Machine learning in genomic classification and stratification of neuropsychiatric disorders using whole exome sequencing data

Sameer Sardaar

Department of Human Genetics
McGill University
Montreal, Quebec, Canada
April 2021

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science in Human Genetics

Abstract

Neuropsychiatric disorders have a major genetic component to their etiology and manifest with varying and complex set of clinical features. This renders their classification and treatment challenging. Recent advances in sequencing technologies and increasing sample sizes have allowed for the discovery of a large number of risk variants associated with these disorders. However, most studies have focused on univariate statistical approaches for the genetic characterization of these complex disorders. Given their polygenic and multifactorial etiology, advanced algorithms that can model higher interactions among variants/genes are needed. The field of machine learning (ML) is concerned with developing algorithms that learn with experience and are able to find complex relationships among input variables for pattern discovery and predictive modelling. ML algorithms developed on genomic data, such as data from whole exome-sequencing technologies, can assist in shedding light into the complex architecture of neuropsychiatric disorders.

This thesis focuses on applications of ML algorithms for classification and stratification of autism (ASD), schizophrenia (SCZ) and bipolar disorder (BD) using whole-exome sequencing data. More specifically, we show that ASD and SCZ patients can be successfully distinguished from each other based on their rare variants identified through whole-exome sequencing, and we identify the top variants and genes contributing most to our best performing model. Similarly, we show in another dataset that BD patients can be distinguished from SCZ and control individuals using all variants as input features and highlight the most important variants in the best performing model. In addition, we implemented a novel topic modelling approach for clustering and classification of SCZ, BD and control individuals that identified a genetically

more homogenous subcluster of SCZ individuals. Lastly, we developed a novel graph/network approach to model genomic mutations in familial datasets and propose a relational graph neural network for node classification of disease status in siblings discordant for ASD.

Résumé

Les troubles neuropsychiatriques ont une composante génétique majeure dans leur étiologie et se manifestent par un ensemble de caractéristiques cliniques variées et complexes. Ceci complique leur classification et leur traitement. Les progrès récents des technologies de séquençage et l'augmentation de la taille des échantillons ont permis la découverte d'un grand nombre de variantes conférant un risque accru pour ces troubles. Cependant, la plupart des études se sont concentrées sur des approches statistiques unidimensionnelles (univariées) pour la caractérisation génétique de ces troubles complexes. Compte tenu de leur étiologie polygénique et multifactorielle, des algorithmes avancés capables de modéliser des interactions plus élevées entre les variantes/gènes sont nécessaires. Le domaine de l'apprentissage automatique concerne le développement d'algorithmes informatiques qui apprennent avec l'expérience. Ceci peut identifier des relations complexes entre les variables d'entrée pour la découverte de modèles et la modélisation prédictive. Les algorithmes d'apprentissage automatique développés à partir de données génomiques, telles que les données issues de technologies de séquençage d'exomes entières, peuvent éclaircir l'architecture complexe des troubles neuropsychiatriques.

Cette thèse se concentre sur les applications des algorithmes d'apprentissage automatique pour la classification et la stratification de l'autisme (TSA), de la schizophrénie (SCZ) et du trouble bipolaire (BD) à l'aide de données de séquençage de l'exome entier. Plus précisément, nous montrons que les patients atteints de TSA et de SCZ peuvent être discerner, avec succès, les uns des autres en fonction de leurs variantes conférant un risque accru identifiées par séquençage de l'exome entier, et nous identifions les principales variantes et

gènes contribuant le plus à notre modèle le plus performant. De même, nous montrons dans un autre ensemble de données que les patients BD peuvent être discernés des patients SCZ et témoins en utilisant toutes les variantes comme caractéristiques d'entrée et nous mettons en évidence les variantes les plus importantes dans le modèle le plus performant. En outre, nous avons mis en œuvre une nouvelle approche de modélisation de sujet pour le regroupement et la classification des individus SCZ, BD et témoins qui a identifié un sous-regroupement génétiquement plus homogène d'individus avec SCZ. Enfin, nous avons développé une nouvelle approche graphique / réseau pour modéliser les mutations génomiques dans les ensembles de données familiales et nous proposons un réseau neuronal graphique relationnel pour la classification des nœuds de l'état de la maladie chez les frères et sœurs discordants pour les TSA.

Table of Contents

List of Abbreviations

ABCC3	ATP-binding cassette, sub-family C (CFTR/MRP), member 3
ADHD	Attention deficit hyperactivity disorder
AKAP1	A kinase (PRKA) anchor protein 1
ASD	Autism spectrum disorder
BD	Bipolar disorder
CACNA1S	calcium channel, voltage-dependent, L type, alpha 1S subunit
CART	Classification and regression tree
CCDC155	coiled-coil domain containing 155
CDSN	corneodesmosin
CTL	Control
dbGaP	The database of Genotypes and Phenotypes
DNA	Deoxyribonucleic acid
EFHB	EF-hand domain family, member B
ETM	Embedded topic model
FAN1	FANCD2/FANCI-associated nuclease 1
fMRI	Functional magnetic resonance imaging
GBM	Gradient boosting machine
GCN	Graph convolutional network
GNN	Graph neural network
GPU	Graphical processing unit
GWAS	Genome-wide association study
HERC2	HECT and RLD domain containing E3 ubiquitin protein ligase 2
KIF13A	kinesin family member 13A
LDA	Latent Dirichlet allocation
MAF	Minor allele frequency
MDD	Major depressive disorder
METTL23	methyltransferase like 23
ML	Machine learning
MLP	Multi-layer perceptron
MRI	Magnetic resonance imaging
MUC16	mucin 16, cell surface associated
NELBO	Negative evidence lower bound
NN	Neural network
OCD	Obsessive compulsive disorder
PCLO	piccolo presynaptic cytomatrix protein
PGC	Psychiatric Genomic Consortium

PRPF31	pre-mRNA processing factor 31
PRS	Polygenic risk score
PS	Population stratification
QRICH2	glutamine rich 2
R-GCN	Relational graph convolutional network
RNA	Ribonucleic acid
SARM1	sterile alpha and TIR motif containing 1
SCN4A	sodium channel, voltage-gated, type IV, alpha subunit
SCZ	Schizophrenia
SEC24D	SEC24 family member D
SFARI	Simons Foundation Autism Research Initiative
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SSC	Simons Simplex Collection
SVM	Support vector machine
TSPO2	translocator protein 2
TTN	titin
WASL	Wiskott-Aldrich syndrome-like
WES	Whole exome sequencing
WGS	Whole genome sequencing
XGB	Extreme gradient boosting
XGBoost	Extreme gradient boosting

List of Figures

List of Tables

Acknowledgements

I owe an immense amount of gratitude to all the people who have supported me during my Master's. First and foremost, thank you to my supervisor, Dr. Yannis Trakadis. Your guidance and support have helped me become a better researcher, and I would not be here without your full trust in my abilities and the freedom you allowed me to explore different hypotheses.

To my supervisory committee members, Dr. Reihaneh Rabbany and Dr. Jean-Baptiste Rivière, thank you for your invaluable feedback and input for improving and advancing my research. More specifically, I'd like to acknowledge Dr. Rabbany for providing valuable advice in the machine learning and network science aspects of the project and Dr. Rivière for providing genetic and genomic resources assistance. I'd like to thank all my coursework professors for expanding my knowledge and offering helpful resources. I'd also like to extend my special thanks to Ross MacKay and Antonois Daskalakis in the HGEN department for providing administrative support.

To all my friends in Montreal, Ottawa and Edmonton, you know who you are, thank you for enriching my life with your presence. Last but not least, to my family, especially my mom and dad, you have been a constant source of strength and encouragement throughout my life, and I would not be where I am today if not for all the opportunities you provided me growing up.

Contributions

All the work presented in this traditional thesis is the original contribution of the candidate conceptualized together and under the supervision of Dr. Yannis Trakadis. The candidate performed all literature reviews, experiments, and discussion summarized in this thesis. All the material for this research such as the whole-exome sequencing datasets were obtained through Dr. Trakadis's research approval from organizations such as the National Database of Autism Research and the database of Genotypes and Phenotypes. All the computation and digital storage resources required for this research were obtained by Dr. Trakadis for the candidate from Compute Canada and Calcul Québec.

Chapter 1: Introduction

1.1 Neuropsychiatric Genetics

Neuropsychiatric disorders, a set of heterogenous and overlapping disorders which mainly affect behavior, mood, perception and cognition, are one of the largest causes of disability worldwide[1–3]. It has been known for decades before the era of genomic technologies that neuropsychiatric disorders aggregate in families and that they have a major genetic component to their etiology[4]. Psychiatric disorders are clinically diagnosed using the Diagnostic and Statistical Manual of Mental Disorders (DSM)[5] criteria, and they are considered to be polygenic and to follow a complex inheritance. A network of genetic variations across the genome, along with environmental factors, increases one's risk to psychiatric disease. The genetic architecture of most psychiatric disorders consists of many inherited common variants with small effects, as well as, rare, de novo mutations with large effects on risk[6,7]. However, their exact genetic pathophysiology still remains elusive.

This project focuses on three neuropsychiatric conditions, namely, autism spectrum disorder (ASD), schizophrenia (SCZ), and Bipolar Disorder (BD).

Autism Spectrum Disorder

Autism spectrum disorder (ASD) is a set of heterogenous neuropsychiatric disorders, affecting nearly ~1% of the population, with heritability estimated to be around 70-90%[8,9]. Children with ASD show deficits in communication and social behaviour and exhibit repetitive behaviours[9]. Affected children show highly variable clinical features ranging from severe impairment and intellectual disability to above average academic abilities. The prevalence of ASD varies by sex, as males are four times more likely to be diagnosed than females[10].

The exact etiology of ASD is still unknown. The role of de novo and inherited variants, both rare and common, has been well established[11–14]. Through linkage and exome sequencing studies, rare de novo variants in 100-1000 genes have been estimated to play a role for ASD[12,13]. However, most of ASD heritability is now attributed to common inherited variants with small effects[14].

Schizophrenia

Schizophrenia (SCZ) is a chronic neuropsychiatric condition with up to ~0.5-1% prevalence in the population and estimated heritability of ~80% from twin studies[15]. It affects perception, emotion, and cognition, and is characterized by hallucinations, disorganized thinking, incoherent speech and abnormal motor behavior, as well as negative symptoms[16]. SCZ is typically an adult-onset disorder, emerging in the early twenties for males and, slightly later, in mid to late twenties, for females[17].

The role of both common and rare disruptive variants in the genetic architecture of SCZ has been established[18–20]. The largest and most recent GWAS study for SCZ, performed by Psychiatric Genomic Consortium (PGC), reported the association of 108 independent loci which were enriched for genes expressed in the brain[20]. Of the identified loci in the study, 75% correspond to protein-coding genes and an additional 8% are located within 20 kb proximity of a gene. Similarly, studies of rare variants through whole-exome sequencing (WES) have identified a polygenic burden of rare mutations in many genes for SCZ[19].

Bipolar Disorder

Bipolar disorder is a chronic neuropsychiatric disorder with an estimated lifetime prevalence of 1-2%[21]. It's heritability has been estimated to be 70-90% from twin studies and is characterized by recurrent periods of mania and hypomania[22–24]. Each manic and depressive episode can last for several weeks and affects mood, perception, emotion, activity, and energy levels. BD is an adult-onset disorder. Affected individuals have 10-30 times higher rate of suicide and a decreased life expectancy of 9-17 years compared to the general population[25].

The most recent GWAS from PGC has identified 30 loci to be associated with BD, 8 of which had been previously reported for association with SCZ, thus supporting the genetic overlap of the two conditions[26]. The role of rare variants has not been as well-established, but recent family studies using whole-genome sequencing (WGS), whole exome sequencing (WES), and microarray data have provided some preliminary evidence to support this[27,28].

1.1.1 The role of SNP arrays and GWAS in psychiatric genetics

Earlier genetic methods, such as linkage and candidate gene studies, were largely unsuccessful in reliable identification of risk loci for common neuropsychiatric disorders. The invention of SNP arrays that could genotype hundreds of thousands of common variants (SNPs) simultaneously gave rise to genome-wide association studies (GWAS). In these studies, the entire genome is scanned for genetic markers in large numbers of cases and controls, which are then tested for association with the trait of interest using a univariate statistical test[29]. This was a major improvement over candidate gene and linkage methods, as it improved scale of coverage, statistical power, and allowed for an unbiased assessment of the genome[29,30].

GWAS revolutionized the field of psychiatric genetics and allowed for systematic discovery of loci with large sample sizes. With the growing number of samples, GWAS has been successful in identifying a large number of common risk variants. To date, a total of 241 significantly-associated loci have been identified for 10 neuropsychiatric disorders, which include 400 different protein coding genes[1].

Many of the psychiatric disorders are genetically correlated, with some overlapping variants conferring risk to multiple disorders, as evidenced from GWAS results where many of the identified loci have shown association for multiple psychiatric disorders (i.e. pleiotropy)[18,31,32]. The most recent cross-disorder genomic study by Lee et al.[32] reported 109 loci to be associated with at least two disorders, 23 of which shared association with four or more different disorders. The study analyzed 232,964 cases of autism, schizophrenia, bipolar disorder, anorexia, ADHD, major depression, OCD, and Tourette syndrome, as well as 494,162 controls from GWAS studies. It performed pairwise genetic correlation using LD-score regression and showed high correlation of 0.7 between SCZ and BD, 0.22 between ASD and SCZ and 0.14 between BD and ASD.

1.1.2 Polygenic risk score in psychiatric genetics

Polygenic risk score (PRS) is generally used to capture the cumulative *additive* impact of GWAS identified variants for risk prediction and cross-disorder association. When PRS is used for explaining the variance between cases and controls of a particular disorder, Nagelkerke's R^2 (a measure of the proportion of the variance explained) is reported in GWAS studies. So far, in SCZ, R^2 of 0.035 (3.5%), in BD mean R^2 of 0.08 (8%) and in ASD R^2 of 0.0245 (2.45%)

at GWAS significance threshold have been reported[20,26,33]. The lack of success of PRS in clinical settings is not surprising, given it is only focused on one type of common variant (Single Nucleotide Polymorphisms, SNPs) and does not take epistatic interactions of variants into account.

1.1.3 Whole exome sequencing (WES)

Recent advances in sequencing technologies have allowed for the sequencing of the whole exome (i.e. the protein coding part of the DNA), which represents 1-2% of the entire genome. WES is more cost effective than WGS and has facilitated the elucidation of different Mendelian disorders[34]. WES can capture rare variants which cannot be detected by the genotyping arrays commonly used in GWAS. In addition, unlike GWAS, point mutations identified by WES point to specific genes, and can thus be interpreted within the context of their gene.

WES has allowed for identifying variants with high functional impact. It has also uncovered the role of rare and de novo mutations through case control and family studies for neuropsychiatric disorders[35]. For example, it has shown associations of gene-disrupting de novo mutations in ASD for genes expressed in brain tissues[13,28,36,37]. Similarly, for SCZ, the role of excessive gene-disrupting mutations in the postsynaptic genes and calcium ion channel signalling has been reported using WES data[19,38].

Given the amount of data generated by WES, machine learning (ML) analysis of WES data may advance our understanding of psychiatric genetics.

1.2 Machine learning (ML)

The field of machine learning focuses on developing computer algorithms that learn with experience, automatically[39]. ML offers methods for pattern detection and knowledge discovery in high throughput data, rendering it one of the most promising approaches for understanding the human genome[40]. ML algorithms can uncover complex relationships in large dimensional data which would otherwise be impossible for humans to detect.

Owing to the recent improvements in computational hardware such as graphical processing units (GPUs) and advances in deep learning, machine learning has revolutionized many fields such as image and speech recognition, natural language processing, and patient diagnosis based on electronic health records[41–44]. It has also offered methods for analyzing and processing of single cell transcriptomic data, but also for the deconvolution of bulk RNA expression to estimate cell type proportions[45–47]. Another remarkable contribution of ML in health sciences was the major breakthrough from Google DeepMind at the end of 2020 in solving the protein folding problem, accurately predicting the 3D structure of proteins based on their one-dimensional sequence amino acids[48].

Broadly speaking, there are two types of ML algorithms: *supervised* and *unsupervised* ML. In *supervised methods*, the algorithm learns from many examples of input to output labelled data in order to best explain the observed variance in the output through some function or composition of functions of the input features/variables. The trained algorithm can then be leveraged to make predictions on future unknown or unlabeled instances. For example, learning to predict disease status based on labeled or classified samples of cases and controls using a set of feature values, such as genetic variants or clinical symptoms, falls under

supervised machine learning. The target or outcome variable (e.g. disease status) could be a discrete number of classes (e.g. case vs. control) or a continuous variable (e.g. height). In contrast, *unsupervised ML algorithms* do not need labeled data, but rather, look for inherent patterns within the data. An example would be clustering patients into homogenous groups based on their biomarkers, using a similarity or distance function[49,50]. Another example would be topic modelling, which is widely used in natural language processing to find hidden semantic structure (topics) within documents, where each document gets defined by a distribution over the learned latent topics[51].

1.3 Hypothesis and Objectives

Our hypothesis is that advanced machine learning methods can be used for classification and stratification of psychiatric disorders using WES data. Computational methods that go beyond the linear sum of genetic risk factors and take epistatic interaction of many variants into account are needed to properly characterize the polygenic architecture of these disorders and enable subtype identification and risk prediction. Our overall aim was to leverage various machine learning methods on WES data of ASD, SCZ, BD and control subjects to uncover insight into the genetic architecture and etiology of (these) complex neuropsychiatric disorders.

First, given ASD and SCZ are highly heritable disorders, with overlapping genetic risk factors, our first objective was to implement a supervised ML approach to classify patients with ASD and SCZ based on their rare genetic variants and identify important genetic features that distinguish them from each other.

Our second objective was to compare the exomes of individuals with BD, SCZ, and controls, using a different ML approach, called topic modelling. In this approach, we model the exome as a book and try to learn a small number of interpretable latent variables called topics (or dimensions) that could characterize the differences in the conditions targeted. We also explore if these topics could be used to identify homogenous clusters of subjects having the same clinical DSM diagnosis.

Our last objective was to develop a novel method that takes a family-based approach to predict affected and unaffected status in a given psychiatric disease. More specifically, we used exome data from families with ASD and compared affected to unaffected siblings, while integrating the parental genomic information. We modeled family relationships, variant inheritance (i.e. de novo vs. inherited), as well as variant features (e.g. predicted functional type and minor allele frequency) in a large graph and approached the problem as a semi-supervised node classification task for affected status.

Chapter 2: Materials and Methods

2.1 Whole-exome sequencing datasets

2.1.1 Autism WES dataset

*Autism WES data (**NDAR trios and quads**)*: The whole-exome sequencing data for 2,392 families with an affected ASD child were obtained from the National Database of Autism Research (doi: https://doi.org/10.15154/1169318; doi: https://doi.org/10.15154/1169195). Genomic DNA was extracted from whole blood samples. This dataset includes 1800 quads (unaffected parents with one affected and one unaffected child), and 592 trios (unaffected parents and affected child) with the original sequencing data gathered by the Simons Foundation Autism Research Initiative (SFARI) under Simons Simplex Collection (SSC)[52]. Affected children younger than 4 or older than 18 years of age were excluded from the study. In addition, children with some conditions such as severe neurological deficits or genetic evidence of fragile X or Down Syndrome were excluded from the study[52].

2.1.2 Schizophrenia and Bipolar Disorder WES Datasets

*Schizophrenia WES data (**dbGaP trios**)*: This WES data for 623 Bulgarian trios was obtained from the database of Genotypes and Phenotypes (dbGaP) where it is available under phs000687.v1.p1 study ID. Genomic DNA was extracted from whole blood samples. Unrelated families (of parent-offspring trios, with parents not having positive history of schizophrenia) participated in the original study. For probands to be included in the study, they had to have positive history of hospitalization for schizophrenia and to have graduated from school, to exclude probands with intellectual disability from the cohort[53].

Schizophrenia, Bipolar disorders WES data (dbGaP): This dataset accessed through the dbGaP (study phs000473.v2.p2) contains WES data for 12,380 individuals from a Swedish population with 4969 SCZ cases, 1,166 Bipolar disorder cases, and 6,245 controls (CTL). In order for SCZ case individuals to be included as part of the study, they had to be between 18-65 years of age, alive, hospitalized two or more times with SCZ, born in Sweden or other Nordic country and both parents born in Sweden as well[54]. For controls, similar criteria except for never being hospitalized for SCZ in the past were used.

2.1.3 WES data filtering criteria

From the variant call format (VCF) files of ASD (NDAR) and SCZ (dbGaP trios) which were annotated using ANNOVAR according to the reference genome GRCh37 (hg19), we filtered for rare variants (MAF < 0.01) predicted to be functionally important, with genotype quality of 90. To explore the impact of all good quality variants, for the Swedish SCZ/BD/CTL WES cohort, we selected for all variants with mean genotype quality of greater than 60, and the data was annotated using the reference genome GRCh38 (hg 38).

2.1.4 WES data representation

For ML analyses, we represented each genetic variation to take values in one of {0, 1, 2} corresponding to whether the variant is wildtype, heterozygous, or homozygous. Wildtype refers to the non-mutant or standard form of a gene where both alleles correspond to the reference (i.e. 0/0). Heterozygote implies two different forms of an allele have occurred in the same position, where one corresponds to the reference and the other does not (i.e. 0/1). Lastly, in homozygous alternate, both alleles in a particular position do not correspond to the

reference and thus are mutant (i.e. 1/1). The output variable in our ML analyses is the phenotype: affected vs. unaffected, ASD vs. SCZ vs. BD vs. CTL depending on the dataset (i.e. binary or multi-class), also treated as a continuous variable when adjusting for population structure.

2.2 Resources for integration of pathway information and tissue specific gene expression

2.2.1 MSigDB canonical pathways

This annotated dataset of genes to pathways was downloaded from MSigDB (Molecular Signatures Database) which compiles this information from various databases and biomedical literature[55–57]. We converted this information into a numeric matrix of pathway by gene for machine learning analysis. The values of the matrix are binary indicating whether a gene is involved in a particular pathway.

2.2.2 GTEx RNA-Seq data

Tissue-specific gene expression data was downloaded from The Genotype-Tissue Expression (GTEx) project online portal[58]. We downloaded the V8 Gene TPM and filtered for brain tissues only.

2.3 ML analysis of WES data to contrast the genomic architecture of ASD & SCZ

2.3.1 Population stratification correction:

Before training a machine learning algorithm to differentiate between ASD and SCZ samples, we corrected for population stratification. Population stratification can bias an

association or classification task due to the systematic allele frequency differences that exists in different subpopulations driven by non-random mating[59,60]. If not corrected, a ML algorithm could capitalize on such population structure and not learn the underlying genetic differences that are important in differentiating the two disorders. Our focus on rare variants reduced the impact of such a confounder. However, we still performed a principal component analysis (PCA) based stratification correction method (called Eigenstrat) proposed by Price et al.[61] to properly address this.

We applied Eigenstrat on our curated individual x variant matrix of ASD and SCZ samples to infer the top axes of variation which captures population structure differences. Afterwards, we fit a generalized linear model by regressing each variant on the four axes of variation and assign the residuals of this regression to be the corrected values of the particular variant. This essentially removes the impact of population structure from each variant. We do the same for the phenotype, as for a variable to be considered a confounder, it needs to be associated with both the target and input variables[60]. As a result, the genetic variants which used to have integer values of 0,1,2 becomes continuous after this correction. Similarly, the phenotype values, from binary (indicating ASD vs. SCZ), became continuous. After the correction, the adjusted phenotype values all but one fell within the range of -4 to +4, so we capped the phenotype values to this range.

2.3.2 Regularized Gradient Boosted Machines (GBM)

GBM is an ensemble method of weak learners, where each base learner is added to the ensemble iteratively to correct for the errors of previous ones[62,63]. Each base learner in the ensemble can be one of any possible classification and regression trees (CARTs) in the space of

all possible CARTs. GBMs can be trained to optimize any differentiable loss function. Therefore, they can be trained for either classification or regression tasks. We chose GBM, and specifically its extreme gradient boosting implementation (XGBoost)[64], for the problem of ASD vs. SCZ due to its state-of-the-art performance on cartesian data, better interpretability, and parallelized and regularized implementation. Given that our data is large dimensional, we prefer a heavily regularized algorithm to penalize complexity and to avoid overfitting. The regularization and embedded feature selection methodology of XGBoost reduces the number of input variables in the model and ranks them based on their relative importance for predictive power.

*2.3.3 ASD vs. SCZ classification at the variant level (**Variant-Level** Approach):*

In this approach we took the actual genetic variants and their values (i.e. 0,1,2 based on their genotype) to be our features set. We initially trained and optimized XGBoost without any population stratification (PS) correction but then repeated the analysis after correcting for PS using Eigenstrat. After the correction and rounding up/down, the phenotype became continuous valued from initial binary which denoted SCZ and ASD classes. We thus approached the analysis: (1) By keeping the phenotype adjustment as a continuous variable based on Price et al.[61] methodology. (2) By converting the adjusted phenotype to binary. The latter was done by looking at the distribution of the adjusted phenotype values. All SCZ samples had adjusted phenotype values in the interval [-4,-1] and ASD samples in interval [1,4]. Therefore, we assigned the two classes 0 or 1 values based on the two clusters clearly segregating with regards to the two phenotypes. A 70:30 data split was used for training and testing purposes.

2.3.4 ASD vs. SCZ classification at the gene level (Gene-Level Approach):

In this approach, we summed up the corrected variant-level genotype values to their genes, then trained an optimized XGBoost using genes as opposed to variants as the input features. Similar to the variant-level approach above, we took two methodologies for training: regression (when directly using the adjusted phenotype) and classification (after converting the adjusted phenotype to binary). Similarly, we used the 70:30 ratio for splitting our data for training and testing.

2.4 ML analysis of WES data to contrast the genomic architecture of SCZ, BD and controls.

2.4.1 Supervised SCZ vs. BD vs. CTL classification approach

This is an extension of the SCZ vs. ASD approach, but focusing on SCZ vs. BD, and including controls, all from the same population. We first performed a supervised ML analysis using *all* high-quality variants in the exomes of these individuals using a regularized linear model (LASSO) and XGBoost. We kept 25% of the original data for testing, and then used 25% of the remaining 75% for validation. This left ~56% of the data for training purposes. We made sure each set had the same distribution of classes. Given the dataset is unbalanced, especially for BD where its ratio is close to 1:11 in the overall data as shown in supplementary Figure 1, we weighted the loss function based on the inverse frequency of each class. This was used to ensure that the majority class does not dominate the training process in supervised ML. We trained our model on the training set by minimizing the cross-entropy loss, while simultaneously controlling the loss on validation to prevent overfitting. During the training

process, the best model is saved based on its evaluation on the validation set, and then used on the test set for computing final evaluation metrics.

2.4.2 Embedded Topic Modelling Approach:

Topic models have been widely used in natural language processing to discover hidden semantic structure within documents. They model each document as a mixture of 'topics' which need to be inferred from the data, generally in an unsupervised fashion. Inspired by this approach, we modeled the whole exome of each person as a document, the genes as words, and the frequency of mutations in a gene as the frequency of words in a document. Our goal is to infer a small number of latent variables (called topics), which could model the underlying semantic structure of the exomes. Biologically speaking, the latent variables can be thought of as endophenotypes (e.g. a set of symptoms or biological processes/pathways) that can be explained by the underlying structure of genetic variance in the whole exome. We used these interpretable topics to cluster genomic data, as a way to explain phenotypic variation in the corresponding subjects.

The model we used is called Embedded Topic Model (ETM)[65]. ETM is an extension over the common topic modelling approach called latent Dirichlet allocation (LDA) which models each document as a mixture of topics and each topic a distribution of words. ETM simultaneously performs topic modelling and word embedding. The word embedding extension can be thought of as learning a low dimensional representation (i.e. meaning) of words (genes in our case), where similar words end up with similar representation. We trained an ETM model using the negative evidence lower bound (NELBO) loss function.

The extracted latent variables were used for downstream clustering, but also for classification tasks in an XGBoost model with performance being evaluated on held out test data. We performed our ETM analysis in two levels: one at the variant genotype level (where each *variant* can be thought of as a word) and another at the gene level where we aggregate the number of variants within a gene (here, each *gene* is the equivalent of a word in the ETM analysis). In addition, for the gene approach, we further modified the algorithm by including pathway x gene binary matrix as the fixed gene embedding in order to guide the algorithm to learn biologically meaningful topics based on the gene embedding approach mentioned above. We defined gene embedding (or meaning) based on the molecular signature of the different genes, which was extracted from MSigDB and converted into a binary matrix. By fixing the gene embedding matrix and not allowing the ETM to learn it, we are explicitly defining that genes which are more frequently involved in same pathways are semantically ('functionally') more similar.

While implementing ETM models, we experimented with different numbers of latent variables/topics (from 20 -100), as well as various neural network architectures for the encoder in each approach. We used a linear decoder in order to keep the model interpretable. We trained each ETM model in unsupervised fashion for 5,000 to 10000 epochs until the model under training converges.

2.5 ML analysis of WES data to predict ASD status using affected and unaffected siblings

2.5.1 ASD Sib-pair Approach

In the ASD sib-pair approach, we took matched pairs of affected and unaffected siblings and removed SNVs with mutations in less than 5 individuals (sparse features) to significantly reduce data dimension, since such variants are not informative enough in a supervised ML model. This reduced the number of features/variants significantly to ~56,000 (i.e. 3,600 x 56,000 matrix). Then an XGBoost model was trained and optimized through cross validation on 70% of the balanced dataset while using accuracy as the performance measure.

2.5.2 Network based approach and data representation of ASD families and their variants

Using the ASD quad family data, we created a heterogenous graph/network to model the relationships in our data. We used two types of nodes (i.e. individuals and variants) and three types of edges: one type of edge to denote child-parent relationship and two other edge types to connect individuals with their variants (homozygote edge and heterozygote edge). Unlike the traditional ML methods on Cartesian data, here the affected and unaffected child nodes are not independent anymore as they are connected to each other through their parent nodes (Figure 1).

We also enriched our network by including variant information such as minor allele frequency, functional type, and variant type. Each variant node was enriched further by including information about the extent of its gene being expressed in brain tissues. We also encoded genetically meaningful topological structures in our graph as inherited and de novo variants form different structures with human nodes (Figure 2). Therefore, our graph neural

network algorithm not only learns from the information within each node, but also from the embedded topological structure surrounding it.

Figure 1: Mini heterogenous illustrative graph of the proposed network

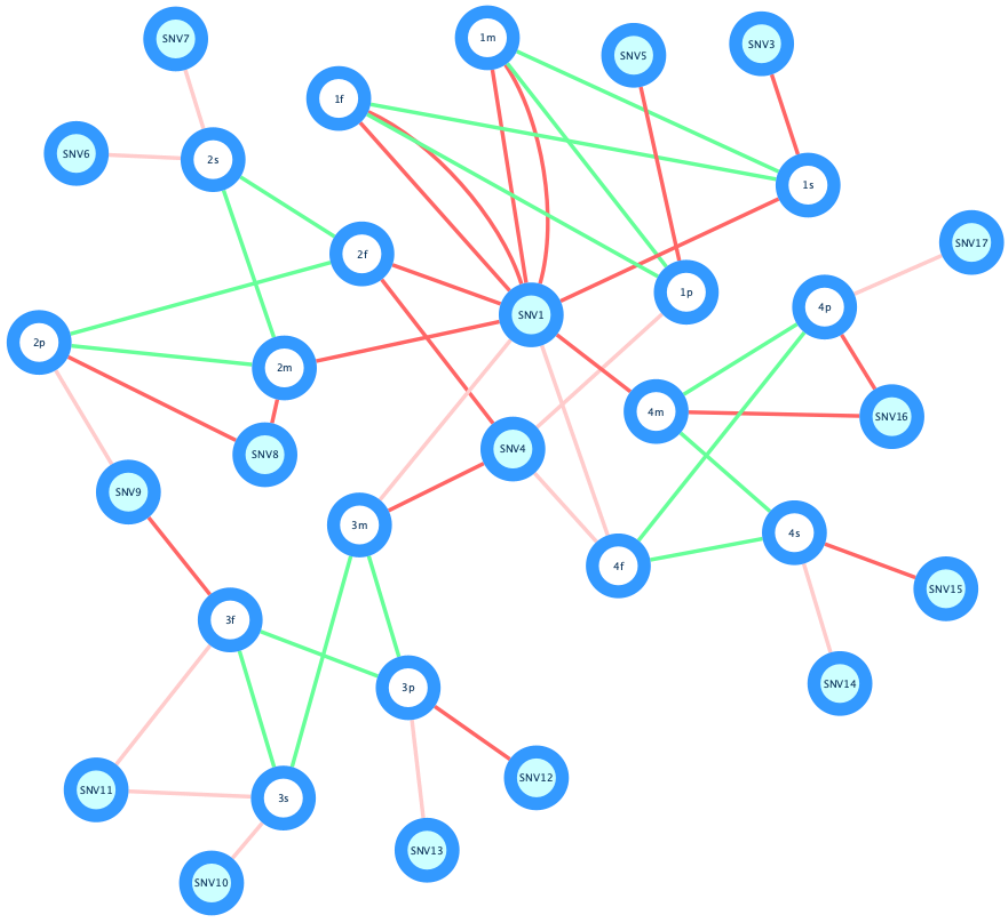


Figure 1: A mini example representation of our graph with three families and sixteen variants. Filled nodes denote variants and white filled nodes denote human nodes. Parent child relationship (edge type) is denoted in green. Human to variant is denoted in two different shades of red (depending on the zygosity of the human in for the particular variant).

Figure 2: Sample homogeneous illustrative graph

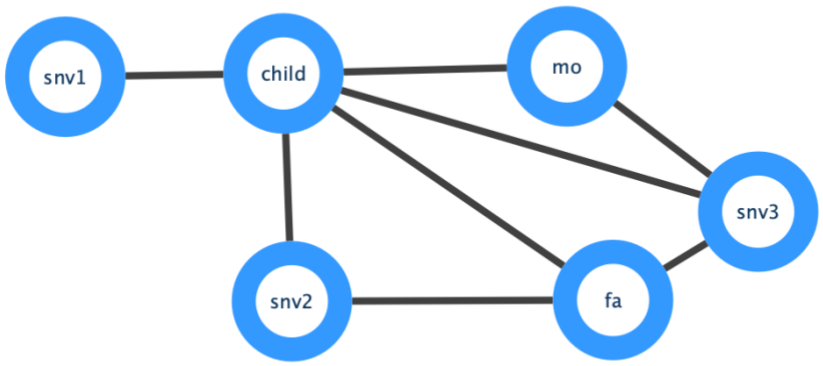


Figure 2: A small homogeneous example of our large heterogeneous graph shown to demonstrate some important topological structures in the graph that translate into important genetic concepts. It's crucial in genetics to understand the possible source (i.e., inheritance pattern) of each mutation in one's genome and how it compares between affected and unaffected children across a population. The graph here is showing the difference among de novo mutation (snv1), a paternally inherited variant (snv2) and a homozygous variant (snv3) where one copy was inherited from each parent. They can be distinguished by the triangle(s) pattern formation, or lack thereof, between the child and each parent with a given SNV.

After defining our network, we set up our problem as a node classification task where we are interested in predicting the disease status of human children nodes (i.e. binary classification) in our proposed graph. To this end, we divide our target nodes (i.e. 1800 nodes) in the graph into training, validation and testing sets of 70:15:15 ratio, where we mask the

labels of validation and testing sets during training. We performed node classification in two ways, one using all rare variants in quad families. In the second approach, we limited our analysis to only variants classified as pathogenic or likely pathogenic variants in ClinVar[66] for the quad families in the network.

2.5.3 Proposed node-classification algorithm for the heterogenous graph/network

Traditional machine learning and deep learning methods have been quite successful at learning from Euclidean (tabular) data but cannot generalize to graphs which are irregular objects, and important notions from deep learning such as convolution are not well-defined on them[67]. Therefore, we are not able to use classes of methods used in our earlier analyses. However, there has been a significant amount of research done lately to extend neural networks to graph structured data and generalize the convolution operation to graphs. One main concept behind these methods is the recursive propagation and aggregation of feature information from node neighborhoods using neural networks[68–70].

However, most of these methods are focused on homogeneous graphs where there is only one type of edges and nodes which is not the case in our ASD network where we have multiple edge and node types. In an effort to generalize graph convolutional neural networks (GCNs) to relational or heterogeneous graphs, Schlichtkrull et al.[70] proposed Relational GCNs (R-GCNs) which extends the GCN convolutional operator to apply relation type specific transformations to the message-passing framework of GCNs. The authors demonstrate the effectiveness of their model on both entity/node classification as well as link prediction tasks among nodes. Similarly, Zitnik et al.[71] proposed Decagon which implements an R-GCN variant

on a heterogeneous drug-drug with protein interaction networks to perform link prediction task of polypharmacy side effects.

Motivated by the R-GCN model, we implemented relation-specific transformations using separate neural networks per canonical relation type instead of a linear transformation proposed in the original model. Therefore, each node's feature content and embeddings get transformed and reduced differently based on the type and direction of the relations on which it's being propagated.

Chapter 3: Results

3.1 ASD vs. SCZ results

The classification algorithm accuracy of ASD vs. SCZ analysis without adjustment for population structure and using variants as features was 86%, with a higher sensitivity of 97% than specificity of 79%. After adjusting for population structure, the accuracy in the *regression approach* (as per section 2.3) was 85.7% *using variants as features*, with 97.8% sensitivity and 78.6% specificity. In the *classification approach*, the accuracy was slightly higher at 87.4% with 99.2% sensitivity and 80.0% specificity (Table 1).

In addition, the *gene-level regression approach* had 88.5% accuracy, with 95.9% sensitivity and 83.2% specificity, while the corrected *gene-level classification approach* performed the best with 91.5% accuracy, with 97.4% sensitivity and 87.1% specificity (Table 1). The top 10 features (variants and genes) from the best performing population-adjusted classification algorithms are shown in Table 2 where we note some overlap in the genes identified using the different approaches. After overlapping the genes utilized in the two models, 151 were shared, with 4 genes in the top 10 of both models shared as well. These 4 genes were *SARM1*, *QRICH2*, *PCLO* and *PRF3*, with the first two being the top 2 of both algorithms.

Table 1: **ASD vs. SCZ analyses results on test data**

Model	Accuracy	Sensitivity	Specificity	Precision	Recall
ASD vs. SCZ approach with no pop. strat. correction, variant-level (Classification)	86.3%	97.1%	79.5%	74.8%	97.1%
ASD vs. SCZ approach with pop. strat. correction, variant-level (Regression)	85.7%	97.7%	78.5%	73.0%	97.7%
ASD vs. SCZ approach with pop. strat. correction, variant-level (Classification)	87.4%	99.2%	80.2%	75.3%	99.2%
ASD vs. SCZ approach with pop. strat. correction, gene-level (Regression)	88.5%	95.9%	83.2%	80.3%	95.9%
ASD vs. SCZ approach with pop. strat. correction, gene-level (Classification)	91.6%	97.4%	87.0%	85.4%	97.4%

Table 1: A summary of the performance of different approaches taken to ASD vs. SCZ on the test set is shown. First, the algorithm performance using variant-level data without correcting for population stratification is shown. After adjusting for population structure, both the genotypes and phenotypes become continuous. Therefore, regression XGBoost algorithms were trained, but for the test set, the resulting continuous values were turned into two classes of -1 and +1 depending on whether the output value was > 0 or not in order to report performance measures in terms of accuracy. For classification tasks, since the continuous phenotypes clustered around -1 and 1, the output variable was turned into binary before training the classifiers.

Table 2: Top 10 features from best-performing variant-level and gene-level approaches to ASD vs. SCZ ML analysis

Variant-level approach	Gene-level approach
rs71373646 (**SARM1**)	**SARM1**
rs6501878 (**QRICH2**)	**QRICH2**
rs34535433 (AKAP1)	**PRPF31**
rs77721383 (**PCLO**)	SEC24D
rs147405274 (TSPO2)	SCN4A
rs11568605 (ABCC3)	CACNA1S
rs41267712 (KIF13A)	CDSN
rs150393409 (FAN1)	HERC2
rs201671744 (CCDC155)	MUC16
rs199870856 (**PRPF31**)	**PCLO**

Table 2: The top 10 features of population-corrected variant-level and gene-level classification approaches to separating ASD vs. SCZ using XGBoost is shown. The variants and genes are shown in order of relative importance inside each algorithm (from most to least important). For variant-level approach, the underlying features used in the algorithm are variants, but their gene is shown in parenthesis for comparison to the gene-level approach. Genes highlighted in bold indicate the genes identified in both approaches (i.e. *SARM1, QRICH2, PCLO, PRPF31*). Combining both lists results in 16 unique genes which make up the top 10 informative features of each algorithm.

3.2 Results of ML analysis of WES data to contrast the genomics of SCZ, BD and controls

3.2.1 Results of supervised classification approach for SCZ vs. BD vs. CTL

As shown in the confusion matrix in Figure 3, analyzing the raw genotype data in a supervised classification of SCZ vs. BD vs. CTL, using XGBoost, the overall accuracy was only 58%. However, very good classification results were noted for the BD samples (precision: 89%, recall: 87%), in the presence of SCZ and CTL samples. Table 3 shows that the performance of LASSO (overall accuracy: 56%), was not as good as that of XGBoost, but again the results were better when classifying BD samples than other classes (precision: 70% precision, recall: 89%).

Figure 3: Confusion matrix of supervised analysis of controls vs. SCZ vs. BD

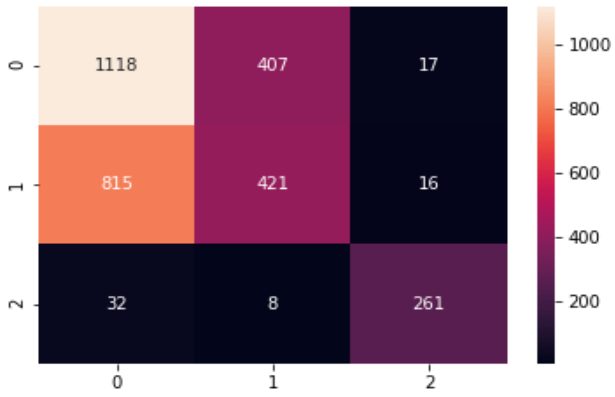


Figure 3. Confusion matrix of the best performing model (XGB) on variant-level genotype data which obtains 58% accuracy with high precision (89%) and recall (87%) for BD class. Classes are denoted as **Control: 0, SCZ: 1, BD: 2.** We can see that the BD samples are highly accurately classified. Out of the 301 samples in the test set, the algorithm classified 261 of them as BD, while 32 as control and 8 as SCZ.

We then took a look into the top features of the XGBoost model to see which features drive the classification results the most. The result for top 20 features is plotted in Figure 4 where we see a few deletions to be quite informative within XGBoost. However, this only quantifies the significance of the top variants within the overall model (by gain). To understand the relationship of each of the top features with the three classes (i.e. CTL, SCZ, and BD), we drew the hierarchical cluster heatmap of these top features against each phenotype (i.e. DSM based diagnosis). Figure 5 shows that the deletions in the top features of the algorithm are enriched mainly in BD samples.

Table 3: Classification results using genotypes and topics approach on SCZ, BD, CTL WES Data

Classifier	Input Data	Accuracy	Precision			Recall		
			CTL	SCZ	BD	CTL	SCZ	BD
XGB	Variant genotypes	**58%**	57%	50%	**89%**	73%	34%	**87%**
LASSO	Variant genotypes	56%	58%	49%	70%	58%	47%	89%
XGB	Gene topics	52.4%	55%	48%	0%	73%	37%	0%
XGB	Variant topics	50.5%	51%	38%	0%	94%	6%	0%

Table 3: A summary of model performance on the test set for SCZ vs. BD vs. CTL using different approaches. XGB denotes gradient boosting model (XGBoost implementation). The variant level XGB classifier generally outperformed models trained on topics of variant and gene level data. It outperformed LASSO linear model as well on variant level data. Both LASSO and XGB performed well in classifying BD samples using variant level genotypes for classification, with 89% precision and 87% recall for XGB, and 70% precision and 89% recall for LASSO.

Figure 4: Relative importance of variants in the SCZ vs. BD vs. CTL classification algorithm

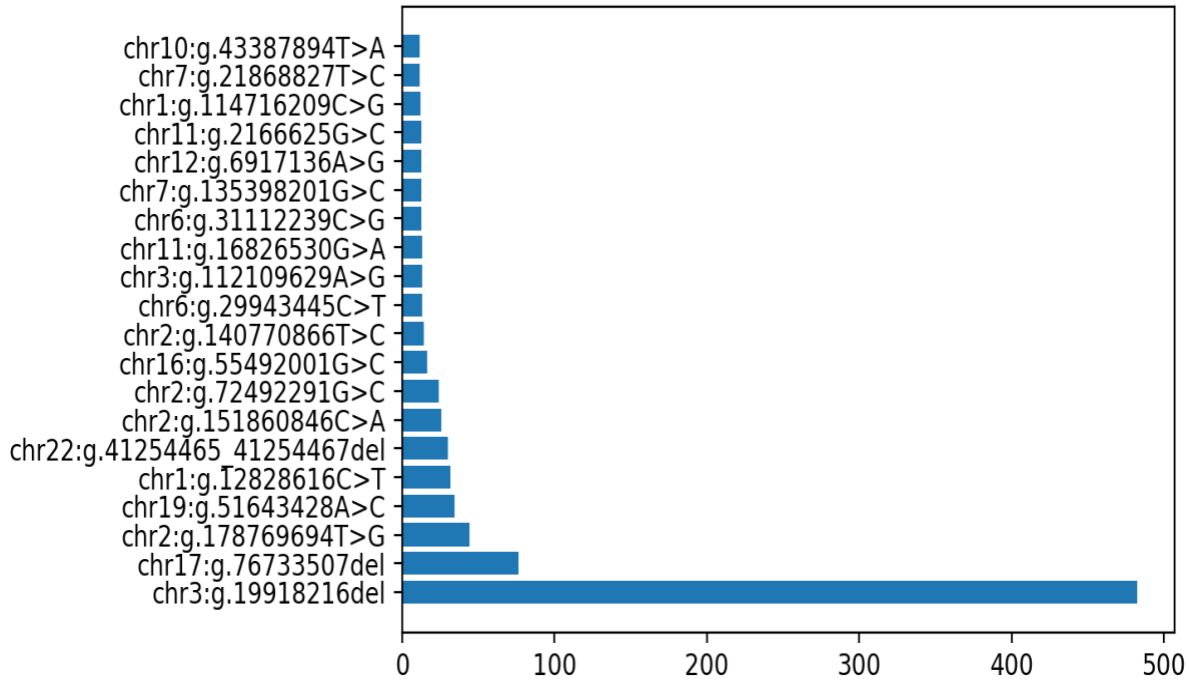


Figure 4: XGBoost model was trained and optimized to predict SCZ vs. BD vs. CTL classes based on all quality variants in the WES data. The model was able to identify BD samples with high accuracy on the test set and here we show the top 20 features of this model based on their relative importance (by gain). The three deletions and 17 single nucleotide variants shown in the figure are quite important in accurate prediction of classes within the model.

Figure 5: Heatmap of the top 20 features of SCZ vs. BD vs. CTL XGBoost model against samples

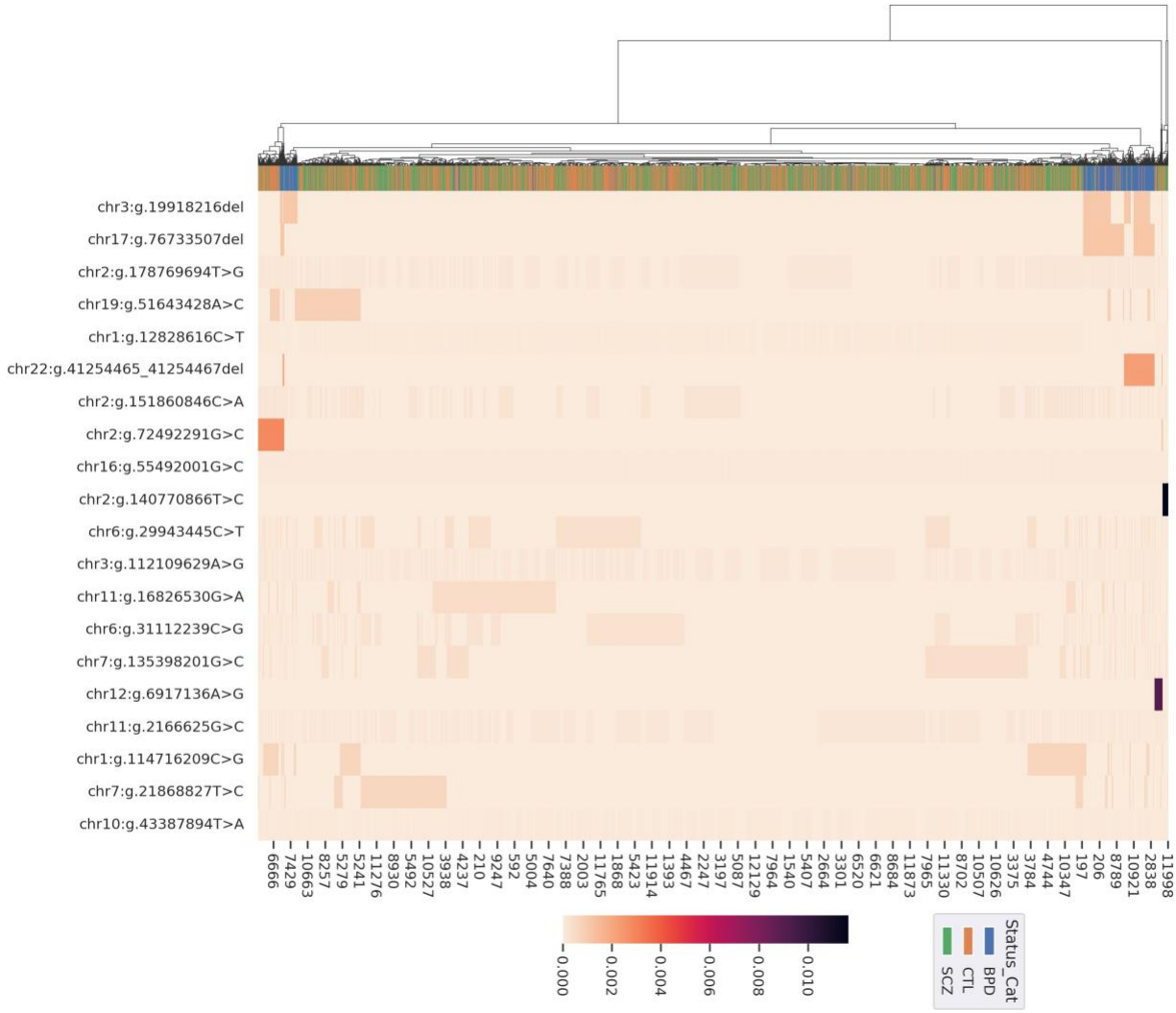


Figure 5: Top 20 features of the XGBoost model of SCZ vs. BD v. CTL are plotted with hierarchical cluster heatmap against all samples in the data to understand where these variants are enriched. As seen in the figure, the three deletions are mainly occurring in the BD samples (blue coloured samples) as indicated by non-zero and normalized genotype values mostly in BD samples.

3.2.2 Results of topic modelling approach

Using the learned topics from the *variant-level* ETM approach for clustering did not produce any interesting results, despite experimenting with different model architectures and hyperparameters. As shown in Figure 6(a), the training for topics quickly plateaued and did not decrease the loss function after 200 epochs. We used the resulting 100 topics from this model in t-SNE and visualized the labeled samples, as shown in Figure 6(b). There were no obvious clusters or subclusters of any specific phenotype.

However, for the *gene-based* ETM approach, three distinct small subclusters, forming mainly of SCZ samples, were noted after the resulting 20 topics from our best model were projected using t-SNE (Figure 7(b)). The remaining subclusters were quite mixed and did not isolate particularly well with regards to a specific clinical phenotype.

In order to understand what drives the clustering of the three SCZ subclusters identified, we delved deeper into uncovering which of the 20 topics show correlation with the clusters identified which consisted almost exclusively of SCZ patients (possibly representing SCZ subtypes). The result is shown in Figure 8, in the hierarchically clustered heatmap for samples and topics. Topic 4 is correlated with the three SCZ subclusters, as shown in the t-SNE in figure 7. To understand which genes have the highest probability under topic 4, we plot the top 5 genes under each of the 20 topics (Figure 9). Looking under topic 4, the most important genes are *MUC16*, *TTN* and *WASL* in order of highest probability.

After fixing the gene embedding with the molecular signature of the genes, the ETM did not yield as well-separated and homogenous subclusters of SCZ as the earlier gene-level topic

approach. However, it still showed formations of more homogenous SCZ clusters than the variant-level topic approach (Figure 10(b)).

Figure 6: Variant-level topic modelling and clustering of individuals

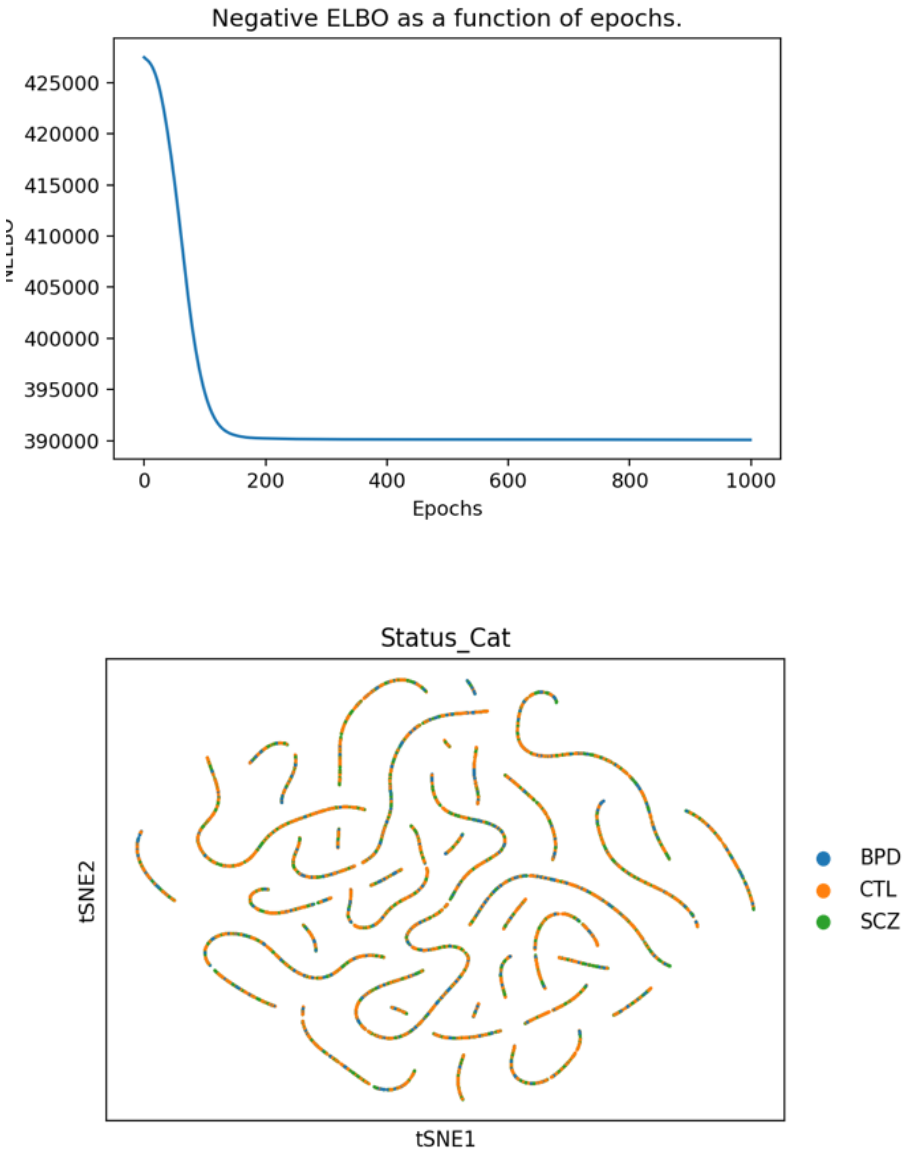


Figure 6*:* *Variant* level ETM model training loss (NELBO) as a function of epochs (top), and the t-SNE plot of the learned topics on samples from the final model (bottom). No clear clustering pattern is forming when the resulting topics is used to look for clusters.

Figure 7: Training and clustering result of gene-level topic modelling (with no fixed gene embedding)

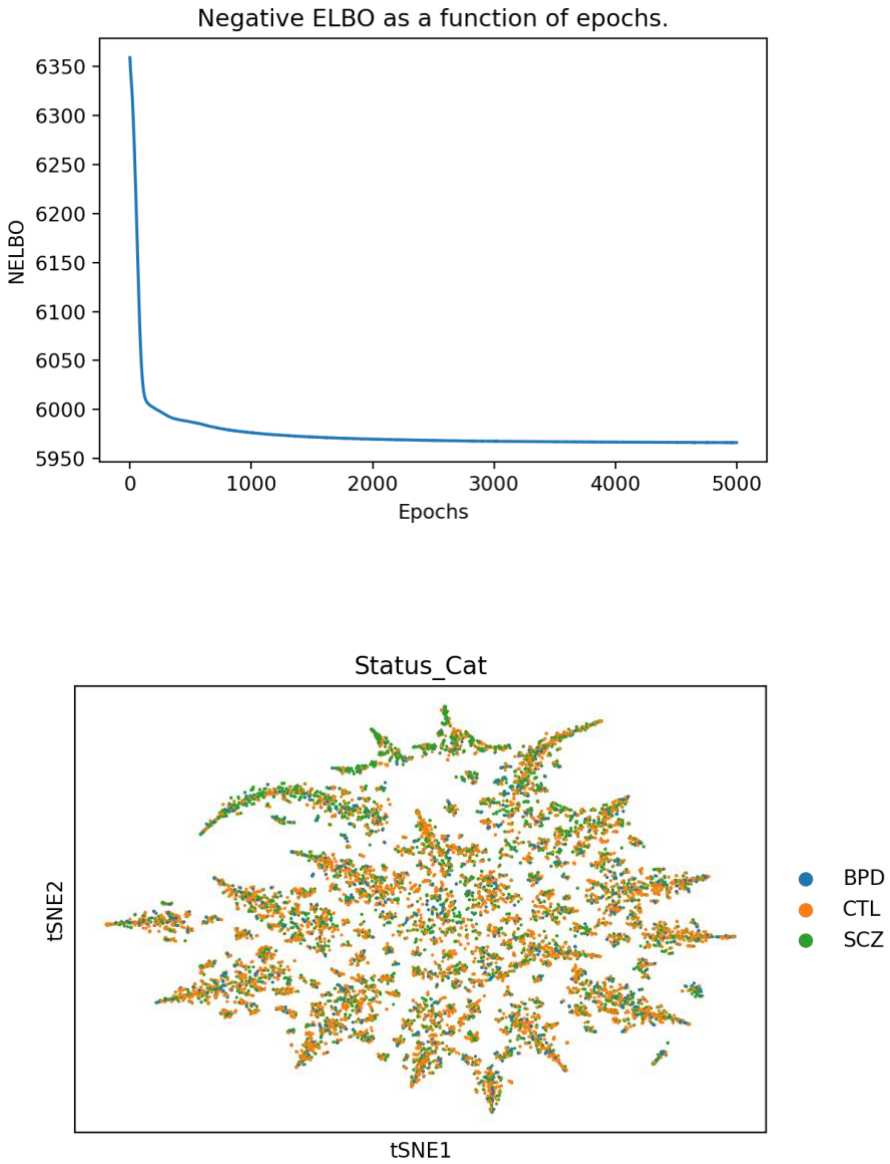


Figure 7: *Gene* level ETM model training loss (NELBO) as a function of epochs (top), and the t-SNE plot of the learned topics from the final model (bottom) with the dots denoting individuals coloured based on their phenotype. The three homogenous SCZ subclusters identified through clustering gene-level topics can be observed in the bottom figure.

Figure 8: Hierarchically clustered heatmap of topics (gene-level approach) and samples

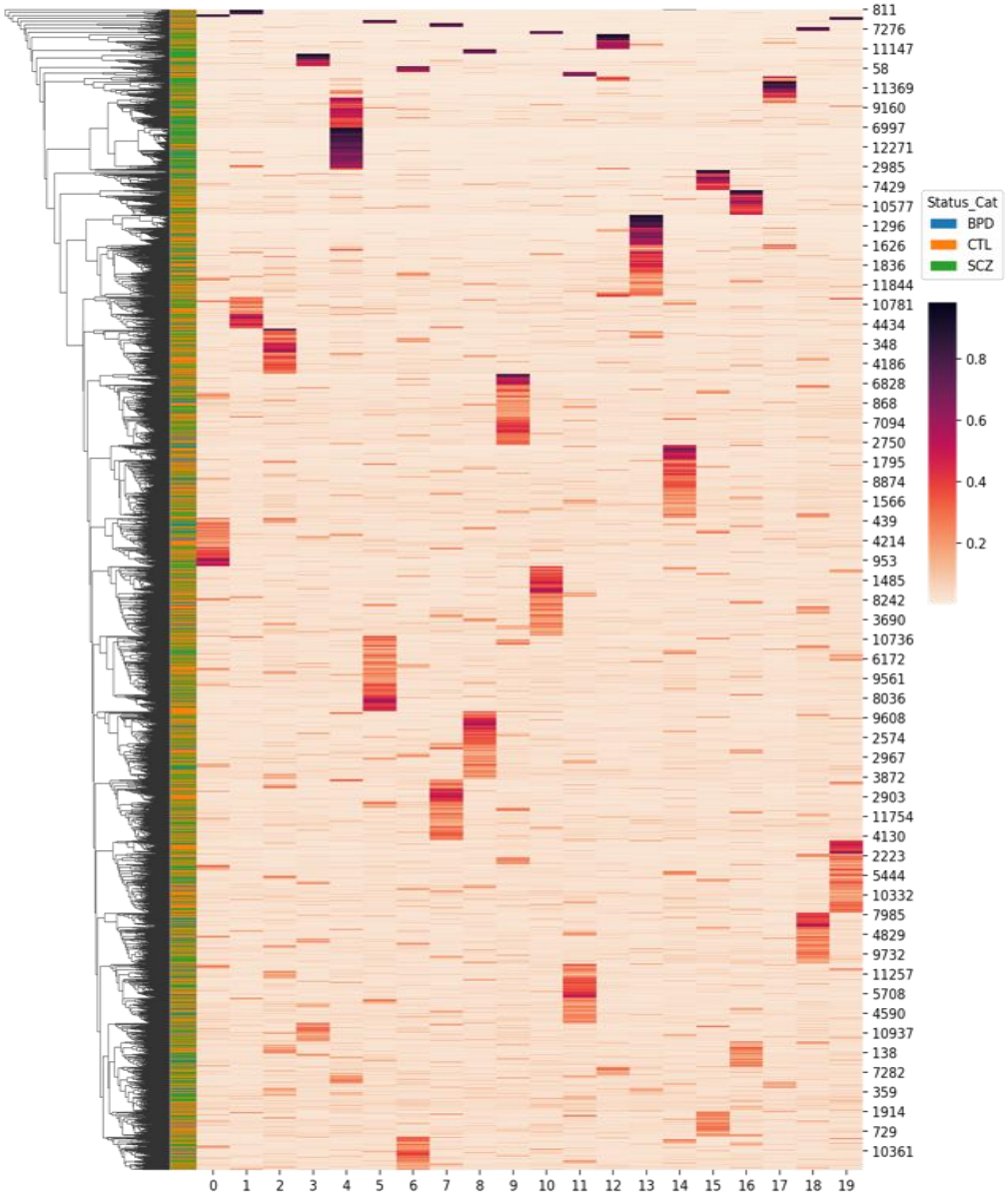


Figure 8: This figure shows the hierarchically clustered heatmap of subjects by topics. As each individual gets defined by a mixture of the 20 inferred topics summing to 1, we are seeing that a large subcluster of SCZ is almost exclusively defined by one topic (topic 4). Individuals in this subcluster are the same as in Figure 7 where they formed three smaller subclusters.

Figure 9: Top 5 genes per each topic under the ETM model for gene-level topic modelling

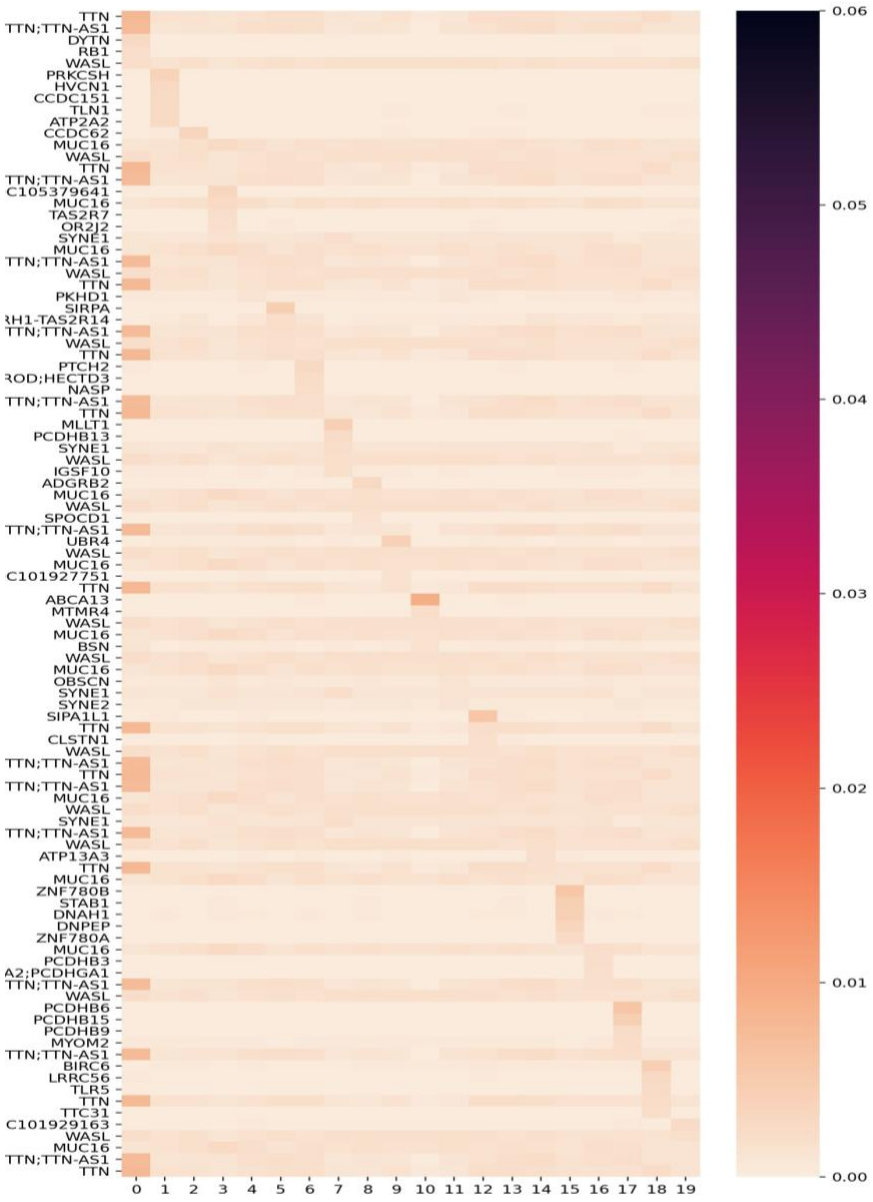


Figure 9: This figure shows only the top 5 genes for each of the 20 topics inferred with the gene level topic modelling. For example, looking for the top genes of topic 4, we see *MUC16*, *TTN*, *WASL and PKHD1* genes to be in its top 5 most important genes. The values of the heatmap are the probability of each gene under the topic. Each topic gets defined by a probability over a number of genes.

Figure 10: Gene-level topic modelling (with fixed gene embedding) and clustering of individuals

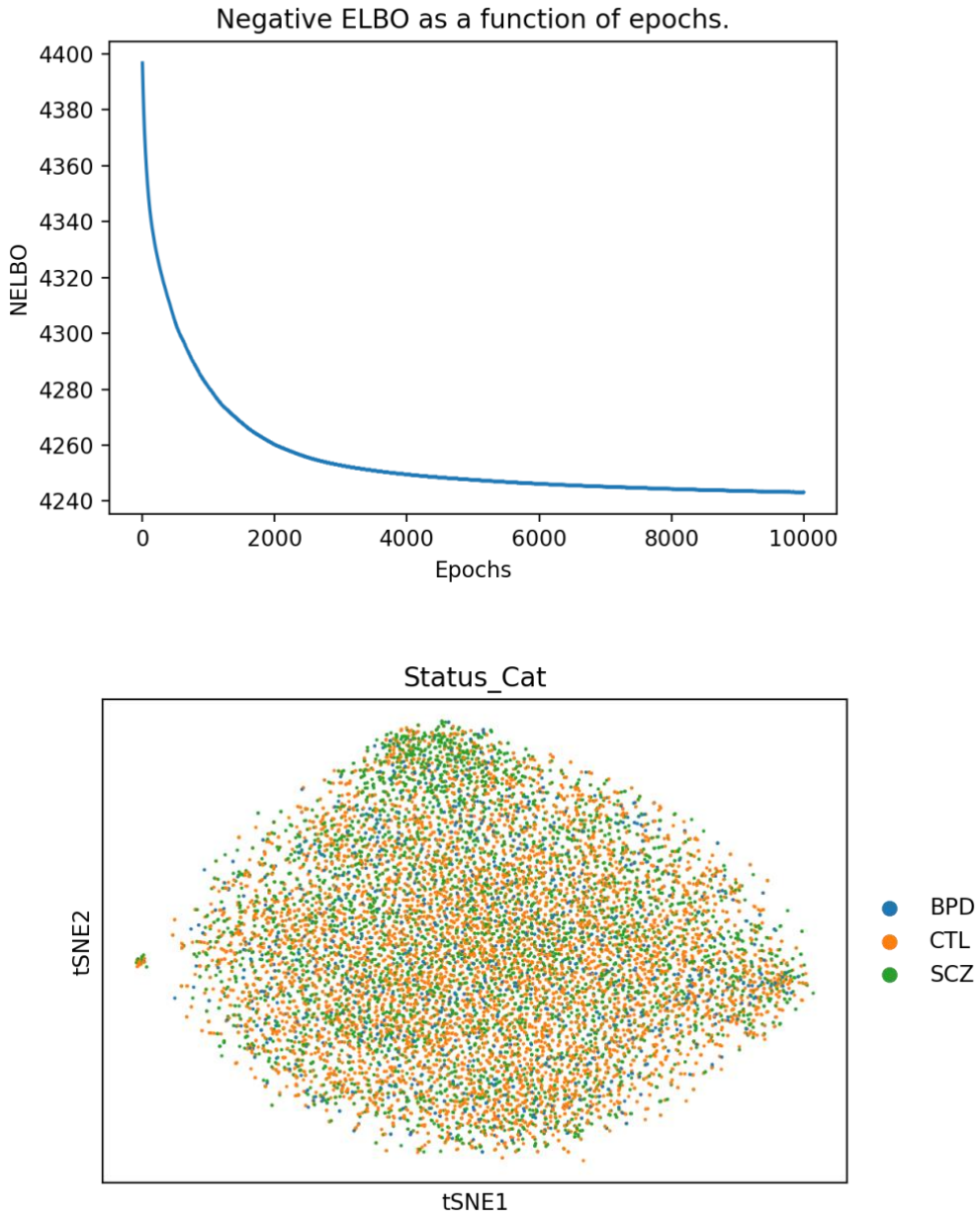


Figure 10: *Gene* level (with fixed gene embedding) ETM model training loss (NELBO) as a function of epochs (top), and the t-SNE plot of the learned topics from the final model (bottom). SCZ clusters (green dots) appearing to form a homogenous cluster at the top.

Lastly, the performance of using the learned topics of variant-level and gene-level (without fixed gene embedding matrix) as input features in a classifier are shown on the test set in Table 3. The overall accuracy of the supervised models on both gene-level and variant level topics is poor. Using the gene-level topics in XGBoost had better predictive power with overall accuracy of 52.4% with the majority class being ~50% of the test data. It's not predicting any of the BD samples correctly. It has better precision (55%) and recall (73%) for the control (majority class) compared to SCZ (48% precision and 37% recall) and BD (0% precision and 0% recall), even though the model was trained to weigh the classes based on their inverse frequency in the training set. The performance of variant-level topics in XGBoost was as good as random with 50.5% overall accuracy, basically defaulting to predicting the majority class in almost all cases.

3.3 Results of ML analysis of WES to predict ASD status in affected and unaffected siblings

3.3.1 Results of ASD Sib-pair Approach

In the ASD sib-pair approach using XGBoost on cartesian data representation, the performance of the model was not significant, as the balanced accuracy was 52.87%, only slightly better than random prediction (50%) (Table 4).

3.3.2 Results of Network Approach

For node classification of disease status within the large network, our relational graph convolutional neural network (R-GCN) model using all variants and families within the network did not outperform random classifier (average accuracy: 50%). Repeating the analysis using only pathogenic and likely pathogenic variants in quad families had similarly low accuracy (52%) equivalent to random guessing, as summarized in Table 4. The original network consisted of

1800 quads. Only 567 families had pathogenic or likely pathogenic variants, for a total of 1100 such variants in the new network.

Table 4: Predicting ASD status in affected and unaffected siblings

Model for ASD Sibpair analysis	Accuracy	Sensitivity	Specificity	Precision	Recall
XGBoost	53%	70%	51%	10%	70%
Graph App. All variants	50%	50%	0	100%	50%
Graph App. Path./L.Path. variants	52%	92%	11%	51%	92%

Table 4. A summary of the performance of different approaches to ASD sib-pair problem is shown here. In the ASD sib-pair problem, we trained an algorithm to differentiate between affected vs. unaffected siblings using XGBoost on genotypes of raw variants, and network approach of node classification where enriched the network with further family and variant features. The latter was performed in two ways, one where all rare variants were used, and another where only pathogenic and likely-pathogenic variants were kept in the network. All three approaches as shown had very low performance on the test set close to ~50% accuracy.

Chapter 4: Discussion

4.1 ASD vs. SCZ ML analysis

In the ASD vs. SCZ supervised ML approach, we were able to train an algorithm using WES data to successfully differentiate the two classes. Even though this is not clinically useful, since ASD and SCZ are clinically very easy to separate based on differential symptoms, it helped us contrast the genomic architecture of these two genetically overlapping diseases. The performance was similar before and after adjustment for population structure, as expected given this analysis focused on rare variants which are not affected as much by population structure as the more common variants.

We focused on the top 10 genes from each approach (16 unique genes) and performed literature review which showed evidence of these genes being previously linked to one or both of ASD and SCZ, as described in detail in our published manuscript[72]. For example, *KIF13A,* a member of the kinesin superfamily proteins which are important for cellular transport and signal transduction has evidence in the literature for a link to SCZ[73–76]. Similarly, Fanconi-associated nuclease 1 (*FAN1*), a DNA repair enzyme, is located in the chromosome 15q13.3 locus, and is associated with increased risk of both ASD and SCZ[77,78].

4.2 ML analysis of WES data to contrast the genomics of SCZ, BD & CTL

We used a similar supervised ML approach as SCZ vs. ASD, and developed a predictive model for SCZ and BD, two closely related complex disorders, in the presence of control samples (multiclass classification). In the supervised variant-level approach, our best performing model had an overall accuracy of 58% on the test set. The boost in performance mainly came

from the model accurately classifying the BD samples, which were identified with 89% precision and 87% recall. Looking under the hood of our classifier model, we extracted the top 20 most predictive features (by gain) and analyzed their correlation with the three clinical phenotypes/DSM-based diagnoses. Among the top 20 variants, there were 17 point-mutations and 3 deletions. The deletions were mainly enriched in BD samples. The most important deletion chr3:19918216del is located in chromosome 3, and particularly within the *EFHB gene.* This deletion has not been previously reported to play a role in BD. However, *EFHB* is involved in calcium ion transportation, the role of which has been well established in psychiatric disorders. The second most important feature was Chr17:76733507del which has also not been previously reported to play a role in BD but it's located within *METTL23 gene*, the product of which acts as a regulator in the transcriptional pathway for cognition[79]. Mutations in this gene have also been reported for association in intellectual disability[80].

We then applied a novel topic modelling method (ETM) to shed further light into the complex genetic architectures of SCZ and BD. More specifically, our goal was to infer a number of latent variables (called topics), which could model the underlying semantic structure of the exomes, and to use these interpretable topics to cluster and explain phenotypic variation in the subjects. We hypothesized that each patient's clinical status can be modelled by a latent distribution of underlying 'topics' (which can be thought of as endophenotypes) caused by complex genetics. And that a mixture of these 'topics' may explain the variance in the observed clinical phenotypes and aid in the identification of genetically homogeneous clusters for precision medicine.

We trained and optimized two different ETM models, one at the gene level using 20 topics, and another at the variant level using 100 topics. We used a higher number of topics in the variant level, due to the significantly larger size of the dataset. However, we still kept the number of topics small (<= 100) to improve interpretability and identify highly important topics in the model. The variant level topics were not very informative in either clustering or supervised classification model. However, the gene level ETM identified three small subclusters consisting predominantly of SCZ samples (Figure 7), which are not driven by population structure (supplementary Figure 2). These patients may constitute a genetically homogeneous subtype of SCZ. After looking at the correlation of each topic with the samples of each phenotypic class, we identified one particular topic to have high association with these subclusters, topic 4. The top 3 genes with highest probability under the topic were *MUC16*, *TTN* and *WASL*. The role of these genes in neuropsychiatric disorders has not been well established. However, *MUC16* was reported to be differentially expressed in the brains of BPD patients vs. controls, albeit not in SCZ vs. controls[81]. *WASL* has been reported to be part of a gene-gene interaction network that is overrepresented in SCZ cases[82], while rare de novo mutations have been reported in *TTN* in patients with autism[83].

Next, using the learned topics above, we performed supervised classification for each variant level and gene level ETM model outputs. Our classifier trained on the gene level topics (accuracy 52.4%) performed only slightly better than using variant level topics (50.5%). Although neither classifier performed well and the performance difference noticed is small (Table 3), if combined with fact that clustering of SCZ samples was better at the gene level approach discussed above (Figure 7) may suggest that dividing up the exome at the gene level

has a more informative semantic power than using the more granular variants level genotypes in the topic approach.

Lastly, in order to guide our ETM model to learn biologically meaningful topics, we modified the algorithm by fixing the gene embedding to the pathway x gene matrix in the gene-level topic modelling approach. This essentially guides the algorithm not to learn gene embedding based on the underlying co-variation of the genes but rather from the biologically defined gene molecular signature. However, the topics learned from this approach were not as informative for downstream clustering (Figure 10 vs. Figure 7), suggesting that gene similarity defined by co-involvement in similar pathways does not have more semantic power in a topic model than using co-occurrence of mutations in the genes.

4.3 ML analysis of WES to predict ASD status in affected and unaffected siblings

The last part of this project focused on predicting ASD status among affected and unaffected siblings using WES rare variants. Our previous approach, shown to be useful in separating BD cases against controls and ASD vs. SCZ, was not useful using the family dataset for ASD. This is likely due to the embedded assumption in most supervised ML algorithms, that each sample is independent of all others (i.i.d. assumption). It is clearly not the case in our ASD dataset, as we are dealing with siblings, not unrelated cases and controls. Each sibling pair is independent, but not each individual. Another factor to keep in mind when interpreting the above results is that, as per current knowledge, for an individual to be affected, a number of hits (in different susceptibility genes) need to accumulate beyond a certain threshold. Overall, there is a very large number of susceptibility genes so the combination of affected genes will

not be the same in different families. We expect that siblings share most of the disease-causing variants with each other but have small differences in their network account for the discordant disease status. Obviously, the genetic background and differences of the different sibpairs vary a lot across different families.

To address the limitations of supervised learning on cartesian representation of sibling data, we created a network/graph representation of the data. As a result, interconnectedness of individuals (family relationships) and variant features were taken into account in the new data structure. More specifically, we placed each sibling pair within the context of their source of genetics (i.e. parents) and each family within the broader context of all other families in a large heterogenous network. As a result, when training a graph neural network for disease-status classification in children, not only will the feature information from the local neighborhood nodes propagate to the children, but also the topological structures that these neighborhood nodes form with the children and their surrounding nodes.

We modified the R-GCN algorithm such that when information from each node gets propagated to its local neighbourhood, a relation-type specific multi-layered neural network learns the best way to transform the incoming data for aggregation at the source node for disease status classification. When limiting the approach to pathogenic and likely pathogenic variants, the approach performed slightly better than when all of the rare variants were included in the network. However, the overall performance was still poor, barely outperforming chance and underperforming XGBoost on Cartesian data format where siblings were considered independent and parent genomics were not taken into account (Table 4). The lack of success could, at least partially, be attributed to the large number of parameters in the model and high

sparsity in the network. It is also possible that our graph neural network architecture could benefit from transformation and aggregation functions which are inspired by genetics rather than using complex neural networks.

4.4 Machine learning in neuropsychiatry

Machine learning algorithms have been used for a wide variety of tasks in neuropsychiatry, with a main focus on disease status prediction and drug treatment response[84]. Most studies have based their models on neuroimaging data[84]. Using neuroimaging data for early diagnosis of neuropsychiatric disorders has grown exponentially since the early 2000's, with applications in Alzheimer's, ASD, SCZ, mood disorders and others, using methods ranging from early support vector machine (SVM) to more recent deep learning methods[85]. For example, recent deep learning models have achieved a balanced accuracy of up to 90% for identifying Alzheimer's and ADHD from MRI and fMRI neuroimaging data[84,86].

However, the use of machine learning methods on genomic data for risk prediction is more nascent, and mostly limited to using GWAS SNPs as indicated in the recent meta-analysis of research papers on the topic. Bracher-Smith et al[87] showed that ML on genetic data were mainly used in Schizophrenia (7 studies), bipolar disorder (5 studies) and ASD (3 studies), with GWAS SNPs data as the main input in almost all studies, and only two using WES data. For example, Krapohl et al.[88] combined summary statistics from 81 GWASs to train an elastic net model to predict genetic risk for educational attainment, general cognitive ability, and BMI outperforming traditional PRS methods. Another study by Cao et al.[89] combined SNPs and neuroimaging data (fMRI voxels) for biomarker discovery in SCZ samples using a generalized

sparse model, and evaluated the effectiveness of their method by training for accuracy on ~200 SCZ case control sample data.

Machine learning methods have also been used to analyze SNP data in precision psychiatry, a field of study concerned with personalized treatment of patients based on their genetic profile and environmental factors[90]. A recent study used a feed-forward fully-connected neural network architecture to predict antidepressant treatment response and remission in MDD patients using genetic (SNPs) and clinical biomarkers as input variables[91]. More specifically, Fernandes et al. used clinical biomarkers and ten SNPs, that had shown significant association with treatment response, as input features to train a neural network to predict treatment response. Their model showed a performance of 0.82 AUC on the test set. Similarly, they trained another neural network to predict remission, which showed a performance of 0.80 AUC. Other studies, such as the one by Chang et al.[92], have used multi-omic data (genetic variant + DNA methylation), along with neuroimaging data, to predict treatment response in MDD patients.

In contrast to the previous studies, our work focused on WES data, different ML approaches and contrasting patients with different genetically overlapping DSM diagnoses and controls. Our datasets were larger; focused on both rare and common variants; tried to integrate knowledge about genes expressed in the brain and gene-pathway information; and tried to tackle family-based data. We developed new approaches and methodologies that can be useful in advancing the field of complex (neuropsychiatric) genetics.

Chapter 5: Conclusion and Future Directions

When designing ML algorithm for WES data, an important problem affecting model performance and the ability to uncover meaningful relationships or insights is how the input data is represented. It is a part of the ML design process where domain expertise can add significant value. For example, addressing questions such as whether one should use the raw sequence of the exome/genome, or focus on individual variants versus on the aggregation of variants at the gene level, can greatly impact model performance. This thesis has offered some insights and novel ways of thinking about analyzing genomic data through ML and contributed towards this interdisciplinary direction of research.

We showed different machine learning approaches to the analysis of the whole exome sequencing data of neuropsychiatric disorders and controls that may help pave the way for methodology development that can enable better objective clinical risk prediction, subtyping, and classification of patients. A good predictive model could help medical geneticists diagnose patients at very high risk in pre-symptomatic stage and potentially allow for preventative measures or early treatment initiation. An example of how genetic factors can be used to this end may be our supervised model that successfully identified BD samples among SCZ and control individuals. In addition, the identification of genetically homogeneous clusters, such as the ones shown for SCZ samples using unsupervised interpretable ML methods, may enable targeted drug development as well as inform clinical trials for precision medicine in neuropsychiatric disorders. Overall, our methods were intentionally designed to focus on more interpretable ML models to help identify important genes and variants in order to enable better characterization and interpretation of the results. As a result, in addition to model

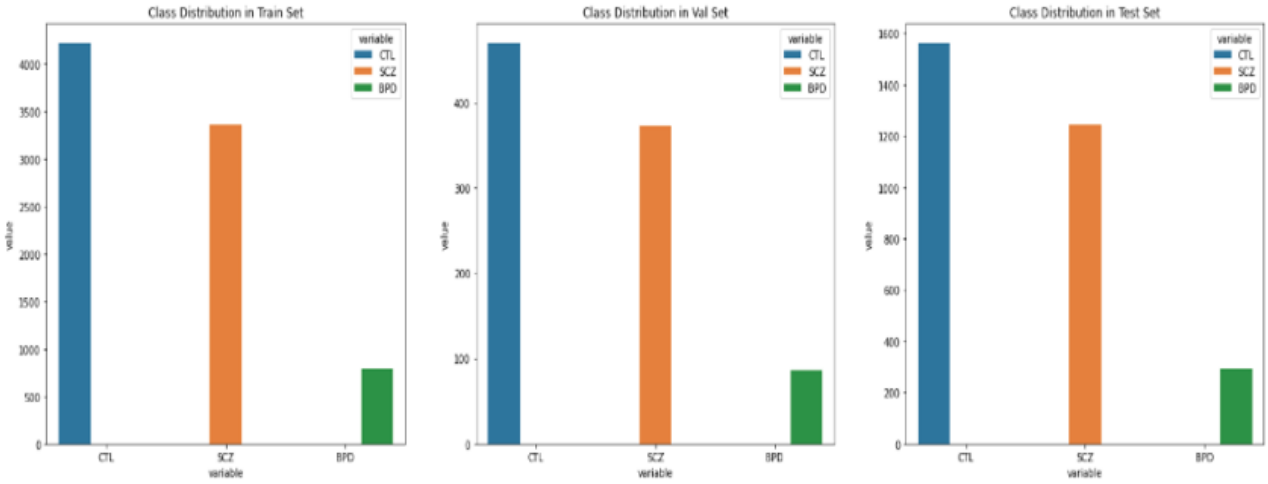
performance, we were able to report and interpret our findings from each approach, shedding further light on the biological characterization of the disorders under study.

It is worth noting that using whole exome sequencing data has its limitation as it is only focused on variants located inside protein-coding genes. Given the already established role of copy number variants (CNVs) and non-coding variants in neuropsychiatric disorders which are not captured through WES, the work summarized in this thesis is missing some important informative variables. Future studies can focus on whole genome sequencing to allow for training of more comprehensive ML models, which can combine different input datasets and be used for precision medicine.

In order to fully characterize the underlying molecular mechanism of neuropsychiatric diseases, we need to integrate several levels of biological information, using up to date technologies (e.g. genome sequencing rather than SNP arrays), and take into account the relationships among dependent biological entities (e.g. multi-omics data). However, integrating different data types brings its own set of challenges as each omic data type has different data modality. Interdependence among different entities such as variant, gene, protein etc. will need to be modeled as well, and methods such as the graph neural networks proposed in the ASD family-approach section may be useful to this end.

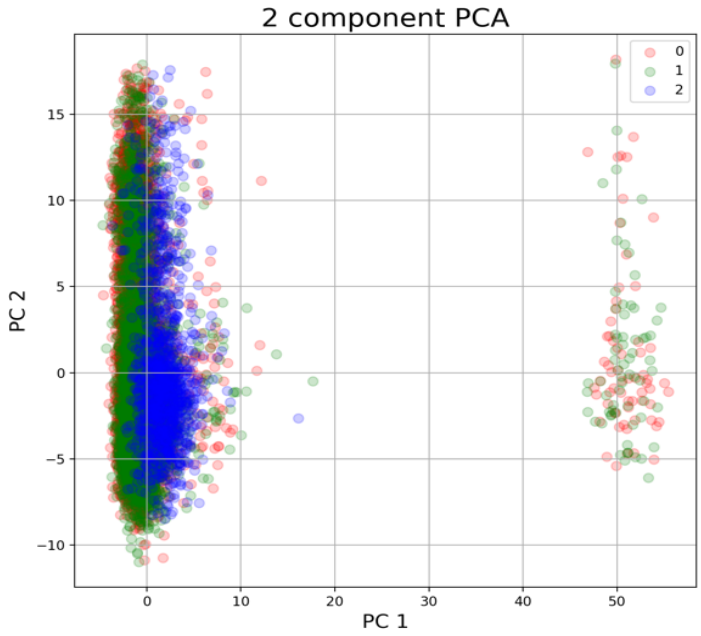
Supplementary Figures

Supplementary Figure 1: Distribution of classes in each training, validation and testing sets



Supplementary Figure 1. Class distribution in training, validation and testing sets for controls (CTL), schizophrenia (SCZ) and bipolar disorder (BD). The bipolar class is an extreme minority within the data. The same distribution of classes as in original population is kept in each of the three sets. During supervised learning analyses, the minority classes are weighted according to their inverse frequency in the population.

Supplementary Figure 2: PCA analysis of Swedish WES



Supplementary Figure 2. Result of population stratification using PCA on raw genotype data of the Swedish WES data shown here. Class 0 is control, class 1 SCZ and class 2 BD. We see overlapped samples from each control, SCZ and BD samples denoting no population stratification as expected given all samples come from a homogenous Swedish population. During plotting, control samples were drawn first, then SCZ and lastly BD resulting in covering each other samples in one large cluster. The small cluster (< 1% of total samples) shown on the right consist of both SCZ and controls.

References

1. Sullivan PF, Geschwind DH. Defining the Genetic, Genomic, Cellular, and Diagnostic Architectures of Psychiatric Disorders. *Cell*. 2019;177(1):162-183. doi:10.1016/j.cell.2019.01.015

2. Eaton WW, Martins SS, Nestadt G, Bienvenu OJ, Clarke D, Alexandre P. The Burden of Mental Disorders. *Epidemiol Rev*. 2008;30:1-14. doi:10.1093/epirev/mxn011

3. Petrou S, Johnson S, Wolke D, Hollis C, Kochhar P, Marlow N. Economic costs and preference-based health-related quality of life outcomes associated with childhood psychiatric disorders. *Br J Psychiatry*. 2010;197(5):395-404. doi:10.1192/bjp.bp.110.081307

4. Burmeister M, McInnis MG, Zöllner S. Psychiatric genetics: progress amid controversy. *Nature Reviews Genetics*. 2008;9(7):527-540. doi:10.1038/nrg2381

5. Barnhill JW. *DSM-5® Clinical Cases*. American Psychiatric Publishing; 2013. doi:10.1176/appi.books.9781585624836

6. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461(7265):747-753. doi:10.1038/nature08494

7. Sullivan PF, Agrawal A, Bulik CM, et al. Psychiatric Genomics: An Update and an Agenda. *Am J Psychiatry*. 2018;175(1):15-27. doi:10.1176/appi.ajp.2017.17030283

8. Grønborg TK, Schendel DE, Parner ET. Recurrence of Autism Spectrum Disorders in Full- and Half-Siblings and Trends Over Time. *JAMA Pediatr*. 2013;167(10):947-953. doi:10.1001/jamapediatrics.2013.2259

9. Bourgeron T. Current knowledge on the genetics of autism and propositions for future research. *Comptes Rendus Biologies*. 2016;339(7):300-307. doi:10.1016/j.crvi.2016.05.004

10. Werling DM, Geschwind DH. Sex differences in autism spectrum disorders. *Curr Opin Neurol*. 2013;26(2):146-153. doi:10.1097/WCO.0b013e32835ee548

11. Iossifov I, O'Roak BJ, Sanders SJ, et al. The contribution of de novo coding mutations to autism spectrum disorder. *Nature*. 2014;515(7526):216-221. doi:10.1038/nature13908

12. Griswold AJ, Dueker ND, Van Booven D, et al. Targeted massively parallel sequencing of autism spectrum disorder-associated genes in a case control cohort reveals rare loss-of-function risk variants. *Molecular Autism*. 2015;6(1):43. doi:10.1186/s13229-015-0034-z

13. Satterstrom FK, Kosmicki JA, Wang J, et al. Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell*. 2020;180(3):568-584.e23. doi:10.1016/j.cell.2019.12.036

14. Gaugler T, Klei L, Sanders SJ, et al. Most genetic risk for autism resides with common variation. *Nature Genetics*. 2014;46(8):881-885. doi:10.1038/ng.3039

15. Sullivan PF, Kendler KS, Neale MC. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch Gen Psychiatry*. 2003;60(12):1187-1192. doi:10.1001/archpsyc.60.12.1187

16. Barnhill JW. Schizophrenia Spectrum and Other Psychotic Disorders. In: *DSM-5 Clinical Cases*. DSM Library. American Psychiatric Publishing; 2013. doi:10.1176/appi.books.9781585624836.jb02

17. Häfner H, Maurer K, Löffler W, et al. The epidemiology of early schizophrenia. Influence of age and gender on onset and early course. *Br J Psychiatry Suppl*. 1994;(23):29-38.

18. Brainstorm Consortium, Anttila V, Bulik-Sullivan B, et al. Analysis of shared heritability in common disorders of the brain. *Science*. 2018;360(6395). doi:10.1126/science.aap8757

19. Purcell SM, Moran JL, Fromer M, et al. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature*. 2014;506(7487):185-190. doi:10.1038/nature12975

20. Ripke S, Neale BM, Corvin A, et al. Biological Insights From 108 Schizophrenia-Associated Genetic Loci. *Nature*. 2014;511(7510):421-427. doi:10.1038/nature13595

21. Rowland TA, Marwaha S. Epidemiology and risk factors for bipolar disorder. *Ther Adv Psychopharmacol*. 2018;8(9):251-269. doi:10.1177/2045125318769235

22. Barnhill JW. Bipolar and Related Disorders. In: *DSM-5 Clinical Cases*. DSM Library. American Psychiatric Publishing; 2013. doi:10.1176/appi.books.9781585624836.jb03

23. Edvardsen J, Torgersen S, Røysamb E, et al. Heritability of bipolar spectrum disorders. Unity or heterogeneity? *Journal of Affective Disorders*. 2008;106(3):229-240. doi:10.1016/j.jad.2007.07.001

24. McGuffin P, Rijsdijk F, Andrew M, Sham P, Katz R, Cardno A. The heritability of bipolar affective disorder and the genetic relationship to unipolar depression. *Arch Gen Psychiatry*. 2003;60(5):497-502. doi:10.1001/archpsyc.60.5.497

25. Dome P, Rihmer Z, Gonda X. Suicide Risk in Bipolar Disorder: A Brief Review. *Medicina (Kaunas)*. 2019;55(8). doi:10.3390/medicina55080403

26. Stahl EA, Breen G, Forstner AJ, et al. Genome-wide association study identifies 30 Loci Associated with Bipolar Disorder. *Nat Genet*. 2019;51(5):793-803. doi:10.1038/s41588-019-0397-8

27. Sul JH, Service SK, Huang AY, et al. Contribution of common and rare variants to bipolar disorder susceptibility in extended pedigrees from population isolates. *Translational Psychiatry*. 2020;10(1):1-10. doi:10.1038/s41398-020-0758-1

28. Forstner AJ, Fischer SB, Schenk LM, et al. Whole-exome sequencing of 81 individuals from 27 multiply affected bipolar disorder families. *Translational Psychiatry*. 2020;10(1):1-10. doi:10.1038/s41398-020-0732-y

29. Psychiatric GWAS Consortium Coordinating Committee, Cichon S, Craddock N, et al. Genomewide association studies: history, rationale, and prospects for psychiatric disorders. *Am J Psychiatry*. 2009;166(5):540-556. doi:10.1176/appi.ajp.2008.08091354

30. Bray NJ, O'Donovan MC. The genetics of neuropsychiatric disorders. *Brain Neurosci Adv*. 2018;2. doi:10.1177/2398212818799271

31. Lee SH, Ripke S, Neale BM, et al. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics*. 2013;45(9):984-994. doi:10.1038/ng.2711

32. Lee PH, Anttila V, Won H, et al. Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders. *Cell*. 2019;179(7):1469-1482.e11. doi:10.1016/j.cell.2019.11.020

33. Grove J, Ripke S, Als TD, et al. Identification of common genetic risk variants for autism spectrum disorder. *Nat Genet*. 2019;51(3):431-444. doi:10.1038/s41588-019-0344-8

34. Bamshad MJ, Ng SB, Bigham AW, et al. Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*. 2011;12(11):745-755. doi:10.1038/nrg3031

35. Wang W, Corominas R, Lin GN. De novo Mutations From Whole Exome Sequencing in Neurodevelopmental and Psychiatric Disorders: From Discovery to Application. *Front Genet*. 2019;10. doi:10.3389/fgene.2019.00258

36. Sanders SJ, Murtha MT, Gupta AR, et al. De novo mutations revealed by whole exome sequencing are strongly associated with autism. *Nature*. 2012;485(7397):237-241. doi:10.1038/nature10945

37. Feliciano P, Zhou X, Astrovskaya I, et al. Exome sequencing of 457 autism families recruited online provides evidence for autism risk genes. *NPJ Genom Med*. 2019;4:19. doi:10.1038/s41525-019-0093-8

38. Gratten J, Wray NR, Keller MC, Visscher PM. Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nat Neurosci*. 2014;17(6):782-790. doi:10.1038/nn.3708

39. Mitchell TM. *Machine Learning*.; 1997.

40. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet*. 2015;16(6):321-332. doi:10.1038/nrg3920

41. Nassif AB, Shahin I, Attili I, Azzeh M, Shaalan K. Speech Recognition Using Deep Neural Networks: A Systematic Review. *IEEE Access*. 2019;7:19143-19165. doi:10.1109/ACCESS.2019.2896880

42. Kłosowski P. Deep Learning for Natural Language Processing and Language Modelling. In: *2018 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*. ; 2018:223-228. doi:10.23919/SPA.2018.8563389

43. Landi I, Glicksberg BS, Lee H-C, et al. Deep representation learning of electronic health records to unlock patient stratification at scale. *NPJ Digit Med*. 2020;3:96. doi:10.1038/s41746-020-0301-z

44. Li Y, Nair P, Lu XH, et al. Inferring multimodal latent topics from electronic health records. *Nature Communications*. 2020;11(1):2536. doi:10.1038/s41467-020-16378-3

45. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat Commun*. 2019;10(1):380. doi:10.1038/s41467-018-08023-x

46. Amodio M, van Dijk D, Srinivasan K, et al. Exploring single-cell data with deep multitasking neural networks. *Nature Methods*. 2019;16(11):1139-1145. doi:10.1038/s41592-019-0576-7

47. Lopez R, Regier J, Cole M, Jordan M, Yosef N. A deep generative model for single-cell RNA sequencing with application to detecting differentially expressed genes. *arXiv:171005086*

[cs, q-bio, stat]. Published online October 16, 2017. Accessed February 5, 2021. http://arxiv.org/abs/1710.05086

48. Callaway E. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature*. 2020;588(7837):203-204. doi:10.1038/d41586-020-03348-4

49. Yu C, Arcos-Burgos M, Licinio J, Wong M-L. A latent genetic subtype of major depression identified by whole-exome genotyping data in a Mexican-American cohort. *Transl Psychiatry*. 2017;7(5):e1134. doi:10.1038/tp.2017.102

50. MacDonald K, Jiang Y, Krishnan A, et al. Patient Stratification Using Metabolomics to Address the Heterogeneity of Psychosis. *Schizophrenia Bulletin Open*. 2020;1(sgaa032). doi:10.1093/schizbullopen/sgaa032

51. Huang Z, Dong W, Duan H. A probabilistic topic model for clinical risk stratification from electronic health records. *Journal of Biomedical Informatics*. 2015;58:28-36. doi:10.1016/j.jbi.2015.09.005

52. Fischbach GD, Lord C. The Simons Simplex Collection: A Resource for Identification of Autism Genetic Risk Factors. *Neuron*. 2010;68(2):192-195. doi:10.1016/j.neuron.2010.10.006

53. dbGaP Study. Accessed February 4, 2021. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000687.v1.p1&phv=197315&phd=&pha=&pht=3695&phvf=&phdf=&phaf=&phtf=&dssp=1&consent=&temp=1

54. dbGaP Study. Accessed February 4, 2021. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000473.v2.p2

55. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*. 2005;102(43):15545-15550. doi:10.1073/pnas.0506580102

56. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. 2011;27(12):1739-1740. doi:10.1093/bioinformatics/btr260

57. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst*. 2015;1(6):417-425. doi:10.1016/j.cels.2015.12.004

58. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013;45(6):580-585. doi:10.1038/ng.2653

59. Hellwege J, Keaton J, Giri A, Gao X, Velez Edwards DR, Edwards TL. Population Stratification in Genetic Association Studies. *Curr Protoc Hum Genet*. 2017;95:1.22.1-1.22.23. doi:10.1002/cphg.48

60. Zhao Y, Chen F, Zhai R, et al. Correction for population stratification in random forest analysis. *International Journal of Epidemiology*. 2012;41(6):1798-1806. doi:10.1093/ije/dys183

61. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*. 2006;38(8):904-909. doi:10.1038/ng1847

62. Friedman JH. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*. 2001;29(5):1189-1232.

63. Natekin A, Knoll A. Gradient boosting machines, a tutorial. *Front Neurorobot*. 2013;7. doi:10.3389/fnbot.2013.00021

64. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. Association for Computing Machinery; 2016:785-794. doi:10.1145/2939672.2939785

65. Dieng AB, Ruiz FJR, Blei DM. Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics*. 2020;8:439-453. doi:10.1162/tacl_a_00325

66. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*. 2018;46(Database issue):D1062-D1067. doi:10.1093/nar/gkx1153

67. Wu Z, Pan S, Chen F, Long G, Zhang C, Yu PS. A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*. 2021;32(1):4-24. doi:10.1109/TNNLS.2020.2978386

68. Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph Attention Networks. *arXiv:171010903 [cs, stat]*. Published online February 4, 2018. Accessed February 5, 2021. http://arxiv.org/abs/1710.10903

69. Hamilton WL, Ying R, Leskovec J. Inductive representation learning on large graphs. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Curran Associates Inc.; 2017:1025-1035.

70. Schlichtkrull M, Kipf TN, Bloem P, van den Berg R, Titov I, Welling M. Modeling Relational Data with Graph Convolutional Networks. In: Gangemi A, Navigli R, Vidal M-E, et al., eds. *The Semantic Web*. Lecture Notes in Computer Science. Springer International Publishing; 2018:593-607. doi:10.1007/978-3-319-93417-4_38

71. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*. 2018;34(13):i457-i466. doi:10.1093/bioinformatics/bty294

72. Sardaar S, Qi B, Dionne-Laporte A, Rouleau GuyA, Rabbany R, Trakadis YJ. Machine learning analysis of exome trios to contrast the genomic architecture of autism and schizophrenia. *BMC Psychiatry*. 2020;20(1):92. doi:10.1186/s12888-020-02503-5

73. Hirokawa N, Tanaka Y. Kinesin superfamily proteins (KIFs): Various functions and their relevance for important phenomena in life and diseases. *Exp Cell Res*. 2015;334(1):16-25. doi:10.1016/j.yexcr.2015.02.016

74. Delevoye C, Heiligenstein X, Ripoll L, et al. BLOC-1 Brings Together the Actin and Microtubule Cytoskeletons to Generate Recycling Endosomes. *Curr Biol*. 2016;26(1):1-13. doi:10.1016/j.cub.2015.11.020

75. Tarabeux J, Champagne N, Brustein E, et al. De novo truncating mutation in Kinesin 17 associated with schizophrenia. *Biol Psychiatry*. 2010;68(7):649-656. doi:10.1016/j.biopsych.2010.04.018

76. Zhou R, Niwa S, Guillaud L, Tong Y, Hirokawa N. A molecular motor, KIF13A, controls anxiety by transporting the serotonin type 1A receptor. *Cell Rep*. 2013;3(2):509-519. doi:10.1016/j.celrep.2013.01.014

77. Ionita-Laza I, Xu B, Makarov V, et al. Scan statistic-based analysis of exome sequencing data identifies FAN1 at 15q13.3 as a susceptibility gene for schizophrenia and autism. *Proc Natl Acad Sci U S A*. 2014;111(1):343-348. doi:10.1073/pnas.1309475110

78. Forsingdal A, Fejgin K, Nielsen V, Werge T, Nielsen J. 15q13.3 homozygous knockout mouse model display epilepsy-, autism- and schizophrenia-related phenotypes. *Translational Psychiatry*. 2016;6(7):e860-e860. doi:10.1038/tp.2016.125

79. Bernkopf M, Webersinke G, Tongsook C, et al. Disruption of the methyltransferase-like 23 gene METTL23 causes mild autosomal recessive intellectual disability. *Hum Mol Genet*. 2014;23(15):4015-4023. doi:10.1093/hmg/ddu115

80. Reiff RE, Ali BR, Baron B, et al. METTL23, a transcriptional partner of GABPA, is essential for human cognition. *Hum Mol Genet*. 2014;23(13):3456-3466. doi:10.1093/hmg/ddu054

81. Darby MM, Yolken RH, Sabunciyan S. Consistently altered expression of gene sets in postmortem brains of individuals with major psychiatric disorders. *Transl Psychiatry*. 2016;6(9):e890. doi:10.1038/tp.2016.173

82. Chang S, Fang K, Zhang K, Wang J. Network-Based Analysis of Schizophrenia Genome-Wide Association Data to Detect the Joint Functional Association Signals. *PLoS One*. 2015;10(7). doi:10.1371/journal.pone.0133404

83. O'Roak BJ, Deriziotis P, Lee C, et al. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet*. 2011;43(6):585-589. doi:10.1038/ng.835

84. Durstewitz D, Koppe G, Meyer-Lindenberg A. Deep neural networks in psychiatry. *Molecular Psychiatry*. 2019;24(11):1583-1598. doi:10.1038/s41380-019-0365-9

85. Davatzikos C. Machine learning in neuroimaging: Progress and challenges. *Neuroimage*. 2019;197:652-656. doi:10.1016/j.neuroimage.2018.10.003

86. Ortiz A, Munilla J, Górriz JM, Ramírez J. Ensembles of Deep Learning Architectures for the Early Diagnosis of the Alzheimer's Disease. *Int J Neural Syst*. 2016;26(7):1650025. doi:10.1142/S0129065716500258

87. Bracher-Smith M, Crawford K, Escott-Price V. Machine learning for genetic prediction of psychiatric disorders: a systematic review. *Molecular Psychiatry*. Published online June 26, 2020:1-10. doi:10.1038/s41380-020-0825-2

88. Krapohl E, Patel H, Newhouse S, et al. Multi-polygenic score approach to trait prediction. *Mol Psychiatry*. 2018;23(5):1368-1374. doi:10.1038/mp.2017.163

89. Cao H, Duan J, Lin D, Shugart YY, Calhoun V, Wang Y-P. Sparse Representation Based Biomarker Selection for Schizophrenia with Integrated Analysis of fMRI and SNPs. *Neuroimage*. 2014;102 Pt 1:220-228. doi:10.1016/j.neuroimage.2014.01.021

90. Fernandes BS, Williams LM, Steiner J, Leboyer M, Carvalho AF, Berk M. The new field of 'precision psychiatry.' *BMC Med*. 2017;15. doi:10.1186/s12916-017-0849-x

91. Lin E, Kuo P-H, Liu Y-L, Yu YW-Y, Yang AC, Tsai S-J. A Deep Learning Approach for Predicting Antidepressant Response in Major Depression Using Clinical and Genetic Biomarkers. *Front Psychiatry*. 2018;9. doi:10.3389/fpsyt.2018.00290

92. Chang B, Choi Y, Jeon M, et al. ARPNet: Antidepressant Response Prediction Network for Major Depressive Disorder. *Genes (Basel)*. 2019;10(11). doi:10.3390/genes10110907