non-causal component $\boldsymbol{g}_{NC}$, and the noise $\tilde{\epsilon}$. We run all three fine-mapping methods over a total of $4 \times 6 \times 5 \times 20 = 2400$ configurations for a comprehensive evaluation.

5.1.6 Results

Varying the Number of Causal Variants Fig. 5-4 illustrates the performance of each method (BEATRICE , FINEMAP, and SuSiE) while increasing the number of causal variants from $d = 1$ to $d = 12$. The points denote the mean performance across all noise configurations (ω^2, p) for fixed d, and the error bars represent the 95% confidence interval across these configurations. We note that BEATRICE achieves a uniformly higher AUPRC than both baseline method, which suggests that BEATRICE can better estimate the PIPs than FINEMAP or SuSiE. BEATRICE

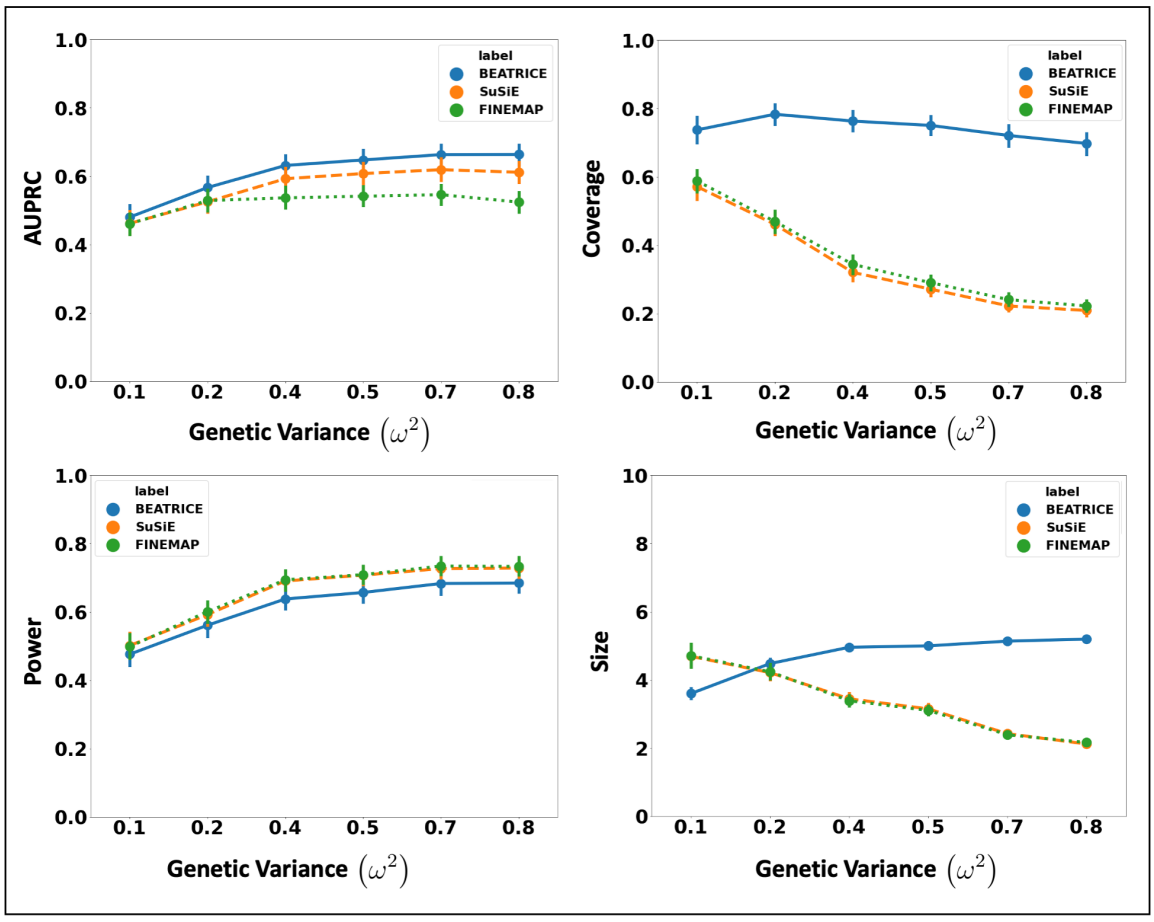


Figure 5-5. The performance metric for increasing phenotype variance explained by genetics. Along the x-axis, we plot the variance explained by genetics (ω^2), and across the y-axis, we plot each metric's mean and confidence interval (95%). We calculate the mean by fixing ω^2 to a specific value $\omega = \omega^*$ and sweep over all the noise settings where $\omega = \omega^*$.

also provides $0.9 - 1.4$ fold increase in coverage than the baselines with similar power, which indicates that the credible sets generated by BEATRICE are more likely to contain a causal variant as compared to SuSiE and FINEMAP. Finally, we note that although FINEMAP and SuSiE identify smaller credible sets, the difference in set size between them and BEATRICE is < 2 variants. Taken together, as the number of causal variants increases, BEATRICE gives us a better estimate of the PIPs and arguably better credible sets. Compared to the baselines BEATRICE does not impose any prior assumptions over the total number of causal variants, which may lead to its improved performance.

Increasing the Genotype Contribution: Fig. 5-5 shows the performance of each method while increasing the genetically-explained variance from $\omega^2 = 0.1$ to $\omega^2 = 0.8$. Similar to above, the points denote the mean performance across all configurations (d, p) for fixed ω^2, and the error bars represent the 95% confidence interval across these configurations. We note that BEATRICE achieves a significantly higher AUPRC than FINEMAP and a slightly higher AUPRC than SuSiE. When evaluating the credible sets, we observe similar trends in coverage (BEATRICE is $0.25 - 2.34$ folds higher) and power (similar performance across methods). While the FINEMAP and SuSiE identify slightly smaller credible, the difference to BEATRICE is only a few variants. Taken together, we submit that BEATRICE achieves the best trade-off across the four performance metrics.

Varying the Contributions of Causal and Non-Causal Variants: Fig. 5-6 illustrates the performance of each method while increasing the contribution of the causal variants from $p = 0.1$ to $p = 0.9$. Once again, the points denote the mean performance across all configurations configurations (d, ω^2) for fixed p, and the error bars represent the 95% confidence interval across these configurations. From an application standpoint, the presence of non-causal variants with small non-zero

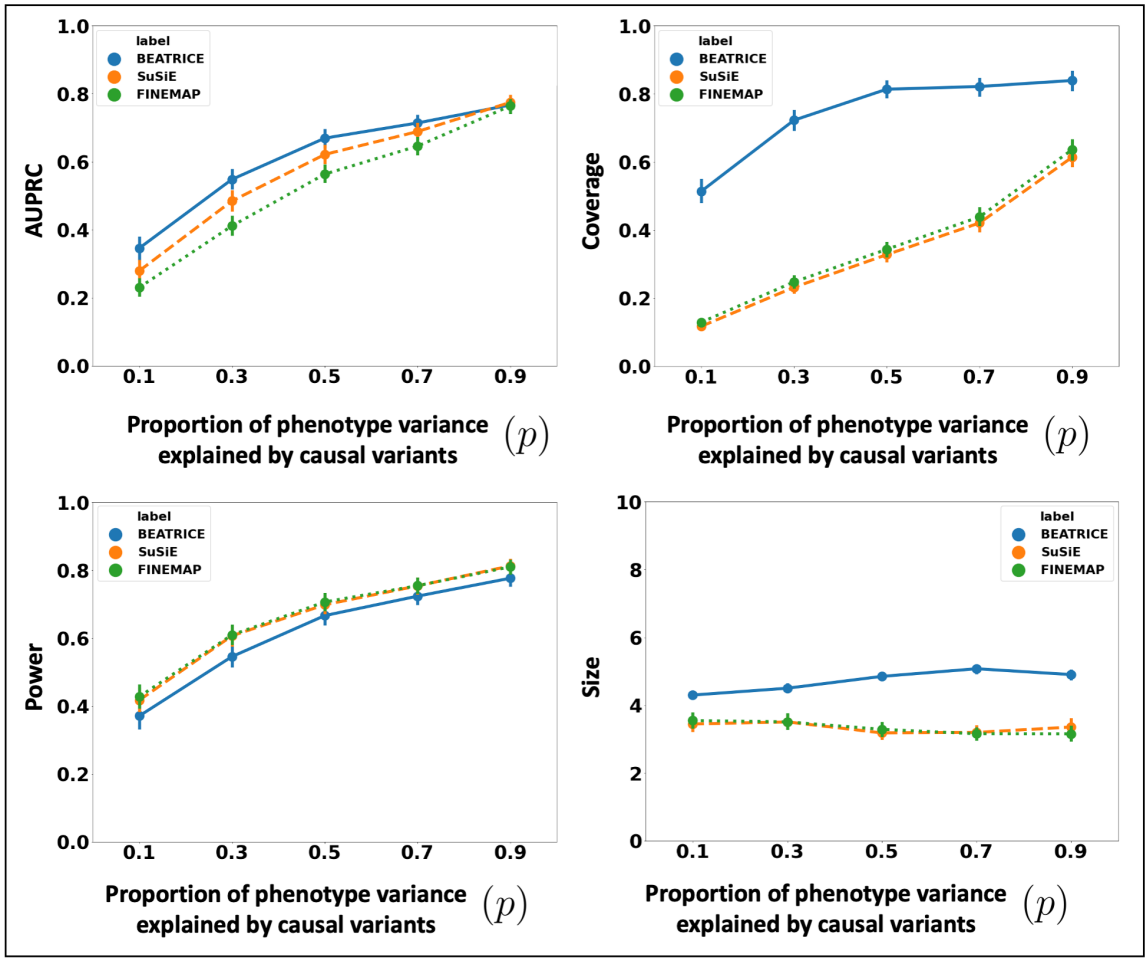


Figure 5-6. The performance metric for multiple levels of noise introduced by non-causal variants. The noise level (p) is explained by the variance ratio of non-causal variants vs. causal variants. Along the x-axis, we plot the noise level (p); across the y-axis, we plot each metric's mean and confidence interval (95%). We calculate the mean by fixing p to a specific value $p = p^*$ and sweep over all the noise settings where $p = p^*$.

effects makes it difficult to detect the true causal variants. Accordingly, we observe a performance boost across all methods when p is larger. Similar to our previous experiments, BEATRICE provides the best AUPRC, with converging performance as $p \to 1$. In addition, BEATRICE identifies better credible sets with significantly higher coverage while maintaining power. Thus, we conclude that BEATRICE is the most robust of the three methods to the presence of noise from non-causal variants. This performance gain may arise from our binary concrete proposal distribution for the causal vector $\boldsymbol{c}$, which provides flexibility to accommodate varying degrees of

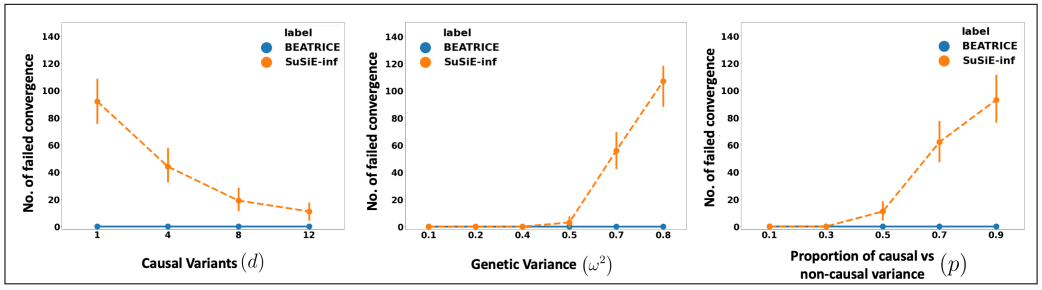


Figure 5-7. Number of non-convergent runs of SuSiE-inf, as compared to BEATRICE.

association.

Additional Comparison with SuSiE-inf: SuSiE-inf is an extension of the SuSiE model that accounts for infinitesimal effects from non-causal variants. In this section, we compare the performance of BEATRICE with SuSiE-inf across the same simulation setting as described in the Section 5.1.5.1 of the main text.

Unlike BEATRICE, we observe that SuSiE-inf fails to converge in multiple cases. Specifically, Fig. 5-7 illustrates the number of experimental settings, for which SuSie-inf fails to converge across each parameter sweep. This problem becomes prominent with increasing SNP heritability, as explained by ω^2. We believe that the SuSiE-inf algorithm, as described in the preprint [202], is still in development with multiple issues with numerical stability.

Fig. 5-8, Fig. 5-9, and Fig. 5-10 shows the performance comparison between BEATRICE and SuSiE-inf. We emphasize that the performance of SuSiE-inf is computed based only on the convergent runs, so these values should be treated as optimistic. In contrast, the performance of BEATRICE is computed across all runs, as we did not face convergence issues with our model. Across different parameter sweeps, we see that the coverage of SuSiE-inf is similar to BEATRICE. However, BEATRICE achieves uniformly better power and AUPRC.

5.1.7 Discussion

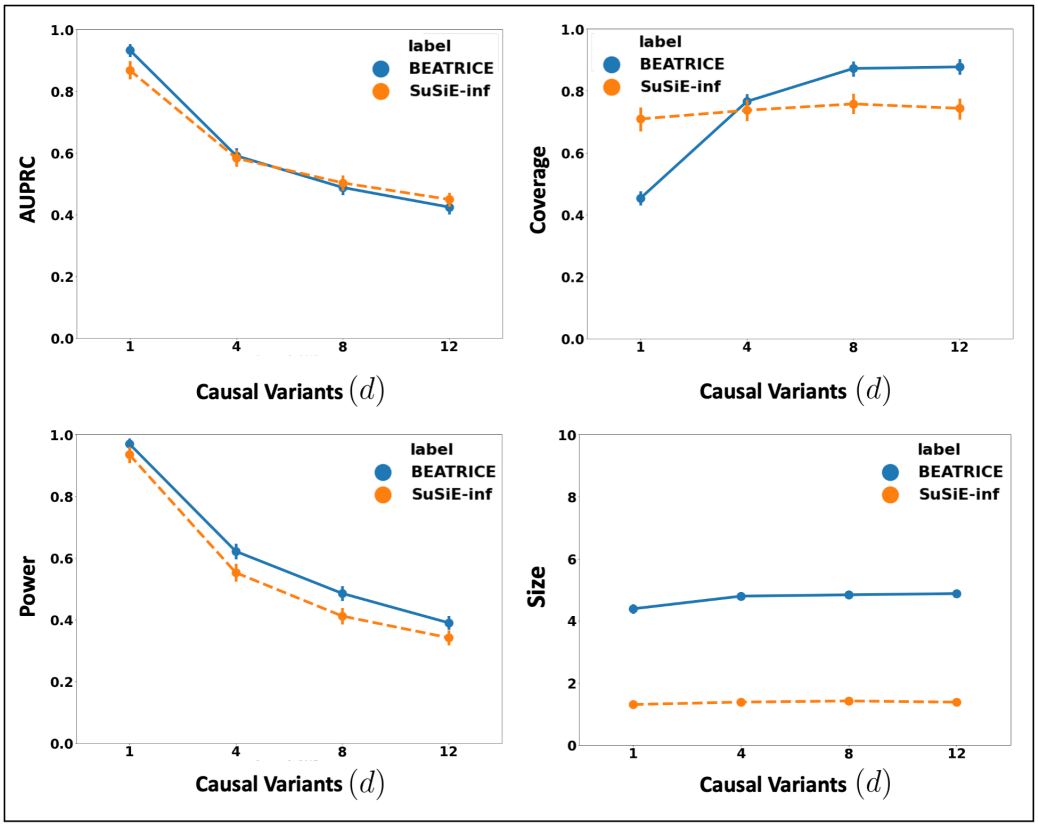


Figure 5-8. Performance metrics of BEATRICE and SuSiE-inf across varying numbers of causal variants. The performance of SuSiE-inf is calculated over the subset of simulation settings in which the algorithm converges; non-convergent settings are omitted from the analysis. The x-axis corresponds to the number of causal variants, and the y-axis plots the mean and confidence interval (95%) of each metric. We calculate the mean by fixing d to a value $d = d^*$ and sweeping over all the noise settings where $d = d^*$.

BEATRICE is a novel, robust, and general purpose tool for fine-mapping that can be used across a variety of studies. One key contribution of BEATRICE over methods like FINEMAP and SuSiE is its ability to discern spurious effects from non-causal variants, including non-causal variants in high LD with true causal variants. Our simulated experiments capture this improved performance by sweeping the proportion of the observed variance attributed to causal (fixed effects) and non-causal (random effects) genetic variants. This parameter $p \in [0, 1]$ is swept over its natural domain, such that $p = 1$ implies that the only link between the genotype and phenotype comes from the causal variants. At this extreme, Fig. 5-6 shows that all methods

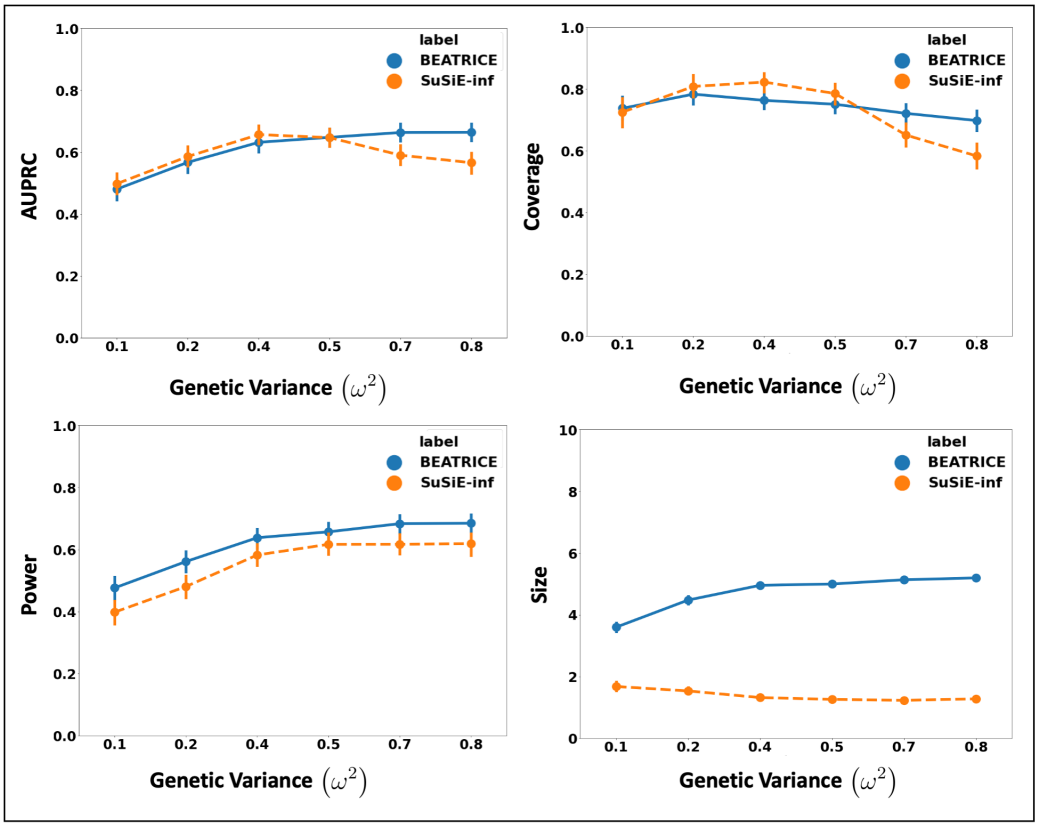


Figure 5-9. Performance metrics of BEATRICE and SuSiE-inf for increasing phenotype variance explained by genetics. The performance of SuSiE-inf is calculated over the subset of simulation settings in which the algorithm converges; non-convergent settings are omitted from the analysis. The x-axis corresponds to the variance explained by genetics (ω^2), and the y-axis plots the mean and confidence interval (95%) of each metric. We calculate the mean by fixing ω^2 to a value $\omega = \omega^*$ and sweeping over all the noise settings where $\omega = \omega^*$.

achieve comparable performance. However, as p decreases, meaning that the effects of non-causal variants increase, BEATRICE outperforms both baselines.

We further probe this behavior by illustrating the element-wise PIPs and the credible sets identified by all three methods under two simulation settings: $\{d = 1, \omega^2 = 0.2, p = 0.9\}$ (Fig. 5-11) and $\{d = 1, \omega^2 = 0.2, p = 0.1\}$ (Fig. 5-12). As seen in Fig. 5-11, the variance explained by the non-causal variants is small, so the causal variant is easy to distinguish and has been correctly identified by all three approaches. In contrast, we see in Fig. 5-12 that when the non-causal variants play a larger role, the causal variant no longer has the maximum GWAS z-score. Here, only BEATRICE

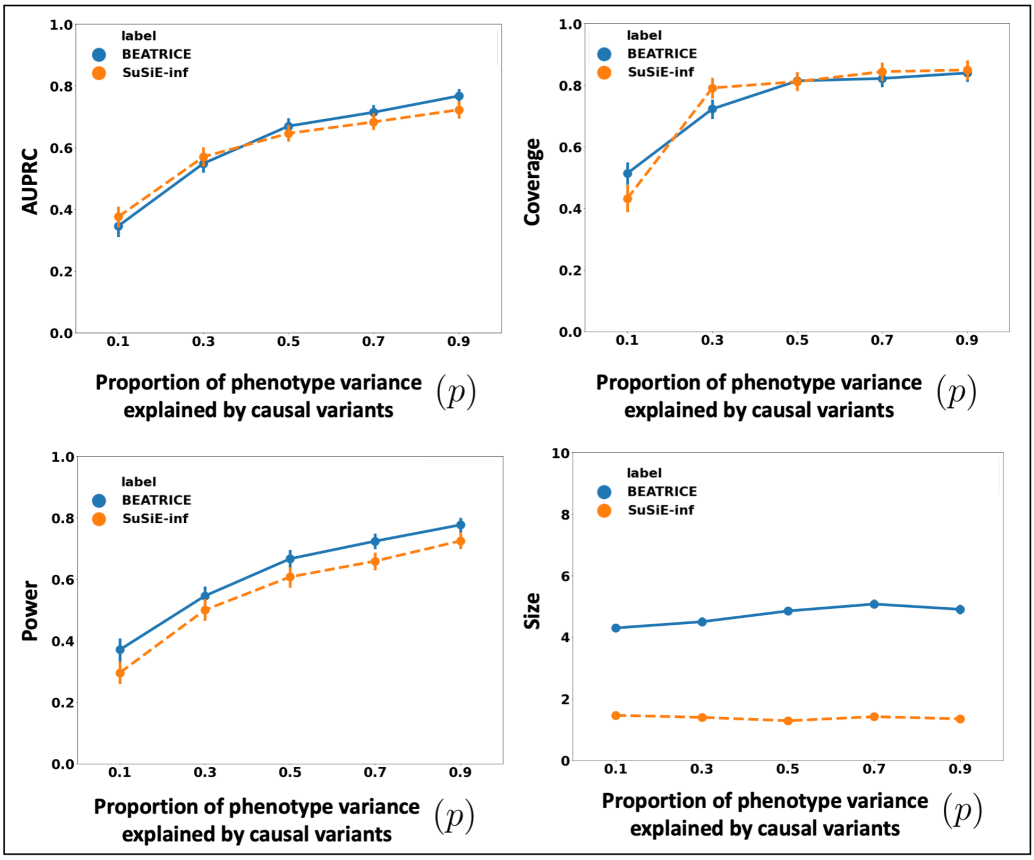


Figure 5-10. Performance metrics of BEATRICE and SuSiE-inf for multiple levels of noise introduced by non-causal variants. The performance of SuSiE-inf is calculated over the subset of simulation settings in which the algorithm converges; non-convergent settings are omitted from the analysis. The x-axis corresponds to the noise level (p), and the y-axis plots the mean and confidence interval (95%) of each metric. We calculate the mean by fixing p to a value $p = p^*$ and sweeping over all the noise settings where $p = p^*$.

correctly identifies the causal variant and assigns it the highest PIP. Both FINEMAP and SuSiE give uncertain predictions, as captured by the large credible sets and multiple high PIPs. We conjecture that BEATRICE takes advantage of the binary concrete distribution to model non-causal variants with non-zero effects, while using the sparsity term of $\mathcal{L}(\cdot)$ to prioritize potentially causal variants.

A second contribution of BEATRICE is our strategic integration of neural networks within a larger statistical framework. Specifically, we use the neural network in Fig. 5-1 as an inference engine to estimate the parameters $\boldsymbol{p}$ of our proposal distribution. In this case, the standard over-parameterization in the neural network

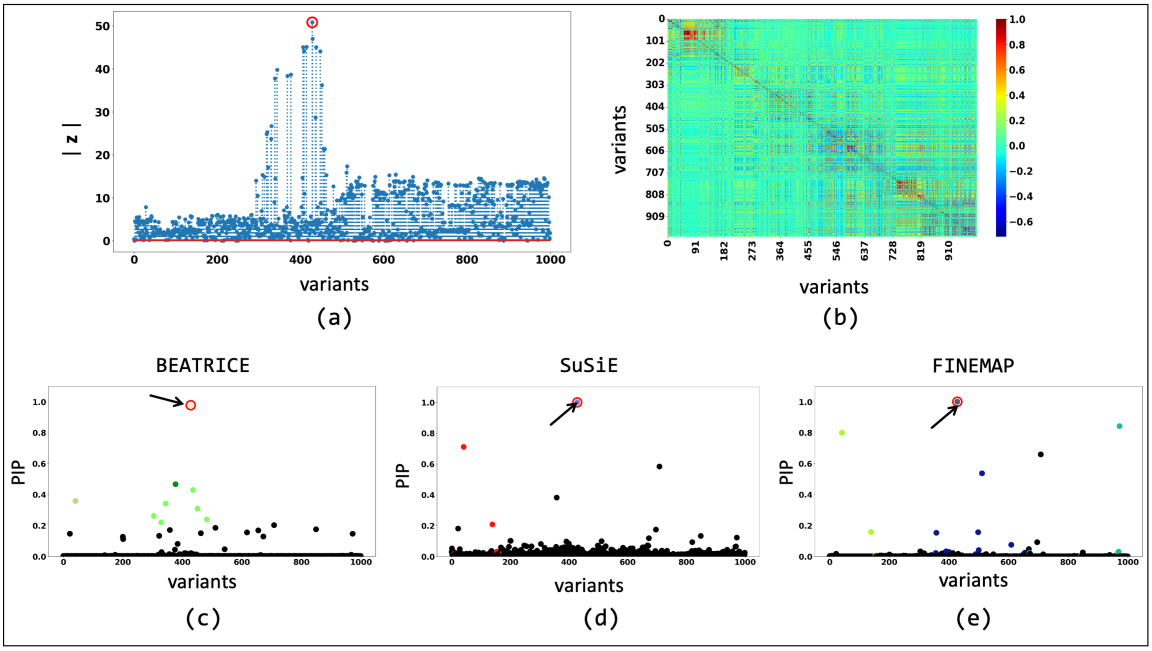


Figure 5-11. The fine-mapping performance of BEATRICE , SuSiE, and FINEMAP at a noise setting of $\{d = 1, \omega^2 = 0.2, p = 0.9\}$. (a) The absolute z-score of each variant as obtained from GWAS. (b) Pairwise correlation between the variants. (c), (d), and (e) are the posterior inclusion probabilities of each variant as identified by BEATRICE , SuSiE, and FINEMAP, respectively. The red circle marked by an arrow shows the location of the causal variant. We have further color-coded the variants based on their assignment to credible sets. The non-black markers represent the variants assigned to a credible set. Additionally, the variants in a credible set are marked by the same color.

helps BEATRICE to manage the complexity of the data while providing a buffer against overfitting. BEATRICE leverage the continuous representation of the causal vectors $\boldsymbol{c}^l$ to backpropagate the gradients through the random sampler and train the network. Additionally, the continuous representation of $\boldsymbol{c}^l$ results in low-variance gradients with respect to the underlying probability map, thus leading to a stable optimization.

Related to the above point, a third contribution of BEATRICE is its ability to efficiently build and evaluate a representative set of causal configurations during the optimization process. This set identifies key regions of the exponential search space to compute the PIPs and credible sets. In particular, we keep track of the sampled vectors at every iteration of the optimization, as described in Section 5.1.4.1.

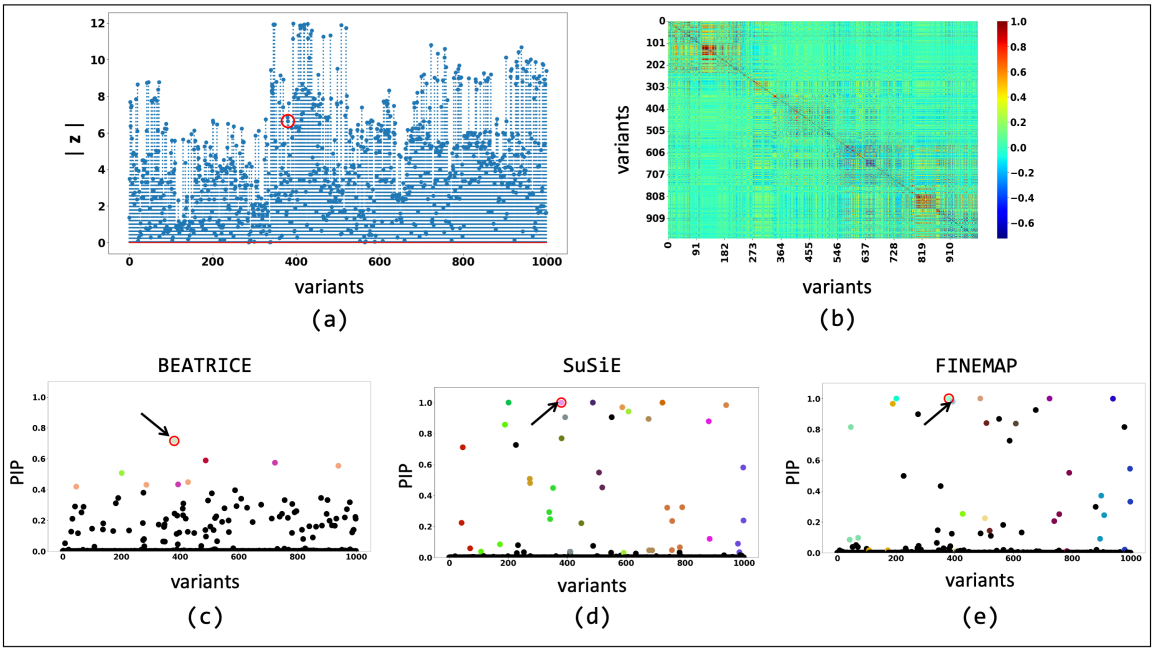


Figure 5-12. The fine-mapping performance of BEATRICE , SuSiE, and FINEMAP at a noise setting of $\{d = 1, \omega^2 = 0.2, p = 0.1\}$. (a) The absolute z-score of each variant as obtained from GWAS. (b) Pairwise correlation between the variants. (c), (d), and (e) are the posterior inclusion probabilities of each variant as identified by BEATRICE , SuSiE, and FINEMAP, respectively. The red circle marked by an arrow shows the location of the causal variant. We have further color-coded the variants based on their assignment to credible sets. The non-black markers represent the variants assigned to a credible set. Additionally, the variants in a credible set are marked by the same color.

By minimizing the KL divergence between the proposal distribution and the true posterior distribution, we ensure that the randomly sampled causal vectors slowly converge to the causal configurations that have non-negligible posterior probability. Our strategy lies in stark contrast with traditional mean-field approaches, where independence assumptions between elements of the proposal distribution do not allow for joint inference of the causal configurations. Furthermore, this strategy allows us to efficiently estimate the PIPs in finite run-time. Fig. 5-13 compares the average run-time of each method across all parameter settings. We observe that the run-time of BEATRICE and SuSiE are less than one minute. In contrast, FINEMAP requires significantly more time to converge.

The final contribution of BEATRICE is its simple and flexible design. Importantly,

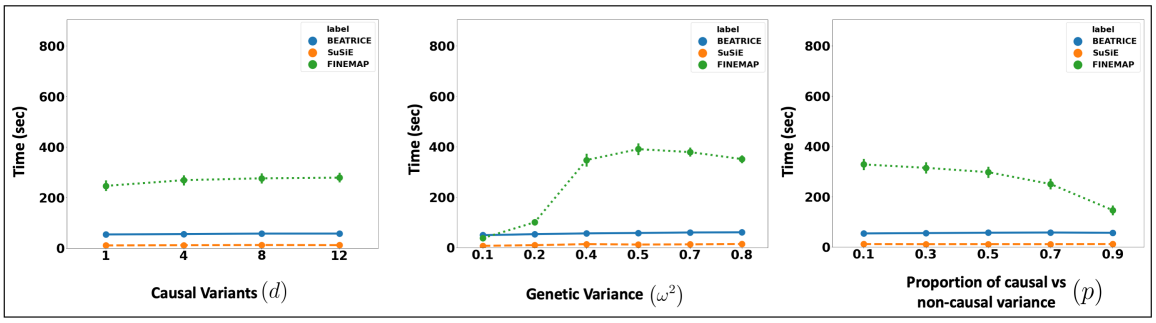
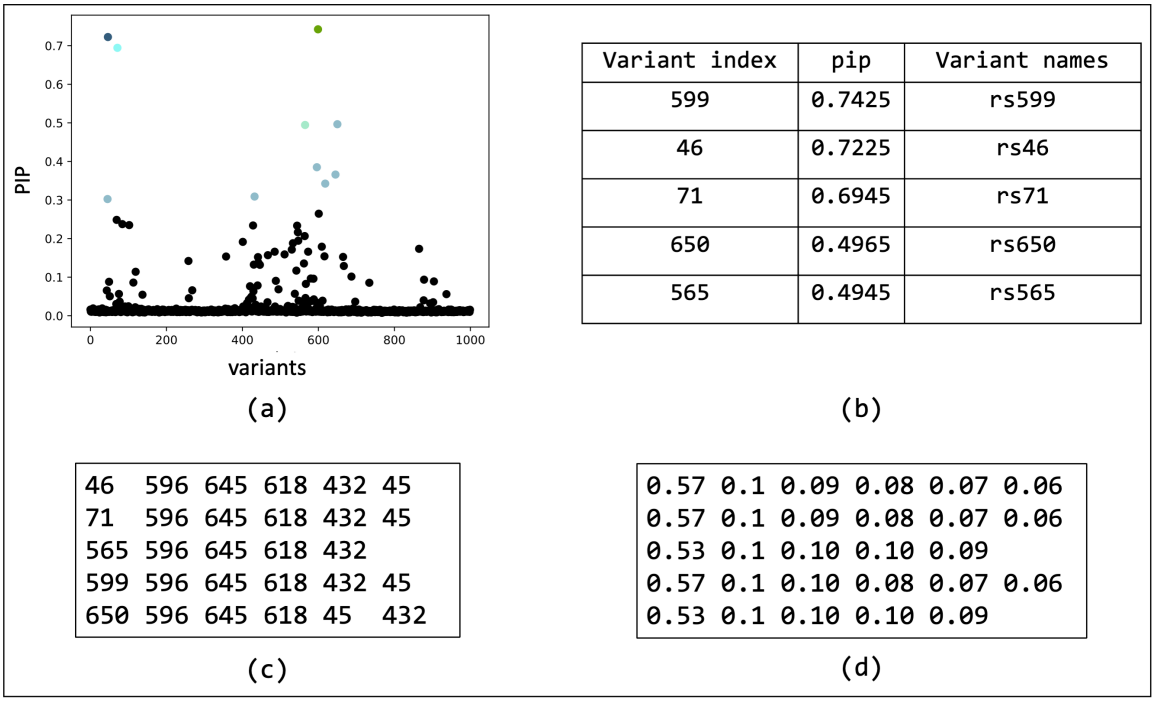


Figure 5-13. The runtime comparison of BEATRICE , SuSiE, and FINEMAP across all the simulation settings.

BEATRICE can easily incorporate priors based on the functional annotations of the variants. Formally, in the current setup, the prior over $\boldsymbol{c}$ is effectively constant, as captured by $p_0 = \frac{1}{m}$. We can integrate prior knowledge simply by modifying the distribution of p_0 across the variants. Thus, BEATRICE is a general-purpose tool for fine-mapping. Going one step further, a recent direction in fine-mapping is to aggregate data across multiple studies to identify causal variants [193]. Here, different LD matrices across studies helps to refine the fine-mapping results. BEATRICE can be applied in this context as well simply by modifying Eq. (5.12) as follows:

$$\mathcal{L}(\phi) = -\frac{1}{S\,L}\sum_{s=1}^{S}\sum_{l=1}^{L}\log\left(N\left(\boldsymbol{z}_s; 0, \boldsymbol{\Sigma}_{X_s} + \boldsymbol{\Sigma}_{X_s}\left(n\sigma^2\boldsymbol{\Sigma}_C^l(\phi)\right)\boldsymbol{\Sigma}_{X_s}\right)\right) + \sum_i \boldsymbol{p}_i \log\left(\frac{\boldsymbol{p}_i}{p_0}\right) + (1-\boldsymbol{p}_i)\log\left(\frac{1-\boldsymbol{p}_i}{1-p_0}\right) \tag{5.19}$$

where s denotes each separate study, S is the total number of studies in the analysis, and $\boldsymbol{z}_s, \boldsymbol{\Sigma}_{X_s}$ are the summary statistics for each study. In this work, we have shown that BEATRICE is highly efficient in handling the complexity that arises due to mutations with infinitesimal effects [188, 203]. Thus, we believe that the advantages of BEATRICE will be more evident when considering polygenic traits and diseases. Additionally, the high coverage and small size of credible sets reported in Fig. 5-4, 5-5, 5-6 show that BEATRICE can successfully prioritize variants in the presence of LD. This property is in stark contrast with the baseline finemapping approaches that generate a large number of credible sets that do not contain a causal variant. Taken



Variant index	pip	Variant names
599	0.7425	rs599
46	0.7225	rs46
71	0.6945	rs71
650	0.4965	rs650
565	0.4945	rs565

```
46  596 645 618 432 45
71  596 645 618 432 45
565 596 645 618 432
599 596 645 618 432 45
650 596 645 618 45  432
```

```
0.57 0.1 0.09 0.08 0.07 0.06
0.57 0.1 0.09 0.08 0.07 0.06
0.53 0.1 0.10 0.10 0.09
0.57 0.1 0.10 0.08 0.07 0.06
0.53 0.1 0.10 0.10 0.09
```

Figure 5-14. Overview of the outputs generated by BEATRICE . (a) The PIPs are displayed and color coded by their assignment to credible sets. (b) A table with the PIPs and the corresponding name of the variants. (c) A text file with the credible sets. Here each row represent a credible set and the entries are indices of the variants present in the credible set. The first column of each row represents the key index. (d) The conditional inclusion probability of each of the credible variants given the key variant.

together, we believe BEATRICE could be useful in eQTL studies, where multiple variants within a locus can show strong association due to the complex LD structure present in the human genome [54]. Additionally, there may be multiple causal variants within a locus, which adds to the complexity of the finemapping problem [18].

Code Availability We have compiled the code for BEATRICE and its dependencies into a docker image, which can be found at https://github.com/sayangsep/Beatrice-Finemapping. We have also provided installation instructions and a detailed description of the usage. The compact packaging will allow any user to directly download and run BEATRICE on their data. Namely, all the user must specify are a directory path to the summary statistics (i.e., z-scores), the LD matrix, and the number of subjects. Fig. 5-14 shows the outputs generated by BEATRICE . The results are

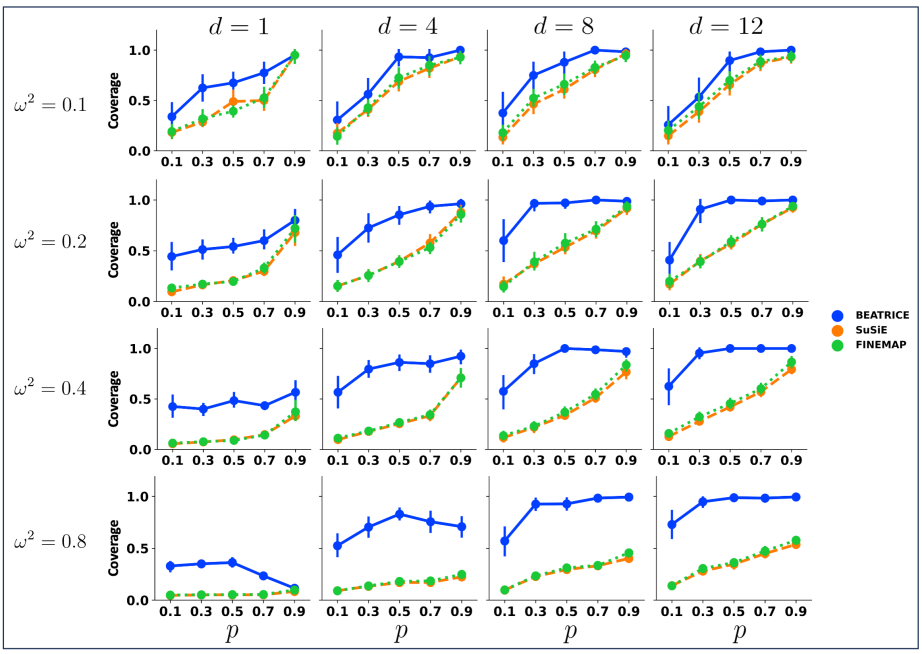


Figure 5-15. Coverage of the credible sets generated by the three models across multiple causal variants $d = [1, 4, 8, 12]$, multiple SNP heritability $\omega^2 = [0.1, 0.2, 0.4, 0.8]$ and multiple infinitesimal effects from non-causal variants $p = [0.1, 0.3, 0.5, 0.7, 0.9]$. Each row and column corresponds to a specific value of ω^2 and d, respectively. In each plot, the y-axis captures coverage, and the x-axis represents p.

output in (1) a PDF document that displays the PIPs and corresponding credible sets, (2) a table with PIPs, (3) a text file with credible sets, and (4) a text file with the conditional inclusion probability of the variants within the credible sets. The user can also generate the neural network losses describe in Eq. (5.12) by adding a flag to the run command.

5.1.8 Summary

We present BEATRICE , a novel Bayesian framework for fine-mapping that identifies potentially causal variants within GWAS risk loci through the shared LD structure. Using a variational approach, we approximate the posterior probability of the causal location(s) via a binary concrete distribution. We leverage the unique properties of binary concrete random variables to build an optimization algorithm that can successfully model variants with differing levels of association. Moreover, we introduce

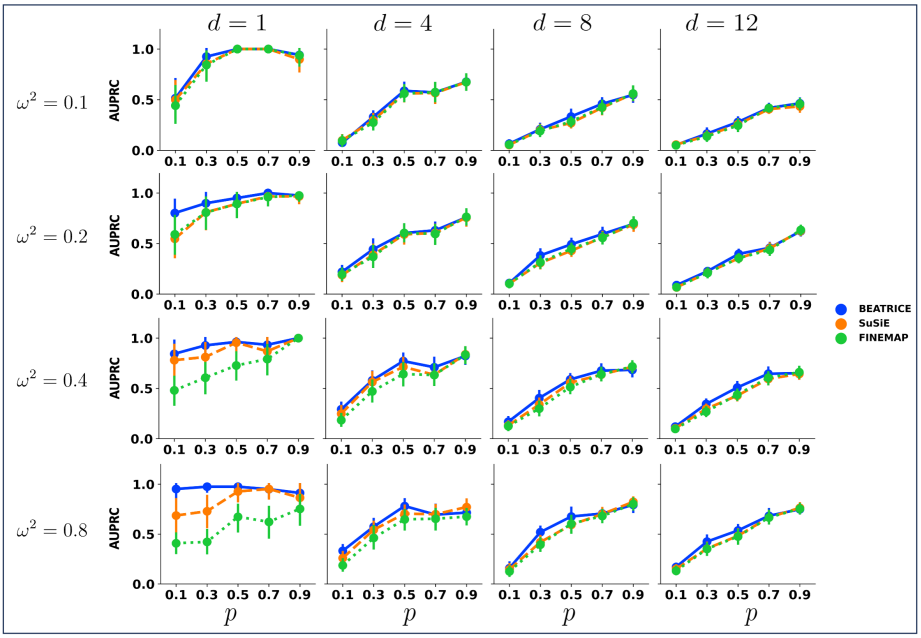


Figure 5-16. AUPRC of PIPs generated by the three models across multiple causal variants $d = [1, 4, 8, 12]$, multiple SNP heritability $\omega^2 = [0.1, 0.2, 0.4, 0.8]$ and multiple infinitesimal effects from non-causal variants $p = [0.1, 0.3, 0.5, 0.7, 0.9]$. Each row and column corresponds to a specific value of ω^2 and d, respectively. In each plot, the y-axis captures AUPRC, and the x-axis represents p.

a new strategy to build a reduced set of causal configurations within the exponential search space that can be neatly folded into our optimization routine. This reduced set is used to approximate the PIPs and identify credible sets. In a detailed simulation study, we compared BEATRICE with two state-of-the-art baselines and demonstrated the advantages of BEATRICE under different noise settings. Finally, our model does not have any prior on the causal variants and is agnostic to the original GWAS study. Hence, BEATRICE is a powerful tool to refine the results of a GWAS or QTL analysis. It is also flexible enough to accommodate a variety of experimental settings.

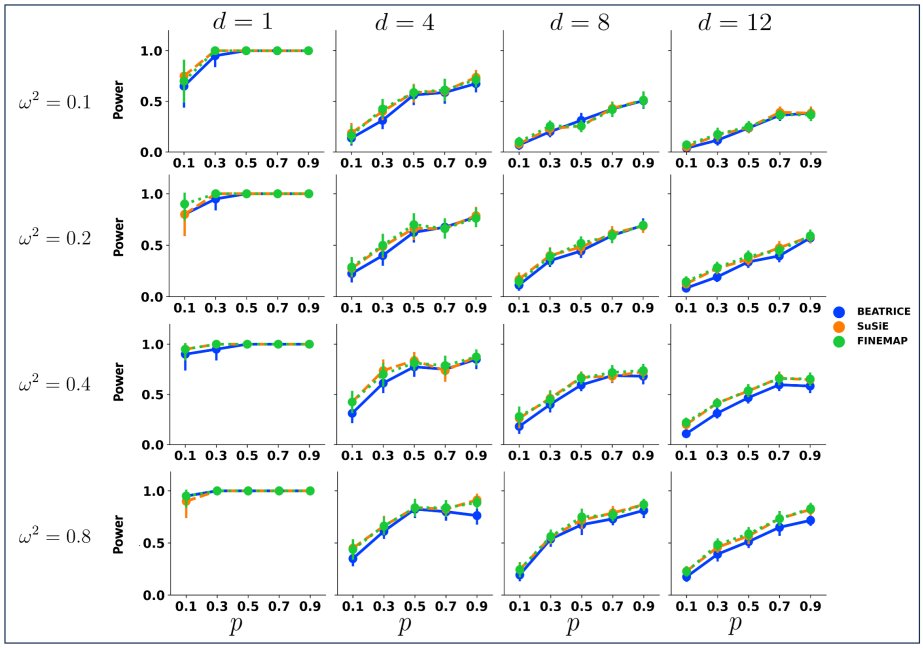


Figure 5-17. Power of the credible sets generated by three models across multiple causal variants $d = [1, 4, 8, 12]$, multiple SNP heritability $\omega^2 = [0.1, 0.2, 0.4, 0.8]$ and multiple infinitesimal effects from non-causal variants $p = [0.1, 0.3, 0.5, 0.7, 0.9]$. Each row and column corresponds to a specific value of ω^2 and d, respectively. In each plot, the y-axis captures power, and the x-axis represents p.

Additional Results

Detailed Comparison Analyses

In this section, we provide detailed comparisons of the models across individual noise setting without averaging the result. Fig. 5-16, Fig. 5-17, and Fig. 5-15 show the performance comparison of AUPRC, power and coverage, respectively. Fig. 5-15 shows a significant improvement in coverage compared to the baselines across noise settings. In addition, BEATRICE shows uniformly better AUPRC in Fig. 5-16. However, in terms of power, all models exhibit similar performance. A high coverage with comparable power suggests that BEATRICE can identify high-quality credible sets that contain causal variants. In contrast, the baselines identify many credible that do not contain a causal variant, ultimately leading to low coverage.

Chapter 6

An Interpretable and Biologically Regularized Approach to Encode High-dimensional Genetic Data in a Deep Learning Framework

The genetic data is high dimensional and complex. After imputation and preprocessing, traditionally, genetic data contain $\sim 300,000$ LD independent [204] genetic variants. Naive implementations of traditional regression-based models or Artificial Neural Networks (ANN) often lead to overfitting. For example, a simple one-layer ANN with ~ 100 nodes in the hidden layer will require us to estimate $\sim$ 3-million parameters. To prevent this, traditional imaging genetics models use a drastically reduced set of genetic features often based on a Genome-Wide Association Study (GWAS) [16, 17, 37] to prevent overfitting and ensuring model stability [205]. In terms of scale, $\sim 300,000$ genetic variants are reduced to ~ 1000 SNPs. In contrast, neuropsychiatric disorders are polygenetic, meaning that they are influenced by numerous genetic variants interacting across many biological pathways. The GWAS sub-selection step effectively removes the downstream information about these interactions [187, 206].

Within the genetics realm, there is a vast literature that associates genetic variants and genes to different biological pathways [56, 207]. Using this information, prior works of [57, 58] have created sparse neural networks to model genetic variants. The

sparse ANN aggregates genetic risk according to the pathways to predict a phenotypic variable. While an important first step, their ANN contains just a single hidden layer, which does not account for the hierarchical and interconnected nature of the biological processes.

The main focus of this chapter is to provide a strategically regularized framework to encode high-dimensional genetic data while accounting for the inherent complex interactions between biological processes. In our approach, we use graph convolutional networks (GCNs) [131] to leverage the interconnected genetic relationships. We construct a sparse hierarchical graph using gene ontology [56], which provides a structurally regularized framework to encode the whole genome genotype data. In addition, we use graph attention to track the information flow through the graph [136]. The attention mechanism focuses on the discriminative interactions between the nodes, thus finding implicated biological processes.

The first part of this chapter is based on our published work at ICLR [64]. Here, we use the graph-based encoding of genetic data in the imaging genetics framework. The encoded genetic representation is combined with imaging data to predict the genetic risk of schizophrenia. Additionally, we use the graph attention module to identify the implicated biological processes to provide insights into the underlying biology of the disorder. The second part of this chapter is about our ongoing work, where we explore the uses of graph-based encoding strategy to generate genetic risk scores that can provide insights about the underlying biological processes. This approach is in stark contrast with Polygenic Risk Score (PRS) based approaches, which generate a cumulative score but fail to provide insights about the implicated pathways. We investigate the utility of this approach to explore the genetic risk of autism.

6.1 GUIDE: A Biologically Interpretable Imaging Genetics Model to Link Genetic Risk Pathways and Neuroimaging Markers of Disease

This work is based on [64], where we introduce an interpretable **G**enetics and m**U**ltimodal Imaging based **DE**ep neural network (**GUIDE**), for whole-brain and whole-genome analysis. Our genetics network uses hierarchical graph convolution and pooling operations to embed subject-level data onto a low-dimensional latent space. The network is based upon the *a priori* knowledge of gene ontology [56]. The hierarchical network implicitly tracks the convergence of genetic risk across well-established biological pathways, while an attention mechanism automatically identifies the salient edges of this network at the subject level. On the imaging front, we use an encoder coupled with Bayesian feature selection to learn multivariate importance scores. The latent embeddings learned by our genetics graph convolutions and imaging encoder are combined for disease prediction. We demonstrate that GUIDE's ontology network and imaging-genetics fusion achieve better classification accuracy than state-of-the-art baselines. More importantly, GUIDE can identify robust and clinically relevant targets in both data domains.

Fig. 6-1 illustrates our GUIDE framework. Inputs are the gene scores $\boldsymbol{g}_n \in \boldsymbol{R}^{G\times 1}$ for each subject n and the corresponding imaging features $\boldsymbol{i}_n^1 \in \boldsymbol{R}^{M_1\times 1}$ and $\boldsymbol{i}_n^2 \in \boldsymbol{R}^{M_2\times 1}$ obtained from two different acquisitions. The gene scores $\boldsymbol{g}_n$ are obtained by grouping the original SNPs according to the nearest gene and aggregating the associated genetic risk weighted by GWAS effect size [17]. The subject diagnosis (phenotype) $y_n \in \{0, 1\}$ is known during training, but is absent during testing.

6.1.1 Embedding Genetic Information as Node Signals

The top portion of Figure 6-1 illustrates our attention-based graph convolution model for genetic embedding. The underlying graph is based on the gene ontological

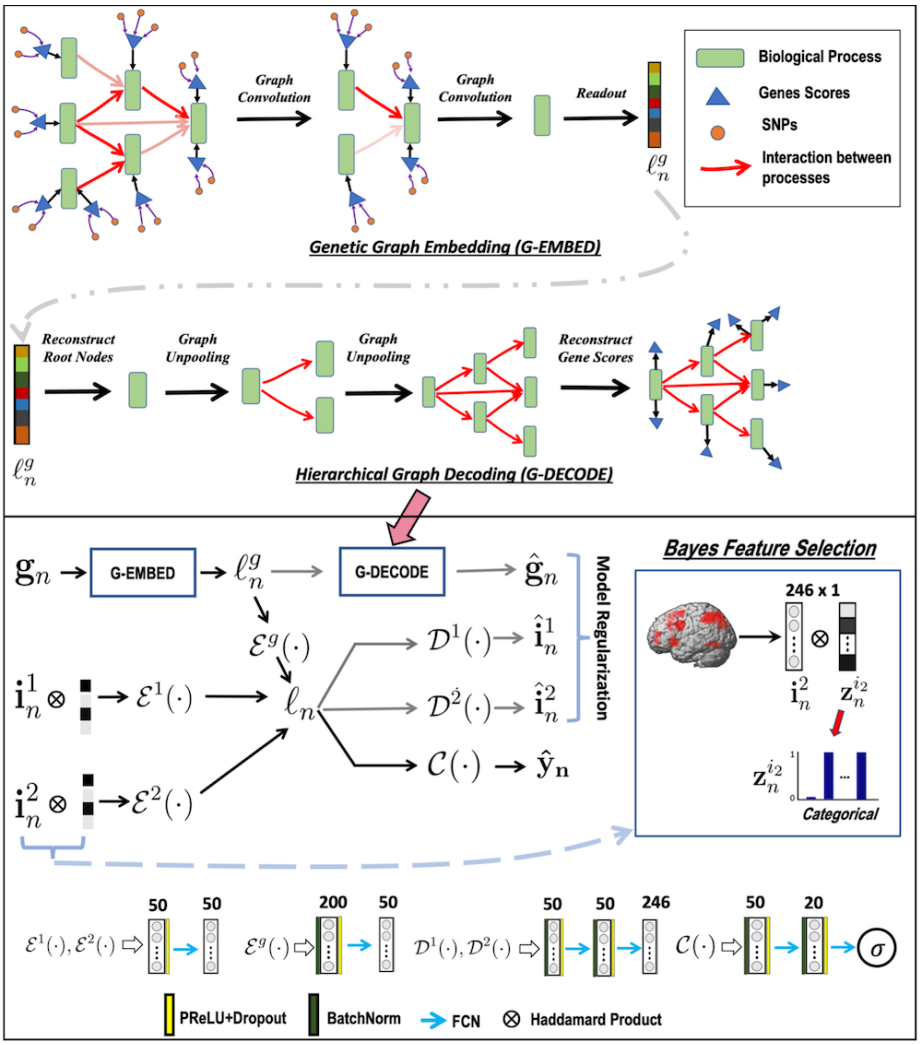


Figure 6-1. Overview of the GUIDE framework. **Top:** Gene embedding using attention based hierarchical graph convolution. We also depict the unpooling operation used as a regularizer. **Bottom:** Imaging and genetics integration; both modalities are coupled for disease classification. The variables $\{\boldsymbol{i}_n^1, \boldsymbol{i}_n^2\}$ correspond to the imaging data, and $\boldsymbol{g}_n$ is the genetic data. $\mathcal{E}(\cdot)$, $\mathcal{D}(\cdot)$, $\mathcal{C}(\cdot)$ are the feature extraction, model regularization, and classification operations, respectively.

hierarchy [56] that effectively maps each gene to different biological processes.

Gene Ontology: The ontology provides a pre-defined hierarchical network of biological processes, which has been curated by experts in biology. From a modeling perspective, the biological processes can be thought of as nodes in a graph. The two main parts of the ontology in [56, 208] are the assignment of genes to these biological

processes (i.e., the node embedding) and the directed edges between the biological processes (i.e., the hierarchical graph).

Node Signal Embedding: After mapping the genes to biological processes, we create the node signals of the gene ontology graph. Mathematically, this ontology gives rise to a sparse binary mapping matrix $\boldsymbol{A}_g \in \{0,1\}^{G \times P}$, where G is the total number of genes, and P is the number of biological processes (i.e., nodes). Given this substrate, GUIDE first learns a projection of the gene scores $\boldsymbol{g}_n$ of the subject n onto P graph nodes. Each node signal is a d-dimensional feature vector, $\boldsymbol{h}_n(p) \in \boldsymbol{R}^{1 \times d}$.

$$\boldsymbol{h}_n(p) = PReLU(\boldsymbol{g}_n^T (\boldsymbol{W}_p \otimes \boldsymbol{A}_g [:, p])), \tag{6.1}$$

where each of the columns of the learned weight matrix $\boldsymbol{W}_p \in \boldsymbol{R}^{G \times d}$ are masked by the nonzero entries in the p^{th} column of $\boldsymbol{A}_g$. We note that this projection step is similar to a single-layer perceptron with only a subset of input nodes connected to each hidden node.

6.1.2 Graph Attention and Hierarchical Pooling

In addition to grouping genes into biological processes, a standard ontology also specifies a hierarchical relationship between the biological processes themselves [209]. An example of this hierarchy is: generation of neurons $\rightarrow$ neurogenesis $\rightarrow$ nervous system development.

We use graph convolution to mimic the flow of information through the hierarchy of biological processes. Here, GUIDE learns two complementary pieces of information. The first is a set of graph convolutional filters that act on each embedded node signal, and the second is a set of subject-specific attention weights that select the discriminative edges. Formally, let the binary matrix $\boldsymbol{A}_p \in \boldsymbol{R}^{P \times P}$ capture the directed

edges in the ontology. Our graph convolution at stage l is:

$$\boldsymbol{h}_n^{l+1}(p) = \sigma\left(\sum_{j \in Child(p)} \boldsymbol{E}_n^l(p,j)\boldsymbol{h}_n^l(j)\boldsymbol{W}^l + \beta_t \boldsymbol{h}_n^l(p)\boldsymbol{W}_s\right), \tag{6.2}$$

where $\boldsymbol{h}_n^l(p) \in \boldsymbol{R}^{1\times d_l}$ is signal for node p at stage l, $\boldsymbol{W}^l \in \boldsymbol{R}^{d_l \times d_{l+1}}$ is the convolutional filter between stages l and $l+1$, β_t is the self-influence for node t, $\boldsymbol{W}_s \in \boldsymbol{R}^{d_l \times d_{l+1}}$ is the convolution filter for self loop, and $\sigma(\cdot)$ is the nonlinearity. The summation in Eq. (6.2) aggregates the influence over all child nodes defined by the graph $\boldsymbol{A}_p$, thus respecting the high-level ontology.

The variables $\boldsymbol{E}_n^l(p,j)$ quantify the influence of the child node j over the parent node p at convolutional stage l. Unlike a standard graph convolutional network, in which the edge weights are fixed, we learn $\boldsymbol{E}_n^l(p,j)$ using a graph attention mechanism. Mathematically, we have

$$\boldsymbol{E}_n^l(p,j) = \frac{\exp\left(\tanh\left(\left[\boldsymbol{h}_n^l(p)\boldsymbol{W}^l \quad \boldsymbol{h}_n^l(j)\boldsymbol{W}^l\right]\cdot \boldsymbol{c}^l\right)\right)}{\sum_{j\in Child(p)}\exp\left(\tanh\left(\left[\boldsymbol{h}_n^l(p)\boldsymbol{W}^l \quad \boldsymbol{h}_n^l(j)\boldsymbol{W}^l\right]\cdot \boldsymbol{c}^l\right)\right)} \tag{6.3}$$

where $\boldsymbol{c}^l$ is a fixed weight vector learned during training. We estimate the self influence variable as $\beta_t = \sigma\left(\boldsymbol{h}_n^l(p)\boldsymbol{W}_s \cdot \boldsymbol{c}_s^l\right)$, where $\sigma(\cdot)$ is sigmoid, and $\boldsymbol{c}_s^l$ is another weight vector learned during training.

Finally, we use hierarchical pooling to coarsen the graph. As formulated in Eq. (6.2), the graph convolution passes information "upwards" from child nodes to parent nodes. From the ontology standpoint, each stage of the hierarchy goes from a lower-level biological process (e.g., neurogenesis) to a higher-level biological process (e.g., nervous system development). Thus, we remove the lowest (leaf) layer from the graph at each stage l and continue this process until we reach the root nodes. The genetic embedding $\ell_n^g \in \boldsymbol{R}^{\mathcal{D}\times 1}$ is obtained by concatenating the signals from each root node.

6.1.3 Bayesian Feature Selection

The bottom branch of Fig. 6-1 shows our embedding and feature selection procedure for the imaging modalities. The feature selection strategy is similar to the strategy

defined in Section 4.2.1. Our dataset in this work contains two fMRI paradigms, leading to inputs $\boldsymbol{i}_n^1$ and $\boldsymbol{i}_n^2$ of each subject n. However, our framework naturally extends to an arbitrary number of modalities.

From a Bayesian viewpoint, the problem of feature selection can be handled by introducing an unobserved binary random vector $\boldsymbol{c}^m$ of the same dimensionality as $\boldsymbol{i}_n^m$ and inferring its posterior probability distribution given the paired training dataset: $\mathcal{D} = \{\boldsymbol{i}_n^m, y_n\}$, where m is the modality and y_n denotes the class label. By defining $\boldsymbol{I}^m = [\boldsymbol{i}_1^m, \ldots, \boldsymbol{i}_n^m]$, $\boldsymbol{y} = [y_1, \ldots, y_n]$, and $\boldsymbol{C}^m = [\boldsymbol{c}_1^m, \ldots, \boldsymbol{c}_n^m]$, we note that the desired posterior distribution $p(\boldsymbol{c}^m|\boldsymbol{I}^m, \boldsymbol{y})$ is intractable. One strategy is to minimize the KL divergence between an approximate distribution $q(\cdot)$ and the true posterior distribution $KL\left(q(\boldsymbol{c}^m)||p(\boldsymbol{c}^m|\boldsymbol{I}^m, \boldsymbol{y})\right)$. Mathematically, this optimization can be written as

$$\underset{q(\cdot)}{\operatorname{argmin}} \quad -E_q\left[\log\left(p\left(\boldsymbol{Y}|\boldsymbol{I}^m, \boldsymbol{C}^m\right)\right)\right] + KL\left(q(\boldsymbol{c}^m)||p(\boldsymbol{c}^m)\right), \tag{6.4}$$

where $p(\boldsymbol{c}^m)$ is a prior over the latent masks. While Eq. (6.4) does not have a closed-form solution, it can be optimized via Monte Carlo integration by sampling the vectors $\boldsymbol{c}_n^m \sim Bernoulli(\boldsymbol{p}^m)$, where $\boldsymbol{p}^m$ parameterizes the approximate distribution $q(\cdot)$, and minimizes the empirical form of Eq. (6.4) [180, 181]. In this case, the first term becomes the binary cross entropy loss where the input features $\boldsymbol{i}_n^m$ are masked according to $\boldsymbol{c}_n^m$. In order to learn the probability maps $\boldsymbol{p}^m$ during training, we replace the binary $\boldsymbol{c}_n^m$ with a continuous relaxation of the Bernoulli distribution:

$$\boldsymbol{c}_n^m = \sigma\left(\frac{\log(\boldsymbol{p}^m) - \log(1-\boldsymbol{p}^m) + \log(\boldsymbol{u}_n^m) - \log(1-\boldsymbol{u}_n^m)}{t}\right), \tag{6.5}$$

where $\boldsymbol{u}_n^m$ is sampled from $Uniform(0, 1)$, the parameter t controls the relaxation from the $\{0, 1\}$ Bernoulli, and the feature selection probabilities $\boldsymbol{p}^m$ are learned during training (see Section 6.1.4).

6.1.4 Multimodal Fusion and Model Regularization

As shown in Fig. 6-1, the Bayesian feature selection step is followed by a cascade of fully connected layers, denoted $\mathcal{E}^m(\cdot)$, to project each imaging modality m onto a low-dimensional latent embedding. Likewise, the genetic embedding ℓ_n^g is passed through a separate fully connected cascade $\mathcal{E}^g(\cdot)$ and onto the same low-dimensional space. To leverage synergies between the imaging and genetics data, we fuse the latent embeddings across modalities to obtain a common representation:

$$\boldsymbol{\ell}_n = \frac{1}{M_n}\left(\mathcal{E}^1(\boldsymbol{i}_n^1 \otimes \boldsymbol{c}_n^1) + \mathcal{E}^2(\boldsymbol{i}_n^2 \otimes \boldsymbol{c}_n^2) + \mathcal{E}^g(\ell_n^g)\right), \tag{6.6}$$

where $\otimes$ is the Hadamard product used in the Bayesian feature selection step, $\ell_n^g =$ G-EMBED($\boldsymbol{g}_n$), where G-EMBED$(\cdot)$ represents the genetics network based on the ontological hierarchy, and M_n is the number of modalities present for subject n. Finally, the latent embedding ℓ_n is input to a classification network to tie the learned biomarkers to patient/control phenotype.

Notice that our fusion strategy encourages the latent embedding ℓ_n for an individual patient to have a consistent scale, even when constructed using a subset of the modalities. Thus, we can accommodate missing data during training by updating individual branches of the network based on which modalities are present. In this way, GUIDE can maximally use and learn from incomplete data.

We introduce three regularizers to stabilize the model. The genetic regularizer reconstructs the gene scores by performing a hierarchical graph decoding (G-DECODE) on ℓ_n^g. This operation unwraps the gene encoding via the same ontology [56]. Likewise, the imaging regularizer decodes the original feature vectors from ℓ_n via the artificial neural networks $\mathcal{D}^m(\cdot)$. Finally, the prior over $\boldsymbol{c}_n^m$ appears as the KL divergence between the learned distribution $Ber(\boldsymbol{p}^m)$ and a binary random vector $Ber(\boldsymbol{p}_0)$ with

small entries to enforce sparsity. Our loss function during training is:

$$\mathcal{L}(\boldsymbol{i}_1, \boldsymbol{i}_2, \boldsymbol{g}) = -\sum_{n=1}^{N} \left(y_n \log(\hat{y}_n) + (1 - y_n) \log(1 - \hat{y}_n)\right)$$
$$+ \sum_{m=1}^{2} KL_{\mathcal{M}} \left(Ber(\boldsymbol{p}^m) || Ber(\boldsymbol{p}_0)\right) + \lambda_I \sum_{n=1}^{N_1} ||\boldsymbol{i}_n^1 - \mathcal{D}^1(\ell_n)||_2^2 + \lambda_I \sum_{n=1}^{N_2} ||\boldsymbol{i}_n^2 - \mathcal{D}^2(\ell_n)||_2^2$$
$$+ \lambda_G \sum_{n=1}^{N_g} ||\boldsymbol{g}_n - \text{G-DECODE}(\ell_n^g)||_2^2 \tag{6.7}$$

where $\hat{y}$ is the class prediction, N is the total number of subjects, N_m is the number of subjects with modality m present, and the hyper-parameters $\{\lambda_I, \lambda_G\}$ control the contributions of the data reconstruction errors. The function $KL_{\mathcal{M}}(\cdot||\cdot)$ in Eq. (6.7) averages the element-wise KL divergences across the input feature dimension, thus maintaining the scale of the prior term regardless of dimensionality.

The first two terms of Eq. (6.7) correspond to the classification task and the feature sparsity penalty, which are empirical translations of Eq. (6.4). The final three terms are the reconstruction losses, which act as regularizers to ensure that the latent embedding captures the original data distribution.

Training Strategy: As described in Section 6.1.7, our dataset consists of 1848 subjects with only genetics data and an additional 208 subjects with both imaging and genetics data. Given the high genetics dimensionality, we pretrain the G-EMBED and G-DECODE branches and classifier of GUIDE using the 1848 genetics-only subjects. These 1848 subjects are divided into a training and validation set, the latter of which is used for early stopping. We use the pretrained model to warm start GUIDE framework and perform 10-fold nested CV over the 208 imaging-genetics cohort.

We learn the Bayesian feature selection probabilities during training by sampling the random vectors $\boldsymbol{c}_n^1, \boldsymbol{c}_n^2$ during each forward pass using Eq. (6.5) and using them to mask the inputs $\boldsymbol{i}_n^1, \boldsymbol{i}_n^2$ for patient n. Finally, if a data modality is missing for subject n, we simply fix the corresponding encoder-decoder branch and update the remaining branches and predictor network using backpropagation.

Implementation Details: We use the first five layers of the gene ontology [56] to construct G-ENCODE. In total, this network encompasses 13595 biological processes organized from 2836 leaf nodes to 3276 root nodes. We perform a grid search over three order of magnitude and fix the hyperparameters $\lambda_I = 3 \times 10^{-3}, \lambda_G = 10^{-5}$. We fix the signal dimensionality at $d = 2$ for the gene embedding and $d_l = 5$ for the subsequent graph convolution and deconvolution operations based on the genetics-only subjects. Likewise, the non-linearity in Eq. (6.2) selected to be a $LayerNorm$, followed by $PReLU$ and $Dropout$, once again using the genetics-only subjects. The Bernoulli prior over $\boldsymbol{p}^m$ is set at $p_0 = 0.001$, which is consistent with [62]. We train GUIDE using ADAM with an initial learning rate of 0.0002 and decay of 0.7 every 50 epochs. Our code is implemented using Matlab 2019b and PyTorch 3.7. Training the full model takes roughly 17hrs on a 4.9GB Nvidia K80 GPU.

6.1.5 Baseline Comparison Methods

We compare GUIDE with three conventional imaging-genetics methods and single modality versions of our framework. In each case, the hyperparameters are optimized using a grid search.

Parallel ICA + RF: We concatenate the imaging modalities to single vector $\boldsymbol{i}_n = [\boldsymbol{i}_n^{1^T} \quad \boldsymbol{i}_n^{2^T}]^T$ and perform parallel ICA (p-ICA) [115] with the gene scores $\boldsymbol{g}_n$. Since p-ICA cannot handle missing modalities, we fit a multivariate regression model to impute a missing imaging modality from the available ones. Specifically, if $\boldsymbol{i}_1^n$ is absent, we impute it as: $\boldsymbol{i}_1^n = \boldsymbol{\beta}\, \boldsymbol{i}_2^n$, where $\boldsymbol{\beta}$ is the regression coefficient matrix obtained from training data. After imputation, we use p-ICA to decompose the imaging and genetics data into independent but interrelated networks:

$$\boldsymbol{i}_n = \boldsymbol{S}\, \boldsymbol{e}_n \quad \text{and} \quad \boldsymbol{g_n} = \boldsymbol{W}\, \boldsymbol{f}_n$$

where $\boldsymbol{S}, \boldsymbol{W}$ are independent source matrices and the $\boldsymbol{e}_n, \boldsymbol{f}_n$ are loading vectors. We concatenate the loading matrices $\left[\boldsymbol{e}_n^T, \boldsymbol{f}_n^T\right]$ and use it as the input feature vector for a random forest classifier.

During training, we apply p-ICA to just the training data to estimate the sources $\{\boldsymbol{S}_{train}, \boldsymbol{W}_{train}\}$. We use these estimated sources to obtain the loading matrices for the test data as follows:

$$\boldsymbol{i}_{test} = \boldsymbol{S}_{train}\boldsymbol{e} \quad \text{and} \quad \boldsymbol{g}_{test} = \boldsymbol{W}_{train}\boldsymbol{f}$$

The loading scores obtained from p-ICA are fed to a random forest model for classification. Our hyperparameter tuning optimizes the number and depth of the decision trees. We control the tree depth by setting the minimum number of observations per leaf node. After the grid search, these parameters are set to $\{\text{No. trees} = 10000, MinleafSize = 50\}$.

G-MIND: The G-MIND architecture by Ghosal *et al.* [62] is designed for a 1242 genetics input. Thus, we introduce a fully-connected layer to project the high-dimensional gene scores $\boldsymbol{g}_n$ onto a 1242 dimensional vector for input to G-MIND [62]. We evaluate both random weight initialization and pretraining the genetics branch of G-MIND with the 1848 genetics-only subjects.

Ghosal *et al.* [62] selected the hyperparameters as powers of 10 such that the rescaled terms in the loss function lie within the same order of magnitude [1-10]. This criterion is intuitive (i.e., equal importance is given to both the imaging and genetic data), and it is not performance driven. We use a similar strategy and fix the hyperparameters of genetic reconstruction loss, imaging reconstruction loss, classification loss, and sparsity loss to $0.0001, 0.1, 1, 0.001$, respectively.

G-MIND (Sub-selection): The G-MIND architecture relies on a subselection of SNPs based on the p-values reported in a prior GWAS analysis of schizophrenia [17].

Here, we subselect the same set of SNPs and feed them to the G-MIND model. This baseline captures the performance change when the ontology-based representation is replaced with the GWAS sub-selection. We use the hyperparameters reported in [62].

Single Modality Prediction: We consider two versions of GUIDE. The first consists of the genetics branches and classifier, and the second consists of just the imaging branches and classifier. We optimize these networks architectures using repeated 10-fold cross validation outlined in Section 6.1.4. These baselines probe the advantages of integrating imaging and genetics data modalities in a single framework. The hyperparameters for single modality prediction are the same as the full model.

GUIDE (Random Dropout): The feature selection layer of our original GUIDE framework both regularizes the model and identifies potential imaging biomarkers. In this baseline, we replace the Bayesian feature selection layer with random dropout.

Hierarchical GCN: GUIDE utilizes an ontology to flow the information through the graph. The gene ontology network is curated based on *a priori* information. In this baseline, we replace the ontology based graph with a random graph. The construction of this graph is explained in Section 6.1.6.2. This baseline captures the performance gain for embedding biological knowledge into our architecture.

GCN + MaxPooling: We compare our model with a standard GCN introduced by Kipf & Welling [131]. In this baseline, we generate a weighted adjacency matrix ($\boldsymbol{A} \in \boldsymbol{R}^{13908\times13908}$) using absolute correlation values between the gene scores from the pretraining data (i.e., 1848 genetic only subjects). This adjacency matrix acts as an undirected graph between the nodes. Unlike GUIDE, the nodes represent each gene score instead of a biological process. We also perform hierarchical pooling with a max pooling operation [210] to reduce the data dimension between the graph convolution

layers. This is in stark contrast to GUIDE where we use biological knowledge to consolidate the information flow through the network. Here, we note that the adjacency matrix $\boldsymbol{A}$ is not sparse, resulting in a significantly higher computational overhead than GUIDE.

In this baseline, we use two graph convolutional layers, each followed by a max-pooling layer. The max-pooling layer reduces the number of nodes by approx 50% such that after the final layer, the reduced data is of the same dimension as ℓ_g in GUIDE. This architecture balances the computational requirements and the data representation ability of the model. Finally, we fix the readout layers, the imaging modules, and the classification module to the same architecture of GUIDE.

6.1.6 Evaluation Strategy

We conduct a comprehensive evaluation of our framework that includes influence of embedding *a priori* biological information, biomarker reproducibility, and classification performance.

6.1.6.1 Ablation Study

In this section, we evaluate the gain from three novel components of GUIDE: graph attention, feature selection, and the decoder for regularization. Specifically, the graph attention encourages GUIDE to focus on the discriminative interaction patterns in the genetic data, the Bayesian feature selection identifies the most predictive brain regions and the decoders ensure that the low-dimensional embedding faithfully captures the original data distribution.

6.1.6.2 Influence of the Ontology-Based Hierarchy

In GUIDE, we assume that processing the data according to an established gene ontology will extract more robust and discriminative biomarkers. To test this assumption,

we compare our gene ontology network to random networks and to an unstructured model. The random networks contain the same number of nodes in each layer as the ontology-based graph. However, we randomly permute edges between the parents and children nodes, and between the genetic inputs $\boldsymbol{g}_n$ and nodes. The corresponding unstructured model is a fully-connected ANN with the same number of parameters, but no inherent structure between layers. As a benchmark, we compare the deep learning models with the gold-standard Polygenetic Risk Score (PRS) for schizophrenia developed by the PGC consortium [17]. Broadly, the PRS is a weighted combination of the risk alleles for schizophrenia, as determined by a large GWAS study. We run a logistic regression on the (scalar) PRS to determine class membership.

6.1.6.3 Classification Performance

We use repeated 10-fold cross validation (CV) on the 208 imaging-genetics subjects to quantify the performance. The models are trained using 8 folds, and the remaining two folds are reserved for validation and testing. We report accuracy, sensitivity, specificity, Area Under the ROC Curve (AUROC), and Area Under the precision-recall Curve (AUPRC). The operating point is chosen by minimizing the classification error on the validation set. We further use DeLong tests to compare statistical difference in AUROC between GUIDE and the baselines.

6.1.6.4 Reproducibility of Feature Importance Maps

The probability maps $\boldsymbol{p}_m$ capture the importance of each feature of modality m. We evaluate both the predictive performance and the reproducibility of our Bayesian feature selection (BFS) scheme. We extract the the top-K features of $\boldsymbol{p}_m$ learned during each training fold of our repeated 10-fold CV setup and encode this information as a binary indicator vector, where the entry '1' indicates that the feature is among the top-K. We compare the BFS features with Kernel SHAP [130]. The background

values for K-SHAP are fixed to the average input data across the training set. We use the validation set to select the top K K-SHAP features and encode them in an binary indicator vector.

To quantify predictive performance, we mask our test data from each fold using the BFS and K-SHAP indicator vectors and send it through GUIDE for patient versus control classification. To quantify the reproducibility of the top-K features, we calculate the pairwise *cosine* similarity between all the binary vectors across the folds as identified by either BFS or K-SHAP. The distribution of similarities tells us how often the same imaging features are selected across subsets of the data.

6.1.6.5 Discovering of Biological Pathways

The attention layer of G-ENCODE provides a subject-specific measure of "information flow" through the network. We first trace all possible paths between leaf and root nodes to identify discriminative paths in the network. The importance of edges along these paths are used in a logistic regression framework to predict the subject class label. We then perform a likelihood ratio test and obtain a p-value (FDR corrected) for each path in terms of patient/control differentiation. For robustness, we repeat this experiment 10 times using subsets of 90% of our total dataset $1848 + 208$ subjects and select paths that achieve $p < 0.05$ in at least 7 of the 10 subsets.

6.1.7 Results

6.1.7.1 Data and Preprocessing

We evaluate GUIDE on a study of schizophrenia provided by Lieber Institute for Brain Development (LIBD) Institution that contains SNP data and two fMRI paradigms.

Illumina Bead Chipset including 510K/ 610K/660K/2.5M is used for genotyping. Quality control and imputation were performed using PLINK and IMPUTE2. The resulting 102K linkage disequilibrium independent ($r^2 < 0.1$) indexed SNPs are

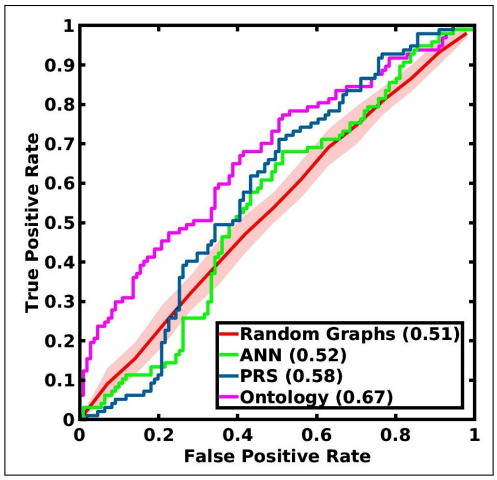


Figure 6-2. ROCs for the PRS (blue), unstructured ANN (green) and the structured models where G-EMBED and G-DECODE use either random graphs (red) or the gene ontology network (magenta). The AUROC is given in parentheses.

grouped to the nearest gene (within 50kb basepairs) [211]. The 13,908 dimensional input genes cores are computed as the weighted average of the SNPs using GWAS effect size [17]. We note that the GWAS was performed on a separate dataset that did not include our site.

Imaging data include two task-fMRI paradigms. The first paradigm is a working memory task (N-back), and the second is a simple declarative memory task (SDMT). The data modalities are explained in Section 2.4.1. After preprocessing we use Brainnetome atlas [143] to define extract brain activation maps from 246 cortical and subcortical regions. The inputs $\boldsymbol{i}_n^1, \boldsymbol{i}_n^2$ correspond to the average contrast over voxels in the respective region. Additional details are reported in Section 2.4.1.

Our imaging dataset is incomplete, as many subjects were only scanned using one of the two fMRI paradigms. Table 6-II reports the breakdown of patients and controls for each configuration. Finally, Table 6-I reports the demographic information for the cohort. In each case, the patient and control groups were matched on age, IQ (WRAT score), and years of education.

In total, our dataset contains 1848 subjects (792 schizophrenia, 1056 control) with

Demographic	LIBD	
	N-back	SDMT
Sex (M/F)	113/47	70/40
Age (years)	31 ± 10	31 ± 10
Education (years)	15 ± 2	12 ± 2
IQ	105 ± 10	104 ± 10

Table 6-I. Demographic information for subjects provided by LIBD Institution.

Data Sets							
SNP-Only		N-back + SNP		SDMT + SNP		All Modalities	
Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls
793	1056	42	56	17	31	38	24

Table 6-II. Breakdown by patients and controls for each configuration.

just SNP data and 208 subjects with imaging and SNP data, divided as follows: 98 subjects (42 schizophrenia, 56 control) with SNP and N-back data, 48 subjects (17 schizophrenia, 31 control) with SNP and SDMT data, and 62 subjects (38 schizophrenia, 24 control) who have all the three data modalities.

6.1.7.2 Benefit of the Gene Ontology Network:

To quantify the value of embedding a biological hierarchy into our model architecture, we first train just the genetics branch and classifier for all three deep networks (unstructured, random graphs, ontology) using the 1848 genetics-only subjects and test them on the SNP data from the other 208 subjects. We sample 10 random graphs in this experiment.

Figure 6-2 illustrates the ROC curve over the test data. As seen, the testing performance is near-chance using both random graphs and the ANN. The gold-standard PRS does slightly better; however, this measure was derived from a much larger consortium dataset. In contrast, our gene ontology network (magenta line in Figure 6-2) achieves the best performance of any method, suggesting that our biologically inspired architecture can extract robust and predictive features from the genetic data.

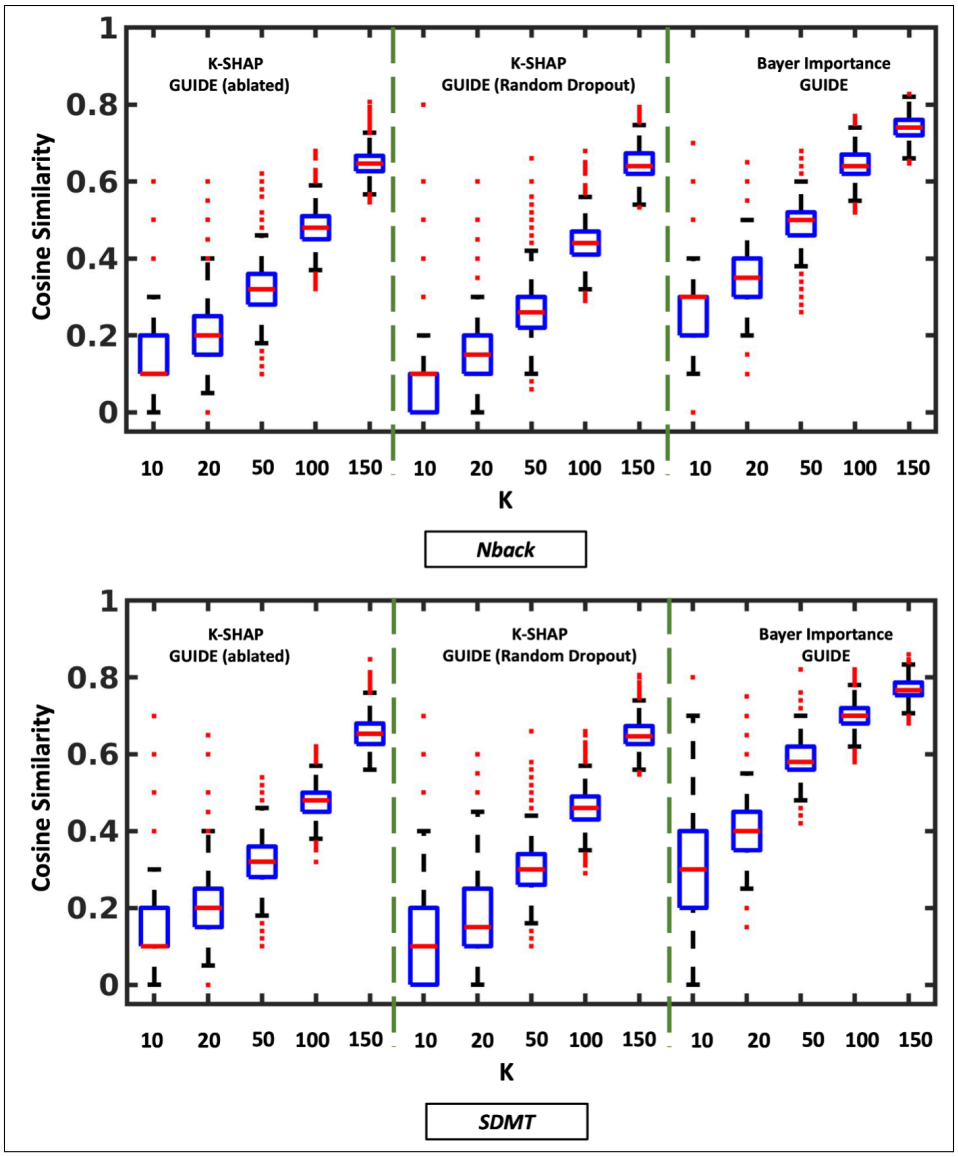


Figure 6-3. The reproducibility of imaging biomarkers when the input layer of GUIDE is trained without dropout, with random dropout, and with Bayesian feature selection.

6.1.7.3 Classification Performance:

Table 6-III reports the 10-fold CV testing performance of all the methods on the multimodal imaging-genetics dataset; we repeat the CV procedure 10 times to obtain standard deviations for each metric. We note that P-ICA and G-MIND with random initialization have relatively poor performance, likely due to the high dimensionality of $\boldsymbol{g}_n$ and low sample size. Pretraining on a separate genetics dataset improves the

Table 6-III. Classification performance (mean ± std) across repeated CV runs. P-values obtained from DeLong test indicate significantly greater AUROC for GUIDE than each of the baselines.

Method \ Perf	Sensitivity	Specificity	Accuracy	AUPRC	AUROC	P-Value
P-ICA	0.30 ± 0.10	$\mathbf{0.80 \pm 0.07}$	0.56 ± 0.03	0.54 ± 0.04	0.59 ± 0.04	$< 10^{-4}$
G-MIND (random)	0.62 ± 0.06	0.65 ± 0.05	0.63 ± 0.02	0.62 ± 0.03	0.67 ± 0.03	$< 10^{-4}$
G-MIND (pretrain)	0.60 ± 0.07	0.66 ± 0.07	0.63 ± 0.03	0.62 ± 0.03	0.68 ± 0.02	$< 10^{-4}$
G-MIND (Sub-selection)	$\mathbf{0.63 \pm 0.08}$	0.67 ± 0.06	0.65 ± 0.01	0.63 ± 0.04	0.70 ± 0.02	$< 10^{-4}$
Hierarchical GCN	0.48 ± 0.17	0.75 ± 0.13	0.62 ± 0.02	0.65 ± 0.02	0.71 ± 0.02	1.8×10^{-4}
GCN+MaxPooling	0.43 ± 0.19	0.76 ± 0.14	0.61 ± 0.03	0.64 ± 0.02	0.69 ± 0.02	$< \times 10^{-4}$
Imaging Only	0.44 ± 0.18	0.76 ± 0.14	0.61 ± 0.01	0.62 ± 0.02	0.66 ± 0.02	$< 10^{-4}$
Genetic Only	0.54 ± 0.15	0.69 ± 0.10	0.62 ± 0.02	0.63 ± 0.03	0.68 ± 0.02	$< 10^{-4}$
GUIDE (Random Dropout)	0.51 ± 0.14	0.79 ± 0.12	0.66 ± 0.02	$\mathbf{0.70 \pm 0.02}$	$\mathbf{0.75 \pm 0.01}$	0.27
GUIDE	0.62 ± 0.04	0.76 ± 0.04	$\mathbf{0.69 \pm 0.01}$	$\mathbf{0.70 \pm 0.03}$	$\mathbf{0.75 \pm 0.01}$	

AUC of G-MIND, highlighting the benefits of increased data. Among all the versions of G-MIND, the subselection of SNPs gives the best performance. We also compare GUIDE with two graph convolution-based models. In the first model, we see that replacing the structured representation of the nodes with a random hierarchical graph reduces performance. Hence, the ontological representation is extracting meaningful information from the data. The second model compares GUIDE with a standard GCN and max-pooling. This GCN uses a dense graph for convolution and relies on a data-driven strategy for dimensionality reduction. Again, we observe poor performance, in this case likely due to the large model size. In comparison, GUIDE operates on a sparse graph and relies on biological knowledge for dimensionality reduction. Finally, we replace the GUIDE BFS layer with random dropout. While the quantitative performance is similar, the feature reproducibility is much higher with BFS (see Fig. 6-3). This result underscores the dual value of BFS for predictive performance and biomarker discovery.

Table 6-IV. Classification performance (mean ± std) across repeated CV runs. P-values obtained from DeLong test indicate significantly greater AUROC for GUIDE than each of the ablated models.

Method \ Perf	Sensitivity	Specificity	Accuracy	AUPRC	AUROC	P-Value
No Attention	0.49 ± 0.09	$\mathbf{0.79 \pm 0.08}$	0.65 ± 0.03	0.70 ± 0.02	0.73 ± 0.02	0.01
No Feature Selection	0.56 ± 0.13	0.73 ± 0.10	0.65 ± 0.03	0.67 ± 0.03	0.74 ± 0.03	0.28
No Decoder	0.58 ± 0.18	0.66 ± 0.17	0.62 ± 0.02	0.65 ± 0.02	0.70 ± 0.01	$< 10^{-4}$
No Attention, No Feature Selection	0.56 ± 0.18	0.72 ± 0.13	0.64 ± 0.02	0.67 ± 0.03	0.73 ± 0.03	0.05
No Attention, No Decoder	0.50 ± 0.16	0.76 ± 0.12	0.64 ± 0.03	0.66 ± 0.02	0.71 ± 0.02	$< 10^{-4}$
No Feature Selection, No Decoder	0.61 ± 0.10	0.69 ± 0.10	0.65 ± 0.03	0.63 ± 0.03	0.72 ± 0.03	0.01
No Attention, No Feature Selection, No Decoder	0.57 ± 0.20	0.68 ± 0.17	0.63 ± 0.03	0.61 ± 0.03	0.71 ± 0.02	4×10^{-4}
GUIDE	$\mathbf{0.62 \pm 0.04}$	$\underline{0.76 \pm 0.04}$	$\mathbf{0.69 \pm 0.01}$	$\mathbf{0.70 \pm 0.03}$	$\mathbf{0.75 \pm 0.01}$	

6.1.7.4 Performance in Ablation Study

Table 6-IV compares the performance of the ablated models with GUIDE. This ablation study allows us to quantify both the improvement of each component when they are incorporated in the model and the degradation in performance when one component is ablated from the full model. For example, removing the decoders causes the classification performance to degrade, regardless of the other components (i.e., GUIDE is better than "No Decoder" and "No Attention, No Feature Selection" is better than "No Attention, No Feature Selection, No Decoder"). We observe similar trends for both graph attention and Bayesian feature selection components. Thus, our ablation study demonstrates that all three components of GUIDE are essential for phenotypic prediction.

6.1.7.5 Reproducibility of BFS Features:

Fig. 6-4 illustrates the classification AUC when the input features are masked according to the top-K importance scores learned by the BFS (solid blue) and K-SHAP (dashed red) procedures. The confidence intervals are obtained across the repeated CV folds. As seen, both feature selection schemes achieve similar AUCs as the number of features

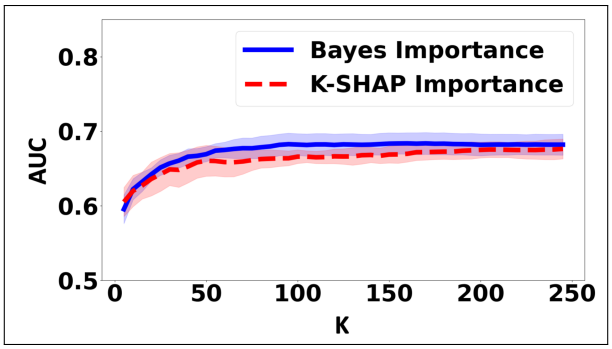


Figure 6-4. Mean AUC and confidence interval when masking the top-K imaging features learned by BFS (solid blue) and K-SHAP (dashed red). K is varied along the x-axis.

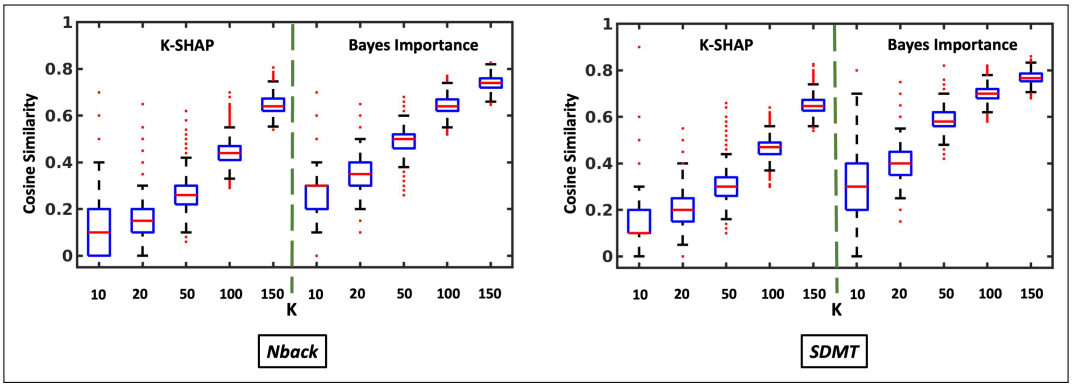


Figure 6-5. The reproducibility of imaging biomarkers when the feature selection has been done using K-SHAP vs Bayesian dropout. **Left** shows the performance on Nback data, **Right** shows the performance on SDMT data

K is varied across its entire range, thus highlighting the robustness of our (simpler) BFS approach.

Fig. 6-5 reports the distribution of *cosine* similarities between the masked feature vectors learned by BFS and K-SHAP for $K = 10, 20, 50, 100, 150$. The repeated CV procedure is run 10 times, yielding 100 total folds and 4950 pairwise comparisons per method. Notice that our BFS procedure achieves significantly higher *cosine* similarity values at each setting for K, which suggests that it selects a more robust set of features that is consistent across subsets of our main cohort.

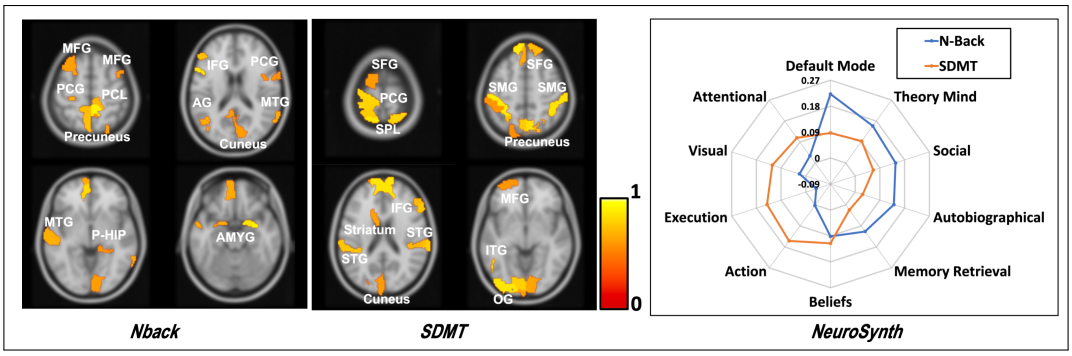


Figure 6-6. Left: The consistent set of brain regions captured by the dropout probabilities $\{\boldsymbol{b}^1, \boldsymbol{b}^2\}$ for $K = 50$. The color bar denotes the selection frequency. **Right:** Brain states associated with the selected regions for each fMRI task, as decoded by Neurosynth.

6.1.7.6 Imaging Biomarkers

Fig. 6-6 illustrates the consistent imaging features that are selected by our method across the folds for $K = 50$. We have colored each brain region according to the selection frequency. For clarity, we have displayed only the top 40% regions. We observe that the N-back biomarkers involve the Middle Frontal Gyrus (MFG), Inferior Frontal Gyrus (IFG), and default mode network which are associated with schizophrenia [4, 212]. The SDMT biomarkers implicate the Supramarginal Gyrus (SMG), Superior Frontal Gyrus (SFG), along with Precuneus and Cuneus. We further interpret the higher order brain states implicated by these regions using Neurosynth [213]. Neurosynth uses a reverse-inference procedure to select a set of "cognitive terms" associated with a set of input coordinates based on how frequently similar patterns have been observed across the fMRI literature. Fig. 6-6 shows that the N-back and SDMT biomarkers are associated with memory retrieval and attention [214, 215], thus verifying that GUIDE captures information relevant to the fMRI tasks.

6.1.7.7 Genetic Pathways

GUIDE contains 14881 "biological paths" between the leaf and root nodes. Using the likelihood ratio test outlined in Section 6.1.6.5, we are able to identify 152 paths

with $p < 0.05$ after FDR correction that appear in at least 7 of the 10 random training datasets. We cluster the biological processes present along these paths into 10 categories based on their semantic similarity. This clustering strategy provides intuition about the higher-level biological functions identified by this model. Here, we use the tf-idf [216] information retrieval scheme to extract keywords in the pathways; we then embed them in a two-dimensional space using t-SNE [217] and apply a k-means clustering algorithm. Fig. 6-7 shows the clusters along with the most frequent keywords within each cluster. As seen, the frequent biological processes involve calcium signaling, regulation of macrophage and immunological synapse formation which have been previously linked to schizophrenia [184, 218, 219]. This exploratory experiment shows that GUIDE can be used to extract discriminative biological information about neuropsychiatric disorders.

6.1.8 Discussion

We introduce a biologically regularized approach to encode millions of genetic variants in a non-linear framework while maintaining interpretability. The first key contribution of GUIDE is a pre-defined network of biological processes to strategically combine

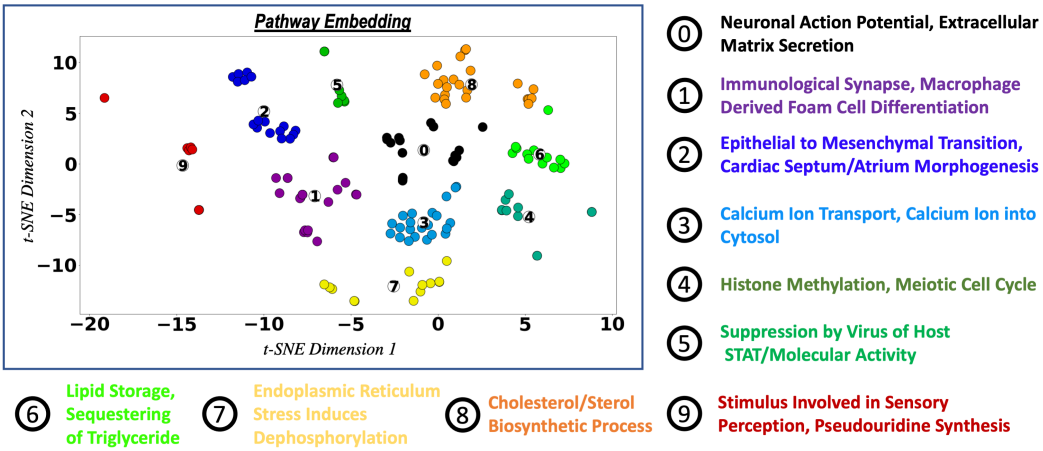


Figure 6-7. Ten different categories of pathways based on their semantic similarity. The key words show the most frequent biological processes within each cluster.

genetic information. The nodes and edges of the graph have a biological meaning. As a result, the edge properties and the node features provide us insights into the information flow through the network. This kind of transparent model has the capability to pinpoint biological mechanisms in a data-driven fashion.

The second contribution is using graph attention to quantify the effect of each node on its parent node. This strategy allows us to identify discriminative edges associated with schizophrenia. This results in the identification of implicated biological pathways that could potentially identify novel therapeutic targets.

The third contribution is the hierarchical convolution combined with hierarchical pooling. In standard polygenic risk score-based models, all the SNP information is combined as a weighted sum. As a result, they lack interpretability. In comparison, GUIDE combines the SNP information in a hierarchical fashion informed by SNP-gene, gene-pathway mapping. In addition, the graph attention controls the relative contributions of each component. This provides a fully automatic and interpretable strategy to combine genetic information and create an interpretable genetic risk score.

Finally, GUIDE provides an end-to-end approach to combine multimodal imaging and genetic data in a single framework. The biologically regularized model restricts the parameter space, leading to model stability and improved prediction. In summary, GUIDE provides a comprehensive framework that can combine multiple data modalities, handle missing data, and parse high-dimensional imaging and genetic data.

6.1.9 Summary

We propose a novel biologically interpretable graph convolutional network that integrates imaging and genetics data to perform disease prediction. This model is able to leverage prior biological information of different connected biological processes to identify patterns from both imaging and genetic data sets. Additionally, the unique use of Bayesian feature selection is able to find a set of clinically relevant features.

The improved classification performance shows the ability of this model to build a comprehensive view about the disorder based on the incomplete information obtained from different modalities. We note that, our framework can easily be applied to other imaging modalities, such as structural and diffusion MRI, simply by adding encoder-decoder branches. In the future, we will apply our framework to other application domains, such as autism and Parkinson's disease.

6.2 GUIDE-PRS:A Biologically Interpretable and Non-linear Approach to Generate Polygenic Risk Scores

In this section, we explore the use of our GUIDE model (shown in Section 6.1) to create non-linear and interpretable polygenic risk scores. Traditional polygenic risk scores combine all the SNP data using a weighted sum. While this approach is easy to interpret and provides the genetic liability of a disorder, it fails to provide insights into the underlying processes. The genetic mutations in the DNA affect genes differently [18, 220, 221]. As a result, the genetic risk could be divided non-uniformly across different biological processes. Prior works [28, 101, 222] have tried to address this issue by incorporating information on biological pathways or gene regulatory networks. However, these approaches rely heavily on prior expertise and handcrafted features. In comparison to these traditional approaches, our genetic encoding strategy (shown in Section 6.1) can automatically combine information using hierarchy of gene ontology and create pathway-specific gene scores. In addition, the graph attention network guides the distribution of genetic risk across different pathways, creating non-linear genetic risk scores.

This section will show our ongoing work on creating interpretable and non-linear genetic risk scores for autism spectrum disorder. The model used in this section is heavily motivated by our genetic encoding strategy shown in Section 6.1. We deviated

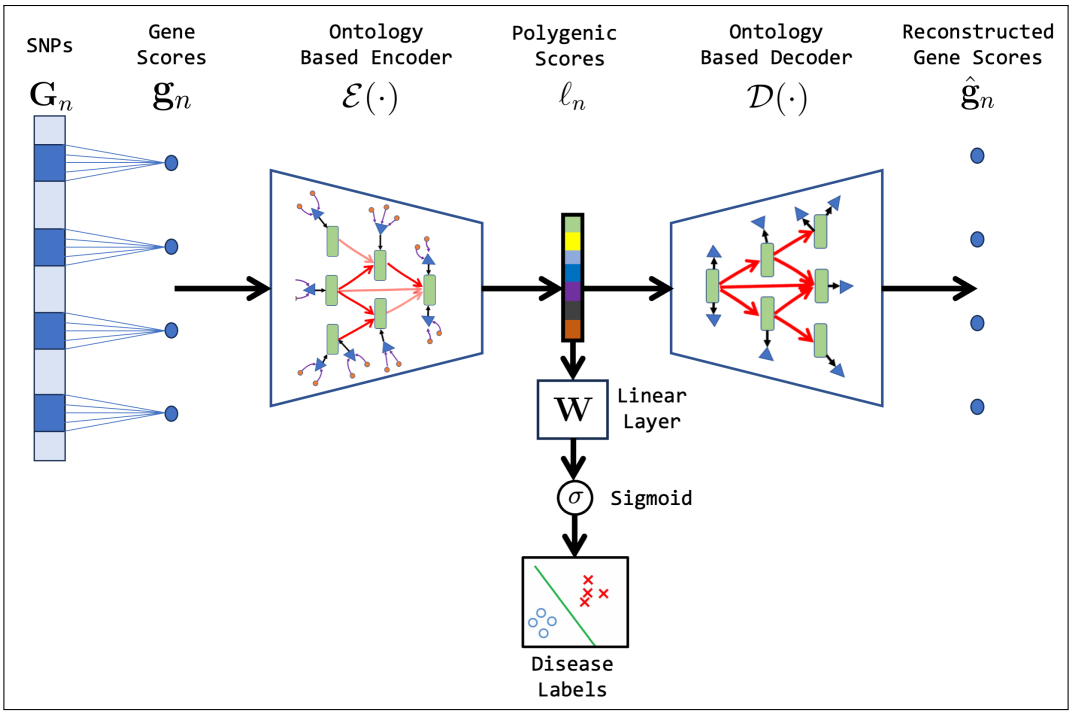


Figure 6-8. The hierarchical encoding strategy to create pathways-specific polygenic risk scores. The SNP data $\boldsymbol{G}_n \in \boldsymbol{R}^{M\times 1}$ from subject n is encoded to create gene scores $\boldsymbol{g}_n \in \boldsymbol{R}^{G\times 1}$. The hierarchical ontology based encoder $\mathcal{E}(\cdot)$ uses graph convolution and graph attention to encode the gene scores and create pathway specific polygenic scores. The polygenic scores $\boldsymbol{\ell}_n \in \boldsymbol{R}^{R\times 1}$ is generated for R root nodes. $\mathcal{D}(\cdot)$ is the hierarchical unpooling operation along the ontology. $\boldsymbol{W}$ is a linear operation to predict class labels from the plygenic scores.

from our previous approach primarily in three places: we provide an additional gene-to-node embedding strategy that restricts the parameter space, we replace the soft-max operation with sigmoid to capture node-node interaction, and finally, we replace the classifier ANN with a linear layer for interpretability.

6.2.1 Hierarchical Encoding of Genetic Data

Fig. 6-8 shows the overall strategy to create the genetic risk scores using hierarchical encoding of the genetic data along a Gene Ontology (GO) [223]. The encoder and the decoder follow a similar strategy described Section 6.1.

Generating Gene Scores: This first level of encoding creates the genetic risk scores from the SNP data. Mathematically, the gene scores are created as:

$$\boldsymbol{g}_n(i) = \sum_{j \in Ne(i)} \omega_j \boldsymbol{G}[n, j] \tag{6.8}$$

where $\boldsymbol{g}_n(i)$ is the i-th gene scores for subject n, $Ne(i)$ is the collection of SNPs in the neighborhood (within $\sim 500K$ basepairs) of the i-th gene, $\boldsymbol{G}$ is the SNP data matrix, and ω_j is the SNP-effect obtained from a separate GWAS study. This encoding strategy creates gene-based liability scores of the underlying disorder.

Gene To Node Mapping: This work explores two types of node signal embedding strategies. The first strategy follows the same approach given in Section 6.1.1. We learn a projection of the gene scores $\boldsymbol{g}_n$ of the subject n onto P graph nodes. Each node signal is a d-dimensional feature vector, $\boldsymbol{h}_n(p) \in \boldsymbol{R}^{1\times d}$.

$$\boldsymbol{h}_n(p) = PReLU(\boldsymbol{g}_n^T (\boldsymbol{W}_p \otimes \boldsymbol{A}_g [:, p])), \tag{6.9}$$

where each of the columns of the learned weight matrix $\boldsymbol{W}_p \in \boldsymbol{R}^{G\times d}$ are masked by the nonzero entries in the p^{th} column of $\boldsymbol{A}_g$. However, note that the embedding weight $\boldsymbol{W}_p$ is unique for each node, resulting in a large unconstrained parameter space.

To address this issue, we introduce a second embedding strategy where the embedding weights are shared across all the nodes. This strategy follows the following structure:

$$\boldsymbol{h}_n(p) = PReLU(\boldsymbol{g}_n^T \left(\tilde{\boldsymbol{W}} \otimes \boldsymbol{A}_g [:, p]\right) \otimes \boldsymbol{\alpha}_p), \tag{6.10}$$

where $\tilde{\boldsymbol{W}}$ is the embedding weight shared across all the nodes. However, to maintain the unique identities of each node, we perform elementwise multiplication of the embeddings with node-specific weight vectors $\boldsymbol{\alpha}_p \in \boldsymbol{R}^{1\times d}$.

Propagation Along Ontological Hierarchy: Gene ontology (GO) [56] provides a hierarchical representation to combine genetic information strategically. Traditionally,

the genes are mapped to multiple low-level biological processes based on their involvement in the specific process. The low-level processes are combined to create high-level biological processes. This hierarchical approach provides us with a framework to distribute genetic risk along different biological processes and create an interpretable genetic risk score.

We use the same graph convolution approach (described in Section 6.1.2) to combine the genetic risk along the biological processes. Formally, let the binary matrix $\boldsymbol{A}_p \in \boldsymbol{R}^{P\times P}$ capture the directed edges in the ontology. Our graph convolution at stage l is:

$$\boldsymbol{h}_n^{l+1}(p) = \sigma\left(\sum_{j\in Child(p)} \boldsymbol{E}_n^l(p,j)\boldsymbol{h}_n^l(j)\boldsymbol{W}^l + \beta_t \boldsymbol{h}_n^l(p)\boldsymbol{W}_s\right), \tag{6.11}$$

where $\boldsymbol{h}_n^l(p) \in \boldsymbol{R}^{1\times d_l}$ is signal for node p at stage l, $\boldsymbol{W}^l \in \boldsymbol{R}^{d_l\times d_{l+1}}$ is the convolutional filter between stages l and $l+1$, β_t is the self-influence for node t, $\boldsymbol{W}_s \in \boldsymbol{R}^{d_l\times d_{l+1}}$ is the convolution filter for self loop, and $\sigma(\cdot)$ is the nonlinearity. Here, note that each node signals $\boldsymbol{h}_n^l(p)$ represents a genetic score generated by aggregating the influence over all child nodes defined by the graph $\boldsymbol{A}_p$.

Previously, the contribution of a child node to its parent node $\boldsymbol{E}_n^l(p,j)$ was captured using a soft-max operation. However, one drawback of soft-max is that if a child node's contribution is small it automatically drives up the contribution of other child nodes. Soft-max fails to capture the absolute importance of a node over its parent nodes. So, in this work, we deviate from soft-max operations and use sigmoid to capture the interactions. Mathematically, we have

$$\boldsymbol{E}_n^l(p,j) = \sigma\left(\begin{bmatrix}\boldsymbol{h}_n^l(p)\boldsymbol{W}^l & \boldsymbol{h}_n^l(j)\boldsymbol{W}^l\end{bmatrix}\cdot \boldsymbol{c}^l\right) \tag{6.12}$$

where $\sigma(\cdot)$ is the sigmoid function. Finally, we use hierarchical pooling to coarsen the graph. As formulated in Eq. (6.11), the graph convolution passes information "upwards" from child nodes to parent nodes. From the ontology standpoint, each stage

of the hierarchy goes from a lower-level biological process (e.g., neurogenesis) to a higher-level biological process (e.g., nervous system development). The final encoded vector, $\boldsymbol{\ell}_n \in \boldsymbol{R}^{R\times 1}$, captures a non-linear representation of genetic risk generated by the root nodes.

In addition to the encoder, we have a decoder branch that reconstructs the gene scores from the encoded latent representation. The decoder branch uses graph unpooling along the same ontological hierarchy. This operation regularizes the model and ensures that the encoded representation contains informative information about the input data.

Subject Classification: The genetic risk scores $\boldsymbol{\ell}_n$ generated by the root nodes are passed through a linear layer, followed by a sigmoid operation to predict the class labels. The classification module ensures that the genetic risk scores are informative and contains discriminative information about the disorder.

Loss Funtion: We train the graph neural network by minimizing a loss function, which is a combination of reconstruction loss and classification loss. Mathematically, the loss can be written as follows:

$$\mathcal{L} = \mathcal{C}(\boldsymbol{y}, \hat{\boldsymbol{y}}; \boldsymbol{W}) + \lambda \sum_n ||\boldsymbol{g}_n - \mathcal{D}\left(\mathcal{E}\left(\boldsymbol{g}_n; \eta\right); \phi\right)||_2^2 \tag{6.13}$$

where $\mathcal{C}(\cdot,\cdot)$ is the binary cross entropy loss between the original class labels $\boldsymbol{y}$ and the predicted labels $\hat{\boldsymbol{y}}$, $\mathcal{E}(\cdot)$ is the encoding operation and $\mathcal{D}(\cdot)$ is the decoding operation parametrized by $\boldsymbol{\eta}$ and $\boldsymbol{\phi}$, respectively. The parameters, $\boldsymbol{W}$ $\boldsymbol{\phi}$ and $\boldsymbol{\eta}$ are learned during training using backpropagation.

Implementation Details: We train the model on SSC data that contains 4217 subjects with probands and pseudo-controls and evaluate the performance on ACE data containing 236 subjects with probands and controls. We use 20% of the SSC data

as validation data for early stopping. We perform a grid search to fix the configuration of the model. We sweep over multiple depths of the ontology network $L = [3, 5, 7]$, two different ontologies (*Cellular Components (CC), Biological Processes (BP)*), two different node embedding strategies, and node feature dimensions $d_l = [2, 5]$. The SSC data lacks "true" controls, so we finalize the model using bootstrapping on the ACE data. We perform 50 bootstrapping trials and use 10% of ACE data as validation to pick the model. The rest 90% of ACE data is used to report the performance of our model across the validation experiments. We train the models with a learning rate of 0.005 using ADAM [186] optimizer. We also want to emphasize that all the models are trained on SSC, and their parameters were fixed during evaluation, so there's no data leakage.

6.2.2 Evaluation Strategies

Classification Performance: We compare the classification performance of our model with a traditional polygenic risk score-based approach. We use the same genetic variants that are used in our model to create the polygenic risk scores. Following the strategy explained in [98] we create the polygenic risk scores as:

$$r_n = \sum_i \mathbf{1}(P(\omega_i) < ths)(\boldsymbol{G}[n, i]\omega_i) \tag{6.14}$$

where r_n is the PRS, $\mathbf{1}(\cdot)$ is the indicator function that decides whether to include a variant i based on a p-value threshold, $P(\omega_i)$ is the p-value of ω_i, ω_i is the estimated effect size obtained from a GWAS, and $\boldsymbol{G}$ is the genotype matrix. We use the polygenic risk score and the 10 principal components obtained from the genotype data as regressors in the logistic regression framework. The p-value threshold was chosen based on the validation performance.

Both the models are trained on the SSC data and evaluated on ACE data. The final classification performance is reported in the form of Area Under ROC (AUROC) and Area Under Precision Recall (AUPRC) curve.

Evaluation of Genetic Risk Scores: In this experiment, we evaluate the importance of the non-linear genetic risk scores captured by the node features of the root nodes. After encoding, the genetic risk scores across R root nodes are represented by the vector $\boldsymbol{\ell}_n \in \boldsymbol{R}^{R\times 1}$. Every element in $\boldsymbol{\ell}_n(i)$ represents a pathway-specific risk score that is generated by combining the genetic risk distributed across the hierarchy of the children nodes of that specific root node. We evaluate the risk scores by performing a two-sample t-test between the probands and the controls. Finally, we report the average $-\log(pvalue)$ across the 50 bootstrap trials for each of the risk scores. This analysis will identify the pathway-specific risk scores that contains informative information about the disorder.

Evaluation of Edge Interactions: The edge interactions in our model are captured by the graph attention matrix $\boldsymbol{E}$. Each matrix element $\boldsymbol{E}_n^l[i,j]$ captures the interactions between the child node j, and the parent node i at layer l. In our multi-layer GCN, each layer has a unique interaction between the child node j, and the parent node i. In this analysis, we select the last interactions between i and j before node j gets pruned by graph pooling. The hypothesis is that the last instance of node j before pruning contains all the information from its child nodes; hence, its interaction with the parent node i is most informative. After selecting the interactions, we create an interaction matrix $\tilde{\boldsymbol{E}}_n$ for individual subjects. Similar to the strategy described in Section 6.1.6.5 we identify all the paths between root and leaf nodes. We concatenate the edge interaction values as features and pass it through a Likelihood Ratio Test (LRT) to identify paths that shows discriminative interactions between probands and controls. This analysis gives a fine-grained understanding of the interaction between biological processes associated with the disorder. We identify the paths that have $pvalue < 0.05$ after FDR correction. We report the frequency of a node across 50 bootstrap trials that is present along paths with significant p-values.

6.2.3 Preliminary Data Analysis

6.2.3.1 Data and Preprocessing

In this analysis, we acquire two genetic datasets from the Simons Simplex Collection (SSC) and the Autism Center of Excellence (ACE). The preliminary data analysis and acquisition are described in Section 2.4.2.

SSC Data: After initial preprocessing, we obtain ~ 2591 simplex family data. Each simplex family contains a proband, an unaffected sibling, and both parents. This family-based data lacks control subjects. Following the strategy defined in [16, 224] we create pseudo-controls from this family data. Finally, we subselect the subjects that belong to the European population for our analysis. The final SSC data contain 4217 subjects (2109 probands and 2108 pseudo-controls.)

ACE Data The ACE data contains 346 subjects among which 236 subjects belong to the European population. We use these subjects to evaluate the models.

Genetic Preprocessing We preprocess the proband-pseudocontrol data from SSC and proband-control data from ACE following the pipeline defined in RICOPILI [147]. After preprocessing, we impute the data using SANGER which uses Haplotype Reference Consortium (release 1.1) as the reference data. We use the overlapping SNPs from ACE and SSC for our analyses. After imputation, we clumped ($r2 < 0.5$, $500Kb$) the data to identify $\sim 400K$ LD independent index SNP. The LD clumping removes redundancy from the highly correlated genetic variants. The final $\sim 400K$ index SNPs are assigned to $\sim$ 17000 genes to create the gene scores. We use PANTHER [142] to map the genes to the nodes of gene ontology. In our experiments, we explore two different ontologies. The first one involves *Cellular Components* (CC) which provides knowledge about the cellular structures in which a gene product performs a function.

Method \ Perf	AUROC	AUPRC
PRS	0.64 ± 0.01	0.64 ± 0.02
GUIDE-PRS (CC)	$\mathbf{0.68 \pm 0.01}$	0.65 ± 0.01
GUIDE-PRS (BP)	0.66 ± 0.005	$\mathbf{0.66 \pm 0.01}$

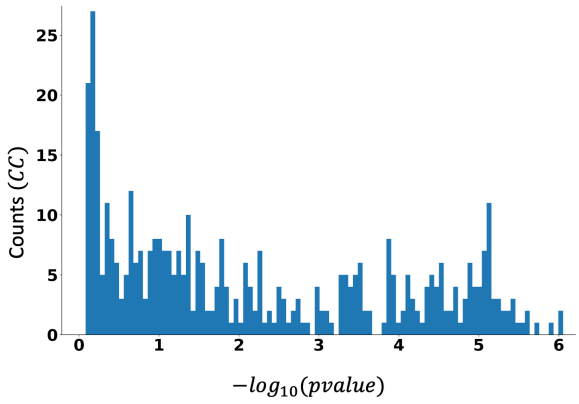
Table 6-V. The classification performance of the models across 50 bootstrap trials on ACE data. The AUROC and AUPRC capture the area under the ROC curve and the area under the precision-recall curve, respectively. GUIDE-PRS (BP) and GUIDE-PRS (CC) are two variants of our model where one is trained using the ontology of *Biological Processes* (BP) and the other is trained using the ontology of *Cellular Components* (CC).

The second one involves *Biological Processes* (BP), which provides knowledge about the genes' involvement in 'biological programs' accomplished by multiple molecular activities.

In order to create the polygenic risk scores, we use SPARK [225] and iPSYCH [16] data to obtain the GWAS summary data. We ensure that there is no subject overlap between the GWAS data and our data. During training, we also perform a GWAS on the SSC training data and combine all the three GWAS summary statistics using METAL [226]. The combined statistics are used to create the gene scores of our model and the polygenic risk scores for the baselines.

6.2.3.2 Results

Classification Performance: Table 6-V shows the classification performance of the models. We again note that all the models are trained on the SSC data and directly evaluated on the ACE data for classification performance. In addition, the model parameters are fixed based on the performance on 10% of ACE data, and the results are reported on the rest of the ACE data across 50 bootstrap trials. Here, we see a clear improvement in classification performance in our approach. This result shows that the ontology-driven hierarchy can combine genetic risk and provide competitive performance in the classification of autism.



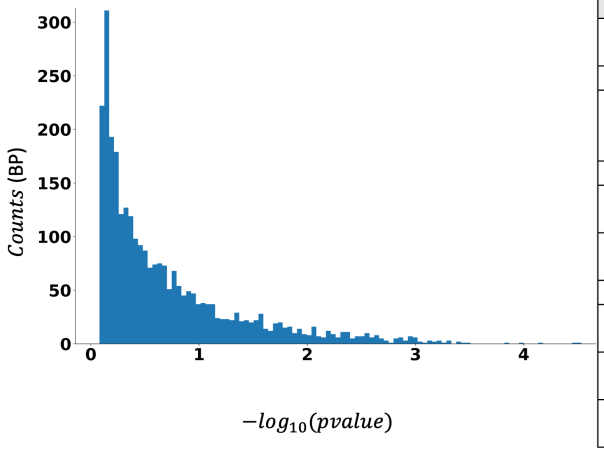
Description	GO Term	-log$_{10}$(pval) FDR
Phenylalanine-tRNA ligase complex	GO:0009328	4.02
Scrib-APC-beta-catenin complex	GO:0034750	4.02
Podocyte foot	GO:0098846	4.04
Amyloid-beta complex	GO:0106003	4.05
PCSK9-LDLR complex	GO:1990666	4.05
CRD-mediated mRNA stability complex	GO:0070937	4.03
BRCA2-MAGE-D1 complex	GO:0033593	4.02
Cellularization cleavage furrow	GO:0110070	4.02
Reelin complex	GO:0110157	4.04
Serine-pyruvate aminotransferase complex	GO:0005969	4.02

Figure 6-9. The histogram of the average $-\log(pvalues)$ across 50 bootstrap trials, testing the significance of the gene scores generated by the root nodes of the ontology of *cellular components*. The p-values are generated by a two-sample t-test. On the left, we show the histogram of the p-values. On the right, we report the description of the top 10 root nodes and their p-values after FDR correction.

Properties of Encoded Genetic Scores: The final layer of our encoding strategy creates a genetic risk score for each root node. The risk scores are the accumulation of all the genetic risks distributed across the child nodes. In this analysis, we explore two ontologies and the properties of the encoded genetic risk scores.

First, we explore the ontology based on the involvement of genes in cellular structures [56]. After model selection, the ontology consists of 5 layers with 1797 nodes and 423 root nodes. We perform a two-sample t-test (described in Section 6.2.2) to obtain a significance level for each risk score. In Fig. 6-9 we show the distribution of average $-\log_{10}(pvalue)$ across 50 bootstrap trials. In addition, we report the description of top 10 root nodes and their mean $-\log_{10}(pvalue)$ after FDR correction. The small p-values show strong evidence that the genetic scores captured by the root nodes contain significant group-level differences. Additionally, some of the cellular components like Amyloid-beta complex [227] and PCSK9 [228] have been previously implicated by autism.

Second, we follow a similar strategy to explore the ontology created by the involve-

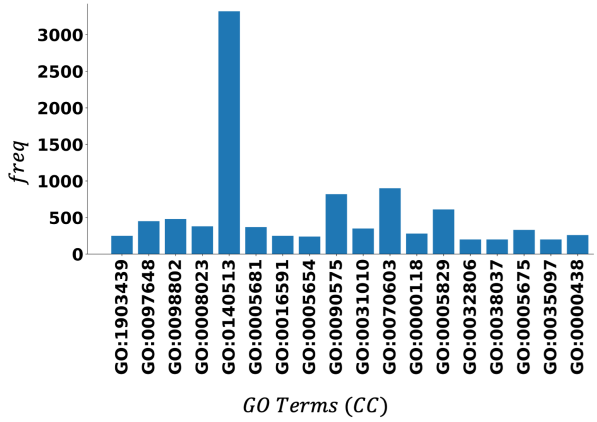


Description	GO Term	-log$_{10}$(pval) FDR
Protein localization to vacuole	GO:0072665	1.26
Seminal vesicle development	GO:0061107	1.16
Apoptotic process in response to mitochondrial fragmentation	GO:0140208	1.01
Isoprenoid transport	GO:0046864	0.91
Response to selenite ion	GO:0072714	0.73
lung connective tissue development	GO:0060427	0.63
Vacuolar sequestering	GO:0043181	0.55
Cyanate metabolic process	GO:0009439	0.54
Response to biotin	GO:0070781	0.54
Subthalamus development	GO:0021539	0.52

Figure 6-10. The histogram of the average $-\log(pvalues)$ across 50 bootstrap trials, testing the significance of the gene scores generated by the root nodes of the ontology of *Biological Processes (BP)*. The p-values are generated by a two-sample t-test. On the left, we show the histogram of the p-values. On the right, we report the description of the top 10 root nodes and their p-values after FDR correction.

ment of genes in larger biological processes. This ontology consists of 14,096 nodes, 5 layers, and 2857 root nodes. Note that this ontology is significantly larger than the previous ontology created by Cellular Components (CC). Like before, we perform the t-test followed by FDR correction. In Fig. 6-10 we show the distribution of p-values and the description of the top 10 root nodes across 50 bootstrap trials. Compared to CC-ontology, here the p-values are relatively large. This could result from the complex interactions between 14,096 nodes. In addition, the genetic risk is divided across 2857 root nodes, so each node only captures a small portion of the genetic risk.

Path Based Analysis: In this analysis, we identify the paths between root and leaf nodes showing differential attention patterns between proband and controls. After performing the likelihood ratio test, we identify the nodes present along each significant path. This analysis gives us a fine-grained understanding of the discriminative paths in the network. In Fig. 6-11 and Fig. 6-12 we show the frequency of the top 20 nodes



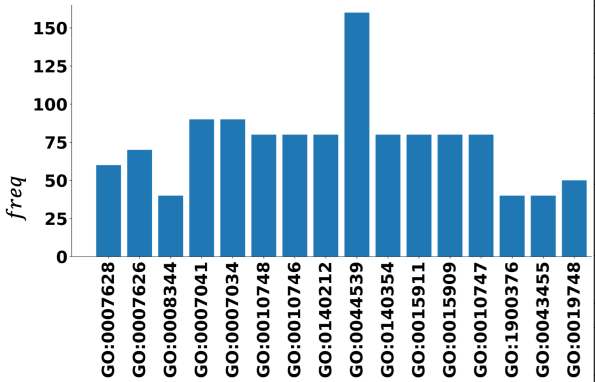
Description	GO Term	*freq*
Nuclear protein-containing complex	GO:0140513	3320
SWI/SNF superfamily-type complex	GO:0070603	900
RNA polymerase II transcription regulator complex	GO:0090575	820
Cytosol	GO:0005829	610
Plasma membrane signaling receptor complex	GO:0098802	480
G protein-coupled receptor complex	GO:0097648	450
Transcription elongation factor complex	GO:0008023	380
Spliceosomal complex	GO:0005681	370
ISWI-type complex	GO:0031010	350
Transcription factor TFIIH holo complex	GO:0005675	330

Figure 6-11. The frequency of the nodes present along a path with significant p-value (< 0.05) across 50 bootstrap trials. The nodes belong to the ontology of *cellular components*. The left image shows the top 20 GO terms and their frequency. The right table gives a brief description of the top 10 GO terms.

that are present along the significant paths across 50 bootstrap trials. We want to note that a node could appear more than 50 times if that node is common across multiple paths. For example, in Fig. 6-9 the GO term *GO:0140513* appeared $3,000$ times across 50 bootstrap trials, which shows that this node is a common node across multiple significant paths. The nodes with high frequency could potentially identify hubs in the network with a strong association with the disorder. As a part of our qualitative exploration, we see that in Fig. 6-11 the path-based analysis identifies GO terms that are involved in transcription [229]. On the other hand, the pathway analysis on the ontology of *biological processes* identifies processes involved in fatty acid transport [230]. This exploratory experiment shows that GUIDE-PRS can be used to extract discriminative biological information about the underlying disorder.

6.2.4 Discussion

We introduce a deep learning based framework that creates multiple genetic risk scores using hierarchical information of biological systems. The main contribution of this



Description	GO Term	*freq*
Long-chain fatty acid import into cell	GO:0044539	160
Lysosomal transport	GO:0007041	90
Vacuolar transport	GO:0007034	90
Negative regulation of long-chain fatty acid import across plasma membrane	GO:0010748	80
Regulation of long-chain fatty acid import across plasma membrane	GO:0010746	80
Regulation of long-chain fatty acid import into cell	GO:0140212	80
Lipid import into cell	GO:0140354	80
Long-chain fatty acid import across plasma membrane	GO:0015911	80
Long-chain fatty acid transport	GO:0015909	80
Positive regulation of long-chain fatty acid import across plasma membrane	GO:0010747	80

Figure 6-12. The frequency of the nodes present along a path with significant p-value (< 0.05) across 50 bootstrap trials. The nodes belong to the ontology of *biological processes*. The left image shows the top 20 GO terms and their frequency. The right table gives a brief description of the top 10 GO terms.

work is the hierarchical graph-based encoder that combines genetic information and creates risk scores associated with the disorder. The graph encoder combines genetic risk and captures complex interactions between biological processes. In comparison to regression or ANN-based approaches, the parameter-sharing property of graph convolution restricts the parameter space, resulting in a highly regularized framework.

The second contribution of this work is the use graph attention to combine the genetic risk. Graph attention provides a data-driven strategy to over-express or suppress the generic risk of specific processes. This approach provides us with a strategy to explore the interaction between nodes in the network. In fact, we show in Fig. 6-12 and Fig. 6-11 that the graph attention can potentially find hubs that are associated with the disorder.

GUIDE-PRS gives us a data-driven and interpretable framework to create genetic risk scores. However, this approach relies heavily on prior biological knowledge, which is often incomplete. For example, our gene-to-node mapping strategy relies on the

knowledge of the functionalities of a gene, but there are many genes whose functions are unknown [231] or cannot be mapped to a node. In those cases, this model removes those genes, thus ignoring valuable genetic information.

6.2.5 Summary

We introduce a novel strategy to create non-linear and interpretable genetic risk scores using graph convolutions. Our model uses the knowledge of gene ontology to combine the genetic risk in a hierarchical fashion. In addition, the model uses graph attention to over-express or suppress the genetic risk of different biological processes, thus providing an automatic and data-driven approach to create interpretable genetic risk scores. In addition, the improved classification performance on autism shows the utility of this model in generating discriminative genetic risk scores. Finally, the exploration of the node-node interactions shows that GUIDE-PRS can be used to extract discriminative biological information about the underlying disorder. We also note that our framework can easily be applied to other clinical traits. In the future, we will apply our framework to identify co-regulated biological processes to explore the shared etiology across multiple clinical traits of autism.

Chapter 7

Discussion and Conclusions

This chapter summarizes the main ideas and frameworks developed in this thesis. We introduce biological knowledge-driven regularized machine learning and deep learning models to integrate and parse imaging and genetics data modalities, with the goal of explaining underlying biology and improving risk prediction of psychiatric disorders. The first part of the thesis (Chapter 3, Chapter 4) introduces models to handle the multifaceted nature of psychiatric disorders. However, in the second part of the thesis (Chapter 5, Chapter 6) we introduce models to parse the complex genetic architectures of neuropsychiatric disorders. Overall, all the models are geared towards providing data-driven solutions to explore interactions between multimodal data while shedding light on unknown causal factors.

In this concluding chapter, we will first summarize the main ideas and the scope of our works. Next, we will provide a brief discussion on potential technical and clinical extensions to the ideas presented.

7.0.1 Overview

In Chapter 3, we introduce a dictionary learning approach to integrate multimodal imaging and genetics data while finding discriminative biomarkers. Our initial approach uses a matrix decomposition framework to identify brain regions with aberrant neural activity while showing a strong association with the polygenic risk score of

schizophrenia. One limitation of this work is that we collapse all the SNP information into a single scalar value, which cannot consider the interactions between the SNPs. To address this limitation, we introduce a generative-discriminative model to integrate SNP-level data with fMRI brain activation maps. Our generative module uses a matrix decomposition framework to integrate multiple data modalities in a single framework. The discriminative module guides the matrix decomposition to find relevant biomarkers and predict disease risk. In addition, we introduce new mathematical and biological knowledge-driven regularization schemes that lead to model stability and improved risk prediction. In Section 3.2.7.4 we demonstrate that our model achieves better classification accuracy than the baselines across all three case-control studies of schizophrenia. In Section 3.2.7.5, we go further and present a strategy to identify a robust set of discriminative biomarkers. Through the meta-analysis, we show that these biomarkers are strongly related to the disease propagation pathway of schizophrenia.

In Chapter 4, we extend our multivariate linear framework to model the non-linear interaction between imaging and genetics using an autoencoder. Our autoencoder frameworks are coupled with Bayesian feature selection and classifiers. The Bayesian module provides interpretability and identifies biomarkers, while the classifiers ensure that the biomarkers contain discriminative information. The autoencoder architecture provides a natural way to integrate new data modalities by adding new encoder-decoder branches. In addition, the autoencoder can handle missing data by freezing the affected part of the network and updating the remaining weights. These properties allow us to take advantage of large data with multiple modalities and missingness.

While the first part of the thesis deals with modeling multiple data modalities, in the next two chapters, we focus on parsing the genetic data and providing insights about the underlying biology of disorders. In Chapter 5, we introduce a deep Bayes variational model to parse complex genetic architectures and identify target variants. We use the

hierarchical Bayesian strategy and a neural network to identify putative causal variants from GWAS summary statistics. Our approach uses a Bayesian variational framework that imposes a binary concrete prior on the set of causal variants. We derive a variational algorithm by minimizing the KL divergence between an approximate density and the posterior probability distribution of the causal configurations. Correspondingly, we use a deep neural network as an inference machine to estimate the parameters of our proposal distribution. The neural network removes the need to handcraft relationships between the input data and the parameter space, thus providing flexibility to handle complex interactions across variants.

The final portion of this thesis deals with the challenges of encoding millions of genetic variants in a single framework while providing interpretability. In Chapter 6, we solve this problem by strategically integrating prior biological knowledge of SNP-gene and gene-pathway interactions in a graph-based framework. Our approach uses graph convolutional networks (GCNs) [131] to leverage the high-dimensional and interconnected genetic relationships. We construct a sparse hierarchical graph using gene ontology [56, 223], which provides a structurally regularized framework to encode the whole genome genotype data. We use this strategy to encode genetic data and integrate it with fMRI data on a population study of schizophrenia. Using detailed ablation and comparative study, we show that this approach uses complete genetic information to improve disease risk prediction. Another critical component of this model is graph attention, which provides an interpretable way to combine genetic risk in a hierarchical fashion. In addition, the graph attention allows us to track the information flow through the graph [136]. In Section 6.1.7.7, we explored the attention values and identified biological processes linked to schizophrenia. Moreover, we take advantage of the graph convolution and graph attention strategy to create interpretable and non-linear genetic risk scores associated with autism. In a preliminary data analysis (Section 6.2.3.2), we found evidence that this approach can identify

potential hubs of biological functions that are associated with the disorder.

7.0.2 Scope and Limitations

Our models introduced in Chapter 3 and Chapter 4 extend the prior works on imaging genetics [11, 13, 14] and provide a joint framework for disease prediction and biomarker identification. The models are not tied to any specific modalities and can be extended to model imaging modalities like structural MRI, and PET with RNA-seq and DNA Methylation data [232, 233]. In addition, the multimodal strategies, coupled with interpretable modules, can be used as novel tools to explore other diseases like Alzheimer's and Parkinson's. However, one key limitation of these approaches is that they cannot model millions of genetic variants in a single framework. The genetic variants are often highly correlated, which leads to problems of singularity, and overfitting [234, 235]. This limitation restricts the model from utilizing complete genetic information about the disorder.

In Section 1, we explain that the genetic risks associated with the polygenic disorder are complex, and often the true signal is hidden behind false positives. Initial finemappiong approaches [39, 89, 96] attempted to identify the true signals from association studies, but their generative assumption often fails to handle spurious signals from non-causal variants. In Chapter 5, we introduce BEATRICE , a robust strategy to handle spurious association signals from non-causal variants. In this work, we have shown that BEATRICE is highly efficient in handling the complexity arising from mutations with infinitesimal effects [188, 203]. Thus, our model can successfully parse polygenic traits and diseases. Additionally, the high coverage and small size of credible sets reported in Fig. 5-4, 5-5, 5-6 show that BEATRICE can successfully prioritize variants in the presence of LD. This property is in stark contrast with the baseline finemapping approaches that generate numerous credible sets that do not contain a causal variant. Taken together, we believe BEATRICE could be useful in

eQTL studies, where multiple variants within a locus can show strong association due to the complex LD structure present in the human genome [54]. Additionally, there may be multiple causal variants within a locus, which adds to the complexity of the finemapping problem [18]. However, one limitation of this model is that the generative process is not geared toward co-localization. Co-localization is a strategy to identify a common set of target variants from GWAS and eQTL studies, with an underlying assumption that a variant appearing in multiple studies is more likely to be causal than others.

In the last Chapter 6, we explore further in the realms of genetics to develop a comprehensive model that can encode whole genome genotype data while providing intuitions about the underlying biology. We introduce a graph convolution model to encode genetic variants using the knowledge of gene ontology. We have used this model to explore an imaging-genetic study of schizophrenia. However, the genetic encoding strategy has a wider application in genetics. For example, we can replace the gene liability scores with single-cell RNA expression [117] data and use the same embedding strategy to identify a target cell associated with a trait. Currently, state-of-the-art approaches use a variation autoencoder [117, 236] for encoding the RNA-seq data. In comparison, our graph-based encoding strategy is interpretable, which can provide further insights into the molecular pathways. Finally, we explore the use of the encoding strategy to create non-linear and interpretable genetic risk scores. Unlike a polygenic risk score, this strategy provides an intuition about the implicated biological functions in a subject-specific label. Each subject has an attention network that can provide insights about the genetic risk along multiple biological processes. Initial evidence in Section 6.2.3.2 shows that this strategy can find potential biological processes and cellular components affected by autism. Additionally, we can use this model to explore the risk shared between multiple disorders like autism, ADHD, and schizophrenia [237]. However, this approach relies heavily on prior biological

knowledge, which is often incomplete. For example, our gene-to-node mapping strategy relies on the knowledge of the functionalities of a gene, but there are many genes whose functions are unknown [231] or cannot be mapped to a node. In those cases, this model removes those genes, thus ignoring valuable genetic information.

7.0.3 Future Extensions

We identify three directions to extend our models for providing insights into the genetic risk associated with complex disorders.

Identification of Colocalized Variants: Most GWAS risk loci lie in the non-coding region [44, 45] of the DNA. A common hypothesis is that the variants alter the individual's genetic risk by affecting the gene expression profile in multiple tissues. Recent approaches [238, 239] try to leverage the GWAS and eQTL studies to identify causal variants, with an underlying assumption that the same causal variant will alter the disease risk and gene expression. The current implementation of BEATRICE cannot accommodate more than one summary statistic. However, we can modify Eq. (5.12) and introduce another likelihood term for eQTL summary statistics originating from the same set of causal variants that influences the GWAS results. From an implementation standpoint, the input to the neural network will now be the concatenation of the summary data, and the neural network output will be a common set of parameters of the binary concrete distribution.

Extending Biological Knowledge with Data-Driven Approaches: The implementation of GUIDE and GUIDE-PRS is restricted by the knowledge of the functionalities of genes in biological processes, which is often incomplete. In order to address these issues, we can take advantage of recent attention-based models [140] that can provide a "soft" assignment for gene-to-node mapping. Instead of hard-coding the mapping function of genes to nodes, we can use this 'soft' assignments technique to

allocate a gene to a node. However, this strategy is unconstrained and may lead to an unstable solution. We can address that by incorporating the hard-coded mapping of genes to nodes as priors. This strategy will ensure that the genes to node mapping will follow the prior, but the model will have the additional flexibility to assign a gene to different nodes.

Explore Shared Genetic Risk Between Disorders: In Chapter 6, we use GUIDE and GUIDE-PRS to explore the biological function underlying schizophrenia and autism. We have viewed the disorders as a binary phenotype. Recent research [36] has shown that these disorders encompass a broad spectrum of phenotypes, with people exhibiting different symptoms to varying degrees. This suggests that the genetic risk for schizophrenia and autism is not due to a single gene or pathway but rather to a complex interplay of multiple genes and pathways. However, genetic interactions often connect genes between functional modules in a coherent manner [240, 241]. We can take advantage of this structure using GUIDE-PRS and potentially identify co-regulated biological processes, taking one step further to identify the underlying causal factors.

In summary, we introduce a suite of machine-learning and deep-learning tools to handle multifaceted disorders while parsing the complex hereditary components. Our models use prior knowledge to develop robust frameworks that are scalable with increasing dimensions and modalities. Additionally, all the models are geared to provide insights into the underlying biology of a disorder. This property makes our models interpretable, thereby serving as invaluable tools for scientific discoveries.

References

1. Belger, A. *et al. The neural circuitry of autism* in *Neurotoxicity Research* **20** (NIH Public Access, 2011), 201–214.
2. Cannon, T. D. *How Schizophrenia Develops: Cognitive and Brain Mechanisms Underlying Onset of Psychosis* in *Trends in Cognitive Sciences* **19** (Elsevier Ltd, 2015), 744–756.
3. Rasetti, R. *et al.* Altered hippocampal-parahippocampal function during stimulus encoding:A potential indicator of genetic liability for schizophrenia. *JAMA Psychiatry* **71,** 236–247 (2014).
4. Callicott, J. H. *et al.* Abnormal fMRI response of the dorsolateral prefrontal cortex in cognitively intact siblings of patients with schizophrenia. *American Journal of Psychiatry* **160,** 709–719 (2003).
5. Vereczkei, A. *et al. Genetic predisposition to schizophrenia: What did we learn and what does the future hold?* in *Neuropsychopharmacologia Hungarica* **13** (Hungarian Association of Psychopharmacology, 2011), 205–210.
6. Frisoni, G. B., Fox, N. C., Jack, C. R., Scheltens, P. & Thompson, P. M. The clinical use of structural MRI in Alzheimer disease. *Nature reviews. Neurology* **6,** 67 (Feb. 2010).
7. Chang, A. S. & Ross, J. S. Diagnostic Neuroradiology: CT, MRI, fMRI, MRS, PET, and Octreotide SPECT. *Meningiomas,* 55–65 (2009).
8. Friston, K. *et al.* Analysis of fMRI Time-Series Revisited. *NeuroImage* **2,** 45–53 (1995).
9. Friston, K. J. (J. *et al. Statistical parametric mapping : the analysis of funtional brain images* 647 (Elsevier/Academic Press, 2007).
10. Worsley, K. J. *et al.* A General Statistical Analysis for fMRI Data. *NeuroImage* **15,** 1–15 (2002).
11. Calhoun, V. D. & Adali, T. Multisubject independent component analysis of fMRI: A decade of intrinsic networks, default mode, and neurodiagnostic discovery. *IEEE Reviews in Biomedical Engineering* **5,** 60–73 (2012).
12. Calhoun, V. D., Adali, T., Pearlson, G. D. & Pekar, J. J. A method for making group inferences from functional MRI data using independent component analysis. *Human brain mapping* **14,** 140–51 (Nov. 2001).
13. Calhoun, V. D., Liu, J. & Adali, T. A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *NeuroImage* **45,** S163 (2009).

14. Kim, M., Won, J. H., Youn, J. & Park, H. Joint-Connectivity-Based Sparse Canonical Correlation Analysis of Imaging Genetics for Detecting Biomarkers of Parkinson's Disease. *IEEE Transactions on Medical Imaging* **39,** 23–34 (Jan. 2020).

15. Pervez, M. T. *et al.* A Comprehensive Review of Performance of Next-Generation Sequencing Platforms. *BioMed Research International* **2022** (2022).

16. Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nature Genetics 2019 51:3* **51,** 431–444 (Feb. 2019).

17. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511,** 421–427 (2014).

18. Abell, N. S. *et al.* Multiple causal variants underlie genetic associations in humans. *Science* **375,** 1247–1254 (Mar. 2022).

19. Lee, H. C. *et al.* Identification of therapeutic targets from genetic association studies using hierarchical component analysis. *BioData Mining* **13** (June 2020).

20. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics 2016 48:3* **48,** 245–252 (Feb. 2016).

21. De Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Computational Biology* **11,** e1004219 (Apr. 2015).

22. Reimand, J. *et al.* Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nature Protocols 2019 14:2* **14,** 482–517 (Jan. 2019).

23. Van Calker, D. & Serchov, T. The "missing heritability"—Problem in psychiatry: Is the interaction of genetics, epigenetics and transposable elements a potential solution? *Neuroscience & Biobehavioral Reviews* **126,** 23–42 (July 2021).

24. Wahbeh, M. H. & Avramopoulos, D. Gene-Environment Interactions in Schizophrenia: A Literature Review. *Genes* **12** (Dec. 2021).

25. Gur, R. E. *et al.* Functional magnetic resonance imaging in schizophrenia. *Dialogues in Clinical Neuroscience* **12,** 333–343 (2010).

26. Karlsgodt, K. H. *et al.* Structural and functional brain abnormalities in schizophrenia. *Current Directions in Psychological Science* **19,** 226–231 (2010).

27. Karam, C. S. *et al.* Signaling pathways in schizophrenia: Emerging targets and therapeutic strategies. *Trends in Pharmacological Sciences* **31,** 381–390 (2010).

28. Rampino, A. *et al.* A Polygenic Risk Score of glutamatergic SNPs associated with schizophrenia predicts attentional behavior and related brain activity in healthy humans. *European Neuropsychopharmacology* **27,** 928–939 (Sept. 2017).

29. Ranlund, S. *et al.* Associations between polygenic risk scores for four psychiatric illnesses and brain structure using multivariate pattern recognition. *NeuroImage : Clinical* **20,** 1026 (Jan. 2018).

30. Chen, Q. *et al.* Schizophrenia polygenic risk score predicts mnemonic hippocampal activity. *Brain* **141,** 1218–1228 (2018).

31. Khundrakpam, B. *et al.* Neural correlates of polygenic risk score for autism spectrum disorders in general population. *Brain Communications* **2** (2020).

32. Sullivan, P. F., Kendler, K. S. & Neale, M. C. Schizophrenia as a Complex Trait: Evidence From a Meta-analysis of Twin Studies. *Archives of General Psychiatry* **60,** 1187–1192 (Dec. 2003).

33. Sandin, S. *et al.* The Heritability of Autism Spectrum Disorder. *JAMA* **318,** 1182–1184 (Sept. 2017).

34. Roofeh, D., Tumuluru, D., Shilpakar, S. & Nimgaonkar, V. L. Genetics of Schizophrenia: Where Has the Heritability Gone? *International Journal of Mental Health* **42,** 5–22 (2013).

35. Hemani, G., Theocharidis, A., Wei, W. & Haley, C. EpiGPU: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. *Bioinformatics* **27,** 1462–1465 (June 2011).

36. Guloksuz, S. & Van Os, J. The slow death of the concept of schizophrenia and the painful birth of the psychosis spectrum. *Psychological Medicine* **48,** 229–244 (Jan. 2018).

37. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature 2022 604:7906* **604** (Apr. 2022).

38. Wang, T. *et al.* Polygenic risk for five psychiatric disorders and cross-disorder and disorder-specific neural connectivity in two independent populations. *NeuroImage : Clinical* **14,** 441 (2017).

39. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198,** 497–508 (Oct. 2014).

40. Maller, J. B. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature Genetics 2012 44:12* **44,** 1294–1301 (12 Oct. 2012).

41. Ng, B. *et al.* Cascading epigenomic analysis for identifying disease genes from the regulatory landscape of GWAS variants. *PLOS Genetics* **17,** e1009918 (Nov. 2021).

42. Joiret, M., Mahachie John, J. M., Gusareva, E. S. & Van Steen, K. Confounding of linkage disequilibrium patterns in large scale DNA based gene-gene interaction studies. *BioData Mining* **12,** 1–23 (June 2019).

43. Brzyski, D. *et al.* Controlling the Rate of GWAS False Discoveries. *Genetics* **205,** 61–75 (1 Jan. 2017).

44. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Human Molecular Genetics* **24,** R102–R110 (Oct. 2015).

45. Cano-Gamez, E. & Trynka, G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Frontiers in Genetics* **11,** 505357 (May 2020).

46. Choi, S. W. *et al.* PRSet: Pathway-based polygenic risk score analyses and software. *PLOS Genetics* **19** (Feb. 2023).

47. Wang *et al.* Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. *Bioinformatics (Oxford, England)* **28,** 229–37 (2012).

48. Batmanghelich, N. K. *et al.* Probabilistic Modeling of Imaging, Genetics and Diagnosis. *IEEE transactions on medical imaging* **35,** 1765–79 (2016).

49. Plitman, E., Patel, R. & Mallar Chakravarty, M. Seeing the bigger picture: multimodal neuroimaging to investigate neuropsychiatric illnesses. *Journal of Psychiatry and Neuroscience* **45,** 147–149 (May 2020).

50. Kang, H. The prevention and handling of the missing data. *Korean Journal of Anesthesiology* **64,** 402–406 (2013).

51. Guan, Y. *et al.* BAYESIAN VARIABLE SELECTION REGRESSION FOR GENOME-WIDE ASSOCIATION STUDIES AND OTHER LARGE-SCALE PROBLEMS. *The Annals of Applied Statistics* **5** (2011).

52. Cui, T. *et al.* Gene–gene interaction detection with deep learning. *Communications Biology 2022 5:1* **5,** 1–12 (Nov. 2022).

53. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82,** 1273–1300 (Dec. 2020).

54. Zou, J. *et al.* Leveraging allelic imbalance to refine fine-mapping for eQTL studies. *PLOS Genetics* **15** (2019).

55. Yang, S. *et al.* Prioritizing genetic variants in GWAS with lasso using permutation-assisted tuning. *Bioinformatics* **36,** 3811–3817 (June 2020).

56. Ashburner, M. *et al.* Gene ontology: Tool for the unification of biology. *Nature Genetics* **25,** 25–29 (2000).

57. Van Hilten, A. *et al.* GenNet framework: Interpretable neural networks for phenotype prediction. *bioRxiv,* 2020.06.19.159152 (2020).

58. Gaudelet, T. *et al.* Unveiling new disease, pathway, and gene associations via multi-scale neural network. *PLOS ONE* **15,** e0231059 (2020).

59. Ghosal, S. *et al.* A generative-predictive framework to capture altered brain activity in fMRI and its association with genetic risk: application to Schizophrenia. *https://doi.org/10.1117/12.2511220* **10949,** 565–575 (Mar. 2019).

60. Ghosal, S. *et al. Bridging Imaging, Genetics, and Diagnosis in a Coupled Low-dimensional Framework* in *MICCAI: Medical Image Computing and Computer Assisted Intervention* **11767 LNCS** (Springer, 2019), 647–655.

61. Ghosal, S. *et al.* A generative-discriminative framework that integrates imaging, genetic, and diagnosis into coupled low dimensional space. *NeuroImage* **238,** 118200 (Sept. 2021).

62. Ghosal, S. *et al. G-MIND: an end-to-end multimodal imaging-genetics framework for biomarker identification and disease classification* in *Medical Imaging 2021: Image Processing* **11596** (SPIE, 2021), 8.

63. Ghosal, S., Schatz, M. & Venkataraman, A. BEATRICE: Bayesian Fine-mapping from Summary Data using Deep Variational Inference. *bioRxiv,* 2003–2023 (2023).

64. Ghosal, S. *et al.* A Biologically Interpretable Graph Convolutional Network to Link Genetic Risk Pathways and Imaging Phenotypes of Disease. *International Conference on Learning Representations.*

65. Gore, J. C. Principles and practice of functional MRI of the human brain. *The Journal of clinical investigation* **112,** 4–9 (July 2003).

66. Heeger, D. J. & Ress, D. What does fMRI tell us about neuronal activity? *Nature Reviews Neuroscience 2002 3:2* **3,** 142–151 (Feb. 2002).

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

68. Van den Heuvel, M. P. & Hulshoff Pol, H. E. Exploring the brain network: A review on resting-state fMRI functional connectivity. *European Neuropsychopharmacology* **20,** 519–534 (Aug. 2010).

69. Calhoun, V. D., Liu, J. & Adali, T. A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *NeuroImage* **45,** S163–S172 (Mar. 2009).

70. Bell, A. J. & Sejnowski, T. J. An Information-Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation* **7,** 1129–1159 (Nov. 1995).

71. Hyvärinen, A. & Oja, E. A Fast Fixed-Point Algorithm for Independent Component Analysis. *Neural Computation* **9,** 1483–1492 (Oct. 1997).

72. Salman, M. S. *et al.* Group ICA for identifying biomarkers in schizophrenia: 'Adaptive' networks via spatially constrained ICA show more sensitivity to group differences than spatio-temporal regression (2019).

73. Holmes, A. P. & Friston, K. J. Generalisability, Random Effects & Population Inference. *NeuroImage* **7,** S754 (May 1998).

74. Wang, G., Muschelli, J. & Lindquist, M. A. Moderated t-tests for group-level fMRI analysis. *NeuroImage* **237,** 118141 (Aug. 2021).

75. Yin, W., Li, L. & Wu, F.-X. Deep learning for brain disorder diagnosis based on fMRI images q.

76. Wang, Z., Childress, A. R., Wang, J. & Detre, J. A. Support vector machine learning-based fMRI data group analysis. *NeuroImage* **36,** 1139–1151 (July 2007).

77. Ben-Hur, A. & Weston, J. *A User's Guide to Support Vector Machines* tech. rep. ().

78. Breiman, L. Random Forests. *Machine Learning* **45,** 5–32 (2001).

79. Srinivasagopalan, S. *et al.* A deep learning approach for diagnosing schizophrenic patients. *Journal of Experimental and Theoretical Artificial Intelligence* **31** (Nov. 2019).

80. Zeng, L. L. *et al.* Multi-Site Diagnostic Classification of Schizophrenia Using Discriminant Deep Learning with Functional Connectivity MRI. *EBioMedicine* **30** (Apr. 2018).

81. Smeland, O. B., Frei, O., Dale, A. M. & Andreassen, O. A. The polygenic architecture of schizophrenia — rethinking pathogenesis and nosology. *Nature Reviews Neurology 2020 16:7* **16,** 366–379 (June 2020).

82. Clarke, T. K. *et al.* Common polygenic risk for autism spectrum disorder (ASD) is associated with cognitive ability in the general population. *Molecular Psychiatry 2016 21:3* **21,** 419–425 (Mar. 2015).

83. Mahmoud, M. *et al.* Structural variant calling: The long and the short of it. *Genome Biology* **20,** 1–14 (Nov. 2019).

84. Hart, J. R., Johnson, M. D. & Barton, J. K. Single-nucleotide polymorphism discovery by targeted DNA photocleavage. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 14040–14044 (Sept. 2004).

85. Uffelmann, E. *et al.* Genome-wide association studies. *Nature Reviews Methods Primers 2021 1:1* **1,** 1–21 (Aug. 2021).

86. Chen, W. *et al.* Fine mapping causal variants with an approximate bayesian method using marginal test statistics. *Genetics* **200,** 719–736 (July 2015).

87. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five Years of GWAS Discovery. *The American Journal of Human Genetics* **90,** 7–24 (Jan. 2012).

88. Spain, S. L. & Barrett, J. C. Strategies for fine-mapping complex traits. *Human Molecular Genetics* **24,** R111–R119 (Oct. 2015).

89. Schaid, D. J. *et al.* From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics 2018 19:8* **19** (2018).

90. Hutchinson, A., Watson, H. & Wallace, C. Improving the coverage of credible sets in Bayesian genetic fine-mapping. *PLOS Computational Biology* **16,** e1007829 (4 Apr. 2020).

91. Cho, S. *et al.* Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis. *BMC proceedings* **3** (Dec. 2009).

92. Sabourin, J. *et al.* Fine-Mapping Additive and Dominant SNP Effects Using Group-LASSO and Fractional Resample Model Averaging. *Genetic Epidemiology* **39** (2015).

93. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58,** 267–288 (1 Jan. 1996).

94. Carbonetto, P. & Stephens, M. Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies. *Bayesian Analysis* **7,** 73–108 (2012).

95. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32,** 1493–1501 (May 2016).

96. Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fine-mapping from summary data with the "Sum of Single Effects" model. *PLOS Genetics* **18,** e1010299 (7 July 2022).

97. Lewis, C. M. & Vassos, E. Polygenic risk scores: From research tools to clinical instruments. *Genome Medicine* **12,** 1–11 (May 2020).

98. Choi, S. W., Mak, T. S. H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols 2020 15:9* **15,** 2759–2772 (July 2020).

99. Takahashi, N. *et al.* Polygenic risk score analysis revealed shared genetic background in attention deficit hyperactivity disorder and narcolepsy. *Translational Psychiatry 2020 10:1* **10,** 1–9 (Aug. 2020).

100. Lin, W. Y., Huang, C. C., Liu, Y. L., Tsai, S. J. & Kuo, P. H. Polygenic approaches to detect gene–environment interactions when external information is unavailable. *Briefings in Bioinformatics* **20,** 2236–2252 (Nov. 2019).

101. Pergola, G., Penzel, N., Sportelli, L. & Bertolino, A. Lessons Learned From Parsing Genetic Risk for Schizophrenia Into Biological Pathways. *Biological Psychiatry* **94,** 121–130 (July 2023).

102. Corley, E. *et al.* Microglial-expressed genetic risk variants, cognitive function and brain volume in patients with schizophrenia and healthy controls. *Translational Psychiatry 2021 11:1* **11,** 1–8 (Sept. 2021).

103. Hariri, A. & Weinberger, D. Imaging genomics. *British Medical Bulletin* **65,** 259–270 (2003).

104. Nathoo, F. S., Kong, L. & Zhu, H. A review of statistical methods in imaging genetics. *Canadian Journal of Statistics* **47,** 108–131. arXiv: 1707.07332 (Mar. 2019).

105. Stein, J. L. *et al.* Voxelwise genome-wide association study (vGWAS). *NeuroImage* **53,** 1160 (Nov. 2010).

106. Hibar, D. P. *et al.* Voxelwise gene-wide association study (vGeneWAS): Multivariate gene-based association testing in 731 elderly subjects. *NeuroImage* **56,** 1875–1891 (June 2011).

107. Kim, K. I. & van de Wiel, M. A. Effects of dependence in high-dimensional multiple testing problems. *BMC Bioinformatics* **9,** 1–12 (Feb. 2008).

108. Liu, J. *et al.* A Review of Multivariate Analyses in Imaging Genetics. *Frontiers in Neuroinformatics* **8,** 29 (2014).

109. Vounou, M., Nichols, T. E. & Montana, G. Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. *NeuroImage* **53,** 1147–1159 (Nov. 2010).

110. Chi, E. C. *et al.* Imaging genetics via sparse canonical correlation analysis. *Proceedings - International Symposium on Biomedical Imaging,* 740–743 (2013).

111. Kim, M. *et al.* Multi-task learning based structured sparse canonical correlation analysis for brain imaging genetics. *Medical Image Analysis* **76,** 102297 (Feb. 2022).

112. Witten, D. M., Tibshirani, R. & Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10,** 515–534 (July 2009).

113. Yan, J. *et al.* JOINT EXPLORATION AND MINING OF MEMORY-RELEVANT BRAIN ANATOMIC AND CONNECTOMIC PATTERNS VIA A THREE-WAY ASSOCIATION MODEL. *Proceedings. IEEE International Symposium on Biomedical Imaging* **2018,** 6 (May 2018).

114. Ke, F., Kong, W. & Wang, S. Identifying Imaging Genetics Biomarkers of Alzheimer's Disease by Multi-Task Sparse Canonical Correlation Analysis and Regression. *Frontiers in Genetics* **12** (Aug. 2021).

115. Pearlson, G. D. *et al.* An introductory review of parallel independent component analysis (p-ICA) and a guide to applying p-ICA to genetic data and imaging phenotypes to identify disease-associated biological pathways and systems in common complex disorders. *Frontiers in Genetics* **6** (2015).

116. Wolf, R. C. *et al.* A Neural Signature of Parkinsonism in Patients With Schizophrenia Spectrum Disorders: A Multimodal MRI Study Using Parallel ICA. *Schizophrenia Bulletin* **46,** 999–1008 (July 2020).

117. Eraslan, G. *et al.* Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics* **20** (July 2019).

118. Shen, L. & Thompson, P. M. Brain Imaging Genomics: Integrated Analysis and Machine Learning. *Proceedings of the IEEE* **108** (Jan. 2020).

119. Najafabadi, M. M. *et al.* Deep learning applications and challenges in big data analytics. *Journal of Big Data* **2,** 1 (2015).

120. Fred Agarap, A. M. Deep Learning using Rectified Linear Units (ReLU). *arXiv,* 1803.08375v2.

121. He, K., Zhang, X., Ren, S. & Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *arXiv,* 1502.01852v1.

122. Ko, W., Jung, W., Jeon, E. & Suk, H. I. A Deep Generative-Discriminative Learning for Multimodal Representation in Imaging Genetics. *IEEE Transactions on Medical Imaging* **41,** 2348–2359 (Sept. 2022).

123. Behrad, F. & Saniee Abadeh, M. An overview of deep learning methods for multimodal medical data mining. *Expert Systems with Applications* **200,** 117006 (Aug. 2022).

124. Patel, S., Park, M. T. M., Knight, J. & Knight, J. Gene Prioritization for Imaging Genetics Studies Using Gene Ontology and a Stratified False Discovery Rate Approach. *Frontiers in Neuroinformatics* **10,** 14 (2016).

125. Dong, G., Liao, G., Liu, H. & Kuang, G. A Review of the Autoencoder and Its Variants: A Comparative Perspective from Target Recognition in Synthetic-Aperture Radar Images. *IEEE Geoscience and Remote Sensing Magazine* **6,** 44–68 (Sept. 2018).

126. Abdelaziz, M., Wang, T. & Elazab, A. Fusing Multimodal and Anatomical Volumes of Interest Features Using Convolutional Auto-Encoder and Convolutional Neural Networks for Alzheimer's Disease Diagnosis. *Frontiers in Aging Neuroscience* **14,** 812870 (Apr. 2022).

127. Venugopalan, J., Tong, L., Hassanzadeh, H. R. & Wang, M. D. Multimodal deep learning models for early detection of Alzheimer's disease stage. *Scientific Reports* **11,** 3254 (Dec. 2021).

128. Jaques, N. *et al. Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction* in *2017 7th International Conference on Affective Computing and Intelligent Interaction, ACII 2017* **2018-January** (Institute of Electrical and Electronics Engineers Inc., 2018), 202–208.

129. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision* **128,** 336–359 (Oct. 2016).

130. Lundberg, S. M. *et al. A Unified Approach to Interpreting Model Predictions* in *Advances in Neural Information Processing Systems* **30** (2017).

131. Kipf, T. N. & Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings.* arXiv: 1609.02907 (Sept. 2016).

132. Lei, D. *et al.* Graph Convolutional Networks Reveal Network-Level Functional Dysconnectivity in Schizophrenia. *Schizophrenia Bulletin* **48,** 881 (July 2022).

133. Fout, A. *et al.* Protein Interface Prediction using Graph Convolutional Networks. *Advances in Neural Information Processing Systems* **30** (2017).

134. Stokes, J. M. *et al.* A Deep Learning Approach to Antibiotic Discovery. *Cell* **180** (Feb. 2020).

135. Yuan, Y. & Bar-Joseph, Z. GCNG: graph convolutional networks for inferring gene interaction from spatial transcriptomics data. *Genome biology* (Dec. 2020).

136. Velicković, P. *et al.* Graph attention networks. *arXiv.* arXiv: 1710.10903 (Oct. 2017).

137. Zhang, J. *et al. Predicting Disease-related RNA Associations based on Graph Convolutional Attention Network* in *Proceedings - 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019* (Institute of Electrical and Electronics Engineers Inc., 2019).

138. Schapke, J. *et al.* EPGAT: Gene Essentiality Prediction With Graph Attention Networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2021).

139. Bruna, J. *et al.* Spectral Networks and Locally Connected Networks on Graphs. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings.* arXiv: 1312.6203 (Dec. 2013).

140. Lee, J. *et al.* Self-Attention Graph Pooling. *36th International Conference on Machine Learning, ICML 2019.* arXiv: 1904.08082 (Apr. 2019).

141. Ying, R. *et al.* Hierarchical Graph Representation Learning with Differentiable Pooling. *Advances in Neural Information Processing Systems* (2018).

142. Mi & others. Protocol Update for Large-scale Genome and Gene Function Analysis with The PANTHER Classification System (v.14.0). *Nature Protocols* **14,** 703–721 (2019).

143. Fan, L. *et al.* The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. *Cerebral cortex (New York, N.Y. : 1991)* **26,** 3508–26 (2016).

144. Sanders, S. J. *et al.* Multiple Recurrent De Novo CNVs, Including Duplications of the 7q11.23 Williams Syndrome Region, Are Strongly Associated with Autism. *Neuron* **70,** 863–885 (June 2011).

145. Lord, C., Rutter, M. & Le Couteur, A. Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of autism and developmental disorders* **24,** 659–685 (Oct. 1994).

146. Lord, C. *et al.* Autism diagnostic observation schedule: a standardized observation of communicative and social behavior. *Journal of autism and developmental disorders* **19,** 185–212 (June 1989).

147. Lam, M. *et al.* RICOPILI: Rapid Imputation for COnsortias PIpeLIne. *Bioinformatics* **36,** 930–933 (Feb. 2020).

148. Lajonchere, C. M. & Consortium, A. Changing the Landscape of Autism Research: The Autism Genetic Resource Exchange. *Neuron* **68,** 187 (Oct. 2010).

149. Won, J. H., Kim, M., Youn, J. & Park, H. Prediction of age at onset in Parkinson's disease using objective specific neuroimaging genetics based on a sparse canonical correlation analysis. *Scientific Reports 2020 10:1* **10,** 1–12 (July 2020).

150. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology 2023,* 1–12 (May 2023).

151. Irofti, P. & Dumitrescu, B. Regularized algorithms for dictionary learning. *IEEE International Conference on Communications* **2016-August,** 439–442 (Aug. 2016).

152. Batmanghelich *et al.* Generative-discriminative basis learning for medical imaging. *IEEE Trans.* **31,** 51–69 (2012).

153. Lai, R. & Osher, S. A splitting method for orthogonality constrained problems. *Journal of Scientific Computing* **58,** 431–449 (2014).

154. Schönemann, P. H. A generalized solution of the orthogonal procrustes problem. *Psychometrika* **31,** 1–10 (Mar. 1966).

155. Ben-Hur, A. & Weston, J. A user's guide to support vector machines. *Methods in molecular biology (Clifton, N.J.)* **609,** 223–239 (2010).

156. Sim, A., Tsagkrasoulis, D. & Montana, G. Random forests on distance matrices for imaging genetics studies. *Statistical Applications in Genetics and Molecular Biology* **12,** 757–786 (Dec. 2013).

157. Roy, M. H. & Larocque, D. Robustness of random forests for regression. *Journal of Nonparametric Statistics* **24,** 993–1006 (2012).

158. Rachakonda, S., Liu, J. & Calhoun, V. *Fusion ICA Toolbox (FIT) Manual* tech. rep. (2012).

159. *Kolmogorov–Smirnov Test* in *The Concise Encyclopedia of Statistics* (Springer New York, New York, NY, 2008), 283–287.

160. Dickinson, D. *et al.* Cognitive Factor Structure and Invariance in People With Schizophrenia, Their Unaffected Siblings, and Controls. *Schizophrenia Bulletin* **37,** 1157–1167 (2011).

161. Sambataro, F. *et al.* Treatment with Olanzapine is Associated with Modulation of The Default Mode Network in Patients with Schizophrenia. *Neuropsychopharmacology* **35,** 904–912 (2010).

162. Tor D., W. NeuroSynth: a new platform for large-scale automated synthesis of human functional neuroimaging data. *Frontiers in Neuroinformatics* **5** (2011).

163. Dayem Ullah, A. Z. *et al.* SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine. *Nucleic Acids Research* **46,** 109–113 (2018).

164. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. **45,** 580–585 (2013).

165. Luciano, M. *et al.* Association analysis in over 329,000 individuals identifies 116 independent variants influencing neuroticism. *Nature Genetics* **50,** 6–11 (2018).

166. Di Giorgio, A. *et al.* Evidence that hippocampal-parahippocampal dysfunction is related to genetic risk for schizophrenia. *Psychological medicine* **43,** 1661–1671 (2013).

167. Zhu, Y. *et al.* Reduced functional connectivity between bilateral precuneus and contralateral parahippocampus in schizotypal personality disorder. *BMC Psychiatry* **17** (Feb. 2017).

168. Goes, F. S. *et al.* Genome-wide association study of schizophrenia in Ashkenazi Jews. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **168,** 649–659 (2015).

169. Grippo, L. & Sciandrone, M. On the convergence of the block nonlinear Gauss-Seidel method under convex constraints. **26,** 127–136 (2000).

170. Li, Q., Zhu, Z. & Tang, G. *Alternating Minimizations Converge to Second-Order Optimal Solutions* in *Proceedings of the 36th International Conference on Machine Learning* **97** (PMLR, 2019), 3935–3943.

171. Stram, D. O. *Tag SNP selection for association studies* in *Genetic Epidemiology* **27** (2004), 365–374.

172. Luvsannyam, E. *et al.* Neurobiology of Schizophrenia: A Comprehensive Review. *Cureus* **14** (Apr. 2022).

173. Quinde-Zlibut, J. M. *et al.* Multifaceted empathy differences in children and adults with autism. *Scientific Reports* **11,** 19503 (2021).

174. Muller-Nedebock, A. C. *et al.* Different pieces of the same puzzle: a multifaceted perspective on the complex biological basis of Parkinson's disease. *npj Parkinson's Disease 2023 9:1* **9,** 1–11 (July 2023).

175. Kay, S. R. *et al.* The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin* **13,** 261–276 (1987).

176. Bowie, C. R. *et al.* Cognitive deficits and functional outcome in schizophrenia. *Neuropsychiatric Disease and Treatment* **2,** 531–536 (2006).

177. Hiremath, C. S. *et al.* Emerging behavioral and neuroimaging biomarkers for early and accurate characterization of autism spectrum disorders: a systematic review. *Translational Psychiatry 2021 11:1* **11,** 1–12 (Jan. 2021).

178. Maddison, C. J. *et al.* The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings.* arXiv: 1611.00712 (Nov. 2016).

179. Jang, E. *et al.* Categorical Reparameterization with Gumbel-Softmax. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings.* arXiv: 1611.01144 (2016).

180. Gal, Y. & Kendall, A. Concrete Dropout. arXiv: 1705.07832v1.

181. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. arXiv: 1312.6114v10.

182. Gal, Y. & Ghahramani, Z. *Dropout as a bayesian approximation: Representing model uncertainty in deep learning* in *international conference on machine learning* (2016), 1050–1059.

183. Dean, B. Is Schizophrenia The Price of Human Central Nervous System Complexity? *Australian and New Zealand Journal of Psychiatry* **43,** 13–24 (2009).

184. *Dysregulation of Neural Calcium Signaling in Alzheimer Disease, Bipolar Disorder and Schizophrenia* in *Prion* **7** (Landes Bioscience, 2013), 2–13.

185. Ben Said, A., Mohamed, A., Elfouly, T., Harras, K. & Wang, Z. J. *Multimodal deep learning approach for Joint EEG-EMG Data compression and classification* in *IEEE Wireless Communications and Networking Conference, WCNC* (Institute of Electrical and Electronics Engineers Inc., 2017).

186. Kingma, D. P. *et al. Adam: A method for stochastic optimization* in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (International Conference on Learning Representations, ICLR, 2015).

187. Docherty, A. R. *et al.* Genome-wide gene pathway analysis of psychotic illness symptom dimensions based on a new schizophrenia-specific model of the OPCRIT. *Schizophrenia Research* **164,** 181–186 (May 2015).

188. Cheng, W., Ramachandran, S. & Crawford, L. Estimation of non-null SNP effect size distributions enables the detection of enriched genes underlying complex traits. *PLOS Genetics* **16,** e1008855 (6 June 2020).

189. Servin, B. & Stephens, M. Imputation-Based Analysis of Association Studies: Candidate Regions and Quantitative Traits. *PLOS Genetics* **3,** e114 (7 July 2007).

190. Pirinen, M., Donnelly, P. & Spencer, C. C. A. EFFICIENT COMPUTATION WITH A LINEAR MIXED MODEL ON LARGE-SCALE DATA SETS WITH APPLICATIONS TO GENETIC STUDIES. *The Annals of Applied Statistics* **7,** 369–390 (2013).

191. Wakefield, J. Bayes factors for Genome-wide association studies: Comparison with P-values. *Genetic Epidemiology* **33,** 79–86 (2009).

192. Kingma, D. P. & Welling, M. *Auto-Encoding Variational Bayes* 2013.

193. LaPierre, N. *et al.* Identifying causal variants by fine mapping across multiple studies. *PLOS Genetics* **17,** e1009733 (Sept. 2021).

194. Hans, C., Dobra, A. & West, M. Shotgun Stochastic Search for "Large p" Regression. *https://doi.org/10.1198/016214507000000121* **102,** 507–516 (478 June 2012).

195. Wang, Z. *et al.* An autoimmune pleiotropic SNP modulates IRF5 alternative promoter usage through ZBTB3-mediated chromatin looping. *Nature Communications 2023 14:1* **14** (Mar. 2023).

196. Albiñana, C. *et al.* Genetic correlates of vitamin D-binding protein and 25-hydroxyvitamin D in neonatal dried blood spots. *Nature Communications 2023 14:1* **14,** 1–16 (Feb. 2023).

197. Li, Y. *et al.* Cross-ancestry genome-wide association study and systems-level integrative analyses implicate new risk genes and therapeutic targets for depression. *medRxiv,* 2023.02.24.23286411 (Mar. 2023).

198. Davis, J. & Goadrich, M. The relationship between precision-recall and ROC curves. *ACM International Conference Proceeding Series* **148,** 233–240 (2006).

199. Dimitromanolakis, A., Xu, J., Krol, A. & Briollais, L. sim1000G: A user-friendly genetic variant simulator in R for unrelated individuals and family-based designs. *BMC Bioinformatics* **20,** 1–9 (1 Jan. 2019).

200. Belmont, J. W. *et al.* The International HapMap Project. *Nature 2004 426:6968* **426** (2003).

201. Meyer, H. V. & Birney, E. PhenotypeSimulator: A comprehensive framework for simulating multi-trait, multi-locus genotype to phenotype relationships. *Bioinformatics* **34,** 2951–2956 (17 Sept. 2018).

202. Cui, R. *et al.* Improving fine-mapping by modeling infinitesimal effects. *bioRxiv* (2022).

203. Zhang, Y. *et al.* Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nature Genetics 2018 50:9* **50,** 1318–1326 (Aug. 2018).

204. Orliac, E. J. *et al.* Improving GWAS discovery and genomic prediction accuracy in biobank data. *Proceedings of the National Academy of Sciences of the United States of America* **119,** e2121279119 (Aug. 2022).

205. Bhattacharjee, S. *et al.* A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *American Journal of Human Genetics* **90,** 821–835 (2012).

206. Jiao, H. *et al.* Genome-wide interaction and pathway association studies for body mass index. *Frontiers in Genetics* **10,** 437351 (May 2019).

207. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research* **45,** D353–D361 (Jan. 2017).

208. Mi, H. *et al.* Large-scale gene function analysis with the panther classification system. *Nature Protocols* **8,** 1551–1566 (2013).

209. Alcocer-Cuarón, C. *et al.* Hierarchical structure of biological systems: A bioengineering approach. *Bioengineered* **5,** 73 (2014).

210. Gao, H. & Ji, S. Graph U-Nets. *CoRR* **abs/1905.05178** (2019).

211. Wong, E. S. *et al.* Interplay of cis and trans mechanisms driving transcription factor binding and gene expression evolution. *Nature Communications* **8** (2017).

212. Wang, H. *et al.* Evidence of a dissociation pattern in default mode subnetwork functional connectivity in schizophrenia. *Scientific Reports 2015 5:1* **5** (2015).

213. Yarkoni, T. *et al.* Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods* **8,** 665–670 (2011).

214. Carter, J. D. *et al.* Attention deficits in schizophrenia — Preliminary evidence of dissociable transient and sustained deficits. *Schizophrenia Research* **122** (2010).

215. Guo, J. *et al.* Memory and Cognition in Schizophrenia. *Molecular psychiatry* **24** (2019).

216. Beel, J. *et al.* Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries* **17** (2016).

217. Van Der Maaten, L. *et al. Visualizing Data using t-SNE* in *Journal of Machine Learning Research* **9** (2008).

218. Ormel, P. R. *et al.* Characterization of macrophages from schizophrenia patients. *npj Schizophrenia* **3,** 41 (2017).

219. Van Kesteren, C. F. *et al.* Immune involvement in the pathogenesis of schizophrenia: A meta-analysis on postmortem brain studies. *Translational Psychiatry* **7** (2017).

220. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics 2015 16:4* **16,** 197–212 (Feb. 2015).

221. Castro, C. P., Diehl, A. G. & Boyle, A. P. Challenges in screening for de novo noncoding variants contributing to genetically complex phenotypes. *Human Genetics and Genomics Advances* **4,** 100210 (July 2023).

222. Cosgrove, D. *et al.* MiR-137-derived polygenic risk: effects on cognitive performance in patients with schizophrenia and controls. *Translational Psychiatry 2017 7:1* **7,** e1012–e1012 (Jan. 2017).

223. Hill, D. P., Smith, B., McAndrews-Hill, M. S. & Blake, J. A. Gene Ontology annotations: What they mean and where they come from. *BMC Bioinformatics* **9,** 1–9 (Apr. 2008).

224. Anney, R. *et al.* A genome-wide scan for common alleles affecting risk for autism. *Human Molecular Genetics* **19,** 4072–4082 (Oct. 2010).

225. Matoba, N. *et al.* Common genetic risk variants identified in the SPARK cohort support DDHD2 as a candidate risk gene for autism. *Translational Psychiatry 2020 10:1* **10,** 1–14 (Aug. 2020).

226. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26,** 2190–2191 (Sept. 2010).

227. Li, X. *et al.* Regressive Autism Spectrum Disorder: High Levels of Total Secreted Amyloid Precursor Protein and Secreted Amyloid Precursor Protein-α in Plasma. *Frontiers in Psychiatry* **13,** 809543 (Mar. 2022).

228. Salem, S., Mosaad, R., Lotfy, R., Ashaat, E. & Ismail, S. PCSK9 Involvement in Autism Etiology: Sequence Variations, Protein Concentration, and Promoter Methylation. *Archives of Medical Research* **54,** 102860 (Sept. 2023).

229. Ayhan, F. & Konopka, G. Regulatory genes and pathways disrupted in autism spectrum disorders. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* **89,** 57–64 (Mar. 2019).

230. Maekawa, M. *et al.* Investigation of the fatty acid transporter-encoding genes SLC27A3 and SLC27A4 in autism. *Scientific Reports 2015 5:1* **5,** 1–15 (Nov. 2015).

231. Rocha, J. J. *et al.* Functional unknomics: Systematic screening of conserved genes of unknown function. *PLOS Biology* **21,** e3002222 (Aug. 2023).

232. Mariani Wigley, I. L. C. *et al.* Neuroimaging and DNA Methylation: An Innovative Approach to Study the Effects of Early Life Stress on Developmental Plasticity. *Frontiers in Psychology* **12,** 672786 (May 2021).

233. Arnatkeviciute, A., Fulcher, B. D., Bellgrove, M. A. & Fornito, A. Imaging Transcriptomics of Brain Disorders. *Biological Psychiatry Global Open Science* **2,** 319–331 (Oct. 2022).

234. Toloşi, L. & Lengauer, T. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics* **27,** 1986–1994 (July 2011).

235. Khaire, U. M. & Dhanalakshmi, R. Stability of feature selection algorithm: A review. *Journal of King Saud University - Computer and Information Sciences* **34,** 1060–1073 (Apr. 2022).

236. Tran, D. *et al.* Fast and precise single-cell data analysis using a hierarchical autoencoder. *Nature Communications 2021 12:1* **12,** 1–10 (Feb. 2021).

237. Carroll, L. S. & Owen, M. J. Genetic overlap between autism, schizophrenia and bipolar disorder. *Genome Medicine* **1,** 102 (2009).

238. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *American Journal of Human Genetics* **99,** 1245–1260 (Dec. 2016).

239. Foley, C. N. *et al.* A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nature Communications 2021 12:1* **12,** 1–18 (Feb. 2021).

240. Fang, G. *et al.* Discovering genetic interactions bridging pathways in genome-wide association studies. *Nature Communications 2019 10:1* **10,** 1–18 (Sept. 2019).

241. VanderSluis, B. *et al.* Integrating genetic and protein-protein interaction networks maps a functional wiring diagram of a cell. *Current opinion in microbiology* **45,** 170 (Oct. 2018).