Advancing the understanding and treatment of psychiatric and other complex diseases through machine learning and omics

Bill Qi

Department of Human Genetics

McGill University, Montréal

December 2023

A thesis submitted to McGill University in partial fulfillment of the requirement of the degree of Doctor of Philosophy

First published on December 12, 2024

Abstract

Complex diseases are common diseases influenced by a combination of numerous genetic as well as environmental factors. Omics data, including genomics, transcriptomics, proteomics, and metabolomics, have contributed to our understanding of the pathophysiology of complex diseases and their treatment. The molecular basis of psychiatric disorders is beginning to be understood under the lens of various genomic-wide association studies, and omics studies. Furthermore, advances in pharmacogenomics have uncovered the associations between omics and medications. However, current approaches still mainly involve the use of univariate approaches, and there is a need for multivariate approaches to extract further insights from omics data.

In this thesis, we focus on psychiatric disorders, including schizophrenia (SCZ) and major depressive disorder (MDD), as our model of complex diseases. Psychiatric disorders have multifactorial etiologies with influence from both genetics and environmental factors, similar to other diseases with "complex inheritance". They display a broad range of symptoms and heterogeneity. Our understanding of the molecular basis of psychiatric disorders is suboptimal, and so is their medical treatment, thus making them important candidate diseases to focus on.

Machine learning is a field that focuses on the development of algorithms that can learn from data to uncover generalizable patterns which are useful for predictive, classification, or clustering purposes. We investigated the potential of ML in identifying and understanding the molecular basis of psychiatric disorders through omics. Furthermore, we extended the scope of ML to the broader context of treatment optimization for different multifactorial diseases based on pharmacogenomics.

More specifically, we applied an ML approach to analyze gene expression data from the database of Genotypes and Phenotypes (dbGaP) to improve our understanding of the pathophysiology of SCZ and MDD. Furthermore, we explored the use of supervised and unsupervised ML analysis of microRNA for disease severity and treatment response prediction in MDD. In our final chapter, we expanded the scope of the application of ML in the broader context of pharmacogenomics through the development of a graph-based approach for prioritizing medications based on individual genetic data from the United Kingdom Biobank (UKBB).

Our research has led to the development of general frameworks for ML analysis of omics data in the context of understanding the molecular basis of complex diseases, and in the context of pharmacogenomics for treatment optimization. With this thesis, we aim to advance the development of methods for understanding complex diseases and optimization of their treatment, which could be used as groundwork for future applications of ML towards more targeted (precision) medicine.

Résumé

Les maladies complexes sont des maladies courantes influencées par une combinaison de nombreux facteurs génétiques et environnementaux. Les données omiques, incluant la génomique, la transcriptomique, la protéomique et la métabolomique, ont contribué à notre compréhension de la physiopathologie des maladies complexes et de leur traitement. La base moléculaire des troubles psychiatriques commence à être comprise sous l'angle de diverses études d'association génomique à grande échelle et d'études omiques. De plus, les progrès en pharmacogénomique ont révélé les associations entre les omiques et les médicaments. Cependant, les approches actuelles impliquent encore principalement l'utilisation d'approches univariées, et il est nécessaire de recourir à des approches multivariées pour extraire de nouvelles informations à partir des données omiques.

Dans cette thèse, nous nous concentrons sur les troubles psychiatriques, notamment la schizophrénie (SCZ) et la dépression majeure (MDD), en tant que modèle de maladies complexes. Ces troubles psychiatriques ont des étiologies multifactorielles influencées à la fois par des facteurs génétiques et environnementaux, similaires à d'autres maladies à "héritage complexe". Ils présentent une large gamme de symptômes et une grande hétérogénéité. Notre compréhension de la base moléculaire de ces troubles est sous-optimale, et leur traitement médical est également insuffisant, ce qui en fait d'importants candidats sur lesquels se concentrer.

L'apprentissage automatique est un domaine qui se concentre sur le développement d'algorithmes capables d'apprendre à partir de données pour découvrir des modèles généralisables qui peuvent être utiles à des fins de prédiction, de classification ou de regroupement. Nous avons étudié le potentiel de l'apprentissage automatique pour identifier et comprendre la base moléculaire des troubles psychiatriques à travers les omiques. De plus, nous

avons élargi le champ d'application de l'apprentissage automatique au contexte plus large de l'optimisation du traitement pour différentes maladies multifactorielles basée sur la pharmacogénomique.

Plus précisément, nous avons utilisé une approche d'apprentissage automatique pour analyser les données d'expression génique de la base de données des génotypes et des phénotypes (dbGaP) afin d'améliorer notre compréhension de la physiopathologie de la SCZ et de la MDD. De plus, nous avons exploré l'utilisation d'analyses d'apprentissage automatique supervisées et non supervisées des microARN pour la prédiction de la gravité de la maladie et de la réponse au traitement dans la MDD. Dans notre dernier chapitre, nous avons élargi le champ d'application de l'utilisation de l'apprentissage automatique dans le contexte plus large de la pharmacogénomique en développant une approche basée sur des graphes pour hiérarchiser les médicaments en fonction des données génétiques individuelles provenant de la Biobanque du Royaume-Uni (UKBB).

Nos recherches ont conduit au développement de cadres généraux pour l'analyse de l'apprentissage automatique des données omiques dans le contexte de la compréhension de la base moléculaire des maladies complexes et dans le contexte de la pharmacogénomique pour l'optimisation du traitement. Avec cette thèse, nous visons à faire progresser le développement de méthodes pour comprendre les maladies complexes et optimiser leur traitement, ce qui pourrait servir de base pour de futures applications de l'apprentissage automatique vers une médecine plus ciblée (de précision).

Table of Contents

List of Abbreviations

ANN	Artificial Neural Network
ATC	Anatomical Therapeutic Chemical
AUC	Area Under the Receiver-Operating Characteristics Curve
BA	Brodmann's Area
BCE	Binary Cross-Entropy
CARTs	Classification and Regression Trees
CIHR	Canadian Institutes of Health Research
COX	Cytochrome C Oxidase
COX6A1	Cytochrome C Oxidase Subunit 6A1
CTL	Control
DAAM2	Dishevelled Associated Activator of Morphogenesis 2
DEGs	Differentially Expressed Genes
DLPFC	Dorsolateral Prefrontal Cortex
DMHUI	Douglas Mental Health University Institute
DNA	Deoxyribonucleic Acid
DNN	Deep Neural Network
DSM	Diagnostic and Statistical Manual of Mental Disorders
DSM-IV	Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition
EDTA	Ethylenediaminetetraacetic Acid
EEG	Electroencephalogram
ENPEP	Glutamyl Aminopeptidase
FDR	False Discovery Rate
FN	False Negative
FP	False Positive
FPR	False-Positive Rate
GABA	Gamma-Aminobutyric Acid
GBM	Gradient Boosted Machines
GCN	Graph Convolutional Network
GEO	Gene Expression Omnibus
GNN	Graph Neural Network
GO	Gene Ontology
GPL	GEO Platform License
GRCh37	Genome Reference Consortium Human Build 37
GRL	Graph Representation Learning
GSA	Gene Set Analysis
GWAS	Genome-Wide Association Study
HBHL	Healthy Brains, Healthy Lives
KEGG	Kyoto Encyclopedia of Genes and Genomes

MADRS	Montgomery-Asberg Depression Rating Scale
MDD	Major Depressive Disorder
MDE	Major Depressive Episode
ML	Machine Learning
MSigDB	Molecular Signatures Database
NARSAD	National Alliance for Research on Schizophrenia and Depression
NESDA	Netherlands Study of Depression and Anxiety
NGS	Next Generation Sequencing
NMDAR	N-Methyl-D-aspartate Receptor
NRES	Non-responder
NTR	Netherland Twin Register
OXPHOS	Oxidative Phosphorylation
PCR	Polymerase Chain Reaction
PGx	Pharmacogenetics
PharmGKB	The Pharmacogenomics Knowledgebase
PrP(C)	Cellular Prion Protein
Q	Phred quality
R	R programming language
RAS	Renin-Angiotensin System
RCT	Randomized Controlled Trial
RES	Responder
RIN	RNA Integrity Number
RNA	Ribonucleic Acid
ROC	Receiver-operating characteristics curve
SCID-I	Structured Clinical Interview for DSM-IV
SCID-IA	Structured Clinical Interview for DSM-IV Axis I Disorders
SCZ	Schizophrenia
SD	Standard Deviation
SGD	Stochastic Gradient Descent
SSD	Subsyndromal Symptomatic Depression
SSRIs	Selective Serotonin Reuptake Inhibitors
SVM	Support Vector Machines
T0	Timepoint 0
T8	Timepoint 8
TGF-β	Transforming Growth Factor-beta
TN	True Negative
TP	True Positive
TPR	True-Positive Rate
UKBB	United Kingdom Biobank
US	United States

XGBoost	Extreme Gradient Boosting
cDNA	Complementary DNA
eQTL	Expression Quantitative Trait Loci
dbGaP	Database of Genotypes and Phenotypes
mRNA	Messenger RNA
p75NTR	p75 Neurotrophin Receptor

List of Figures

Chapter 1:

Chapter 2:

Chapter 3:

Chapter 5:

List of Supplemental Figures

Chapter 5:

List of Tables

List of Supplemental Tables

Chapter 2:

Chapter 3:

Acknowledgements

I am grateful to the numerous individuals who have supported me throughout my graduate studies, providing encouragement, guidance, and assistance at every stage of this journey.

First and foremost, I would like to express my appreciation to my supervisor, Dr. Yannis Trakadis, for believing in me and giving me the opportunity to work under his supervision. I am grateful for his constant guidance and understanding throughout these years. Furthermore, my sincere gratitude goes to my supervisory committee members, Dr. Celia Greenwood and Dr. Jeff Xia, for their invaluable guidance, expertise, and feedback throughout this process.

I would like to acknowledge my fellow lab member, Mr. Sameer Sardaar, for the countless insightful discussions and collaborative efforts we have shared over the years. I would also like to thank my co-authors, Ms. Sonia Boscenco, Ms. Janani Ramamurthy, and Ms. Imane Bennani for their diligent assistance with our manuscripts.

Special thanks to Dr. Laura M. Fiori and Dr. Gustavo Turecki for their collaboration on the microRNA project, a significant contributing chapter in this thesis, which provided valuable perspectives and knowledge to this work.

I am also grateful to the thesis reviewers for dedicating their time to evaluate my thesis.

Furthermore, I would like to extend my appreciation to Mr. Ross MacKay and Ms. Rimi Joshi for their assistance in ensuring an efficient administrative process throughout my graduate studies.

I am also thankful to the Canadian Institutes of Health Research (CIHR) for providing me with a scholarship that allowed me to focus on my research.

Finally, to my friends and family, your belief in me has been a constant source of motivation, and I am always grateful for your presence in my life.

Contribution to Original Knowledge

The work presented in this thesis is a substantial contribution to the development of machine learning (ML) approaches for understanding complex diseases, with novel model selection, evaluation, and interpretation methodologies to ensure the robustness of the insights derived from omics data. Using psychiatric disorders as models of complex diseases, we identified genes, and molecular functions associated with schizophrenia and major depressive disorder (Chapters 2 and 3). Furthermore, our work demonstrated the potential of ML approaches for disease severity monitoring and treatment response prediction (Chapter 4). Lastly, we highlighted the effectiveness of a novel approach based on graph representation learning for integrating biomedical domain knowledge into ML for medication usage prediction based on individual genotype data.

Format of the Thesis

The work presented in this thesis follows the manuscript-based format guidelines of the Department of Graduate and Postdoctoral Studies. The work consists of three published manuscripts (Chapters 2, 3, and 4), as well as a manuscript in preparation for submission (Chapter 5). The manuscript chapters and journals they are published in are as follows:

- Chapter 2: Transcriptomics and machine learning to advance schizophrenia genetics: A case-control study using post-mortem brain data. *Computer Methods and Programs in Biomedicine* 214: 106590.
- Chapter 3: Machine learning and bioinformatic analysis of brain and blood mRNA profiles in major depressive disorder: A case–control study. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 186(2): 101-112.
- Chapter 4: Machine Learning Analysis of Blood microRNA Data in Major Depression: A Case-Control Study for Biomarker Discovery. *International Journal of Neuropsychopharmacology* 23(8): 505-510.
- Chapter 5: Graph representation learning for the prediction of medication usage in the UK Biobank based on pharmacogenetic variants. *(In preparation for submission)*

Contribution of Authors

Bill Qi prepared this thesis under the supervision of Yannis Trakadis. All work included as part of this thesis was performed under the supervision of Dr. Yannis Trakadis. Yannis Trakadis reviewed the thesis and provided constructive feedback.

Chapter 2 is a manuscript authored by Bill Qi, Sonia Boscenco, Janani Ramamurthy, and Yannis, J. Trakadis. Bill Qi performed the bioinformatic and machine learning analyses and drafted the manuscript under the supervision of Yannis Trakadis who conceived and coordinated the project. Bill Qi and Yannis Trakadis designed the original methodology. Sonia Boscenco and Janani Ramamurthy performed the background literature review. All authors reviewed and provided feedback on the manuscript.

Chapter 3 is a manuscript authored by Bill Qi, Janani Ramamurthy, Imane Bennani, and Yannis J. Trakadis. Bill Qi performed the bioinformatic and machine learning analyses and drafted the manuscript under the supervision of Yannis Trakadis who conceived and coordinated the project. Bill Qi and Yannis Trakadis designed the original methodology. Imane Bennani and Janani Ramamurthy performed the background literature review. All authors reviewed and provided feedback on the manuscript.

Chapter 4 is a manuscript authored by Bill Qi; Laura M. Fiori; Gustavo Turecki, and Yannis J. Trakadis. Bill Qi performed the bioinformatic and machine learning analyses and drafted the manuscript under the supervision of Yannis Trakadis who conceived and coordinated the project. Bill Qi and Yannis Trakadis designed the original methodology. Gustavo Turecki coordinated patient recruitment and data collection. Laura Fiori oversaw the production of the microRNA data and put together the Sample Processing and Small RNA-seq sections of the methodology section. All authors reviewed and provided feedback on the manuscript.

Chapter 5 is a manuscript authored by Bill Qi and Yannis J. Trakadis. Bill Qi performed the machine learning analyses and drafted the manuscript under the supervision of Yannis Trakadis who conceived and coordinated the project. Bill Qi and Yannis Trakadis designed the original methodology. All authors reviewed and provided feedback on the manuscript.

Chapter 1. General Introduction

1. Complex diseases

Complex diseases, also commonly referred to as multifactorial or polygenic diseases, have a genetic basis (i.e., are heritable) consisting of the combined effects of and interactions between multiple genetic and environmental factors in their etiology and progression [1, 2]. Examples of complex diseases include cardiovascular disease, diabetes, cancer, autoimmune diseases, as well as psychiatric diseases, and many more. In contrast to Mendelian diseases, which often involve a single gene and well-documented inheritance types [3], complex diseases do not have well-established patterns of inheritance. However, a polygenic inheritance model with a liability determined by the minor effects of many genetic variants has been proposed to explain the underlying inheritance of complex diseases [1]. The polygenic inheritance model has been validated by the identification of numerous genetic loci associations based on genome-wide association studies (GWAS) [2], and the genetic basis of thousands of complex human diseases and traits and diseases have been elucidated [4].

However, GWASs thus far have explained only a limited proportion of the heritability of many complex diseases. This could be due to the fact that GWAS is a univariate approach that considers only the linear effect of each genetic variant separately (i.e., without considering the heterogeneity of effects, non-linear effects, as well as epistasis and gene-environment interactions), which hinders the power for detecting significant loci [5]. These limitations suggest a need for the development of novel multivariate methods to advance our understanding of complex diseases.

2. Role of omics in complex diseases

In complex diseases, the impact of each genetic variant on the disease is complex and interconnected through multiple molecular functions, pathways, and networks, rarely involving only a single factor [6]. Moreover, non-genetic factors including social, environmental, and lifestyle add additional layers of complexities to complex diseases [7].

Effective means of comprehensively quantifying biological molecules are needed to characterize complex diseases, and identify biomarkers associated with outcomes of interest (e.g., disease severity, recurrence, response to treatment, etc.). In contrast to monogenic diseases, complex diseases more often involve the collective effects of hundreds to thousands of genetic and non-genetic factors. Identifying effective means of treating complex diseases is expected to be more challenging, and comprehensive measurements of the biological system are needed to enable a better understanding of the causal mechanisms of disease for the development of better preventive and therapeutic strategies.

Recently, advancements in high-throughput technologies, such as sequencing and mass spectrometry, have enabled the efficient capture of entire classes of genomic, epigenomic, transcriptomic, proteomic, metabolomic, and other "omics" classes of molecules, ranging from genetic variants to RNA (ribonucleic acid) transcripts and protein abundances to products of metabolism [8].

Omics data have been applied to complex disease research to characterize the pathophysiology of complex diseases. As mentioned before, efficient quantification of genetic variants has enabled the identification of the genetic basis of complex diseases using GWAS, as well as leading to the development of disease risk prediction tools such as polygenic risk scores (PRS) [9]. Transcriptomics, which quantifies genome-wide gene expression levels, and other

functional omics classes have enabled a further understanding of the biological mechanisms of complex diseases at the gene and transcript levels through the identification of expression qualitative trait loci (eQTLs) [10], differential gene expression signatures [11], and biological pathways and networks [12].

In addition to the four major omics classes (genomics, transcriptomics, proteomics, and metabolomics), more recent developments include epi-omics (e.g., epigenomic, epitranscriptomics, and epiproteomics), which focuses on the modifications of the existing omics molecules, as well as more specialized knowledge-based omics focused on specific diseases such as immunomics, microbiomics, and redoxomics [13].

Furthermore, omics data could potentially improve the treatment of patients with complex diseases through precision medicine [14]. Precision medicine is an approach towards disease treatment by taking into account individual variability using methods for detailed characterization of patients (e.g., omics) and computational tools to enable more effective treatment and prevention of disease [14]. Pharmacogenomics, as a cornerstone of precision medicine, deals with the study of how variations of genomics and other omics influence response to medications at the individual level with the aim of improving medication efficacy and reducing side effects [15].

Precision medicine driven by omics has led to many advances in the biomedical field. Most notably in the field of oncology, precision medicine has already shown significant potential in improving patient outcomes. For instance, the use of genomics has led to the identification of gene targets for directed therapy, resulting in positive clinical effects in individuals who did not respond to conventional chemotherapy [16]. Applications for resolving disease heterogeneity have also revealed disease subtypes with unique genetic and epigenetic signatures associated

with survival outcomes have been identified in cancer patients [17]. Furthermore, functional omics such as transcriptomics capturing both genetic and non-genetic information relevant to disease have been shown to be useful for informing treatment of breast cancer [18].

Beyond cancers, much of the advancement in precision medicine of non-cancer related complex diseases has been in terms of genomics. Risk prediction methods such as PRS have been found to have strong evidence of clinical utility in cardiovascular diseases for guiding diagnosis and management of disease [19]. Moreover, in various complex diseases, ranging from cardiovascular diseases, diabetes, obesity, allergic diseases, and psychiatric diseases, the use of genomics and other omics along with other data modalities such as clinical and imaging data have been shown to be useful for risk, outcome, and treatment response prediction, as well as for identify subgroups of patients with similar characteristics to reduce the heterogeneity of complex diseases [20].

Individual omics data types have shown some successes, however, the emerging trends in the field of include the use of multi-omics integration [21] as well as single-cell omics [22]. Multi-omics integration can provide a more comprehensive and unbiased view of complex diseases compared to using single omics modalities alone. For example, multi-omics approaches have been used for treatment response prediction in immune-mediated diseases [23]. Furthermore, single-cell omics approaches can provide a higher resolution than current bulk omics approaches, and potentially provide a better understanding of complex diseases and enable precision medicine. For example, single-cell transcriptomics have been used to identify cell-type specific gene signatures to enable a more precise understanding of Alzheimer's disease [24].

While these approaches appear promising, there are still challenges which need to be addressed, particularly in terms of data analysis including sample size, high-dimensionality and

noise, and computational methods such as machine learning would be necessary to leverage omics data to further advance precision medicine [20, 25].

3. Psychiatric disorders

We choose to study psychiatric (or mental) disorders, in particular schizophrenia (SCZ) and major depressive disorder (MDD), under the framework of complex diseases. These selected disorders constitute common conditions that affect thinking, behavior, emotions, and mood. A recent meta-analysis study estimated a median lifetime prevalence of SCZ and other psychotic disorders to be 0.75% [26]. For MDD, the prevalence varies between different demographic categories, however, a recent study estimated the lifetime prevalence to be around 20.6% [27].

These disorders are moderate to highly heritable. For SCZ, an analysis of twin studies has estimated the heritability to be around 81% [28]. For MDD, a similar analysis has estimated the heritability to be around 37% [29]. Furthermore, environmental factors also play a role in these disorders. Early life adversity, such as childhood trauma, has been linked to both SCZ and MDD [30, 31]. Moreover, the interaction between genetic risk factors and environmental stressors may be important in the development of these disorders, suggesting gene-environment interactions [32, 33].

Treatment of psychiatric disorders is suboptimal. For example, 10-30% of patients with SCZ show no response to conventional antipsychotics, with 30-60% of patients experiencing side effects and little improvement [34]. A similar lack of successful treatment has been reported for MDD [35]. Despite the significant genetic basis of these disorders, psychiatry, the specialty concerned with the diagnosis and treatment of mental disorders, has lagged behind in the use of advanced diagnostic and therapeutic technologies relative to other clinical specialties [36]. For

example, the classification and diagnosis of psychiatric disorders are currently based on signs and symptoms criteria documented in the Diagnostic and Statistical Manual of Mental Disorders (DSM) [37], rather than underlying biological factors, as there is a poor understanding of the pathophysiology of these disorders [36].

4. Role of omics in psychiatric disorders

Recent GWAS have identified numerous genetic associations for SCZ [38] and MDD [39], and confirmed the polygenic nature of these psychiatric disorders. Genomics has also yielded risk predictors for these disorders through the development of PRSs [40], and machine learning-based methods [41].

The use of transcriptomic reference panels has enabled the identification of hundreds of transcriptome-wide associations encompassing 157 unique genes in SCZ [42]. Furthermore, the use of gene expression microarrays from cases with SCZ and controls has identified differentially expressed genes of various biological processes including neurotransmission, synaptic plasticity and potentiation, gene regulation, cell cycle progression, etc. [43, 44].

Similarly, proteomic studies have highlighted differences at the peptide and protein level for psychiatric disorders, including replicated differences in insulin-related peptides, interleukins, and brain-derived neurotrophic factor for SCZ, and differences in proteins involved in inflammation, and oxidative stress for MDD [45]. Lastly, at the end of the omics cascade, metabolomics and lipidomics have provided insights into the products of metabolism involved in psychiatric disorders, such as eicosanoids, a class of metabolites involved in inflammatory processes [46].

5. Pharmacogenomics

Pharmacogenomics combines the study of pharmacology and genomics to understand how an individual's genetic variations influence their response to drugs [47]. Pharmacogenomics is particularly relevant to complex diseases. Patients with complex diseases often have significant variations in their responses to drug treatment and experience of side-effects [48]. Pharmacogenomics holds the promise of improving the management of these diseases by enabling more precise and effective drug therapies based on the molecular basis of complex diseases.

Pharmacokinetics and pharmacodynamics are two key aspects of pharmacology that are directly influenced by genetic variation [49]. Pharmacokinetics refers to how a drug is absorbed, distributed, metabolized, and excreted by the body [50], while pharmacodynamics involves the biochemical and physiological effects of drugs and their mechanisms of action [51]. Genetic variations can affect both these processes, leading to individual variations in drug response phenotypes.

6. Machine learning

Machine learning (ML) is a branch of artificial intelligence focused on developing algorithms and methods for learning from data. Broadly, ML aims to identify generalizable patterns and knowledge underlying a series of data observations with the goal of making accurate predictions [52]. The two primary types of ML are supervised and unsupervised. With supervised ML, the goal is to train an algorithm using labeled data. Each data sample consists of a set of input features and a target variable. The algorithm is then trained to predict the target variable based on the values of the input features. Examples of popular supervised ML methods include decision

trees [53], support vector machines [54], and gradient-boosting machines [55]. With unsupervised learning, the data is unlabeled, and the focus is to identify statistical structure and patterns within the input data. Some examples of unsupervised learning include clustering using the K-means algorithm [56], and dimensionality reduction using the uniform manifold approximation and projection algorithm [57].

In contrast to classical statistical methods (e.g., null hypothesis testing, parameter estimation), ML focuses on making accurate predictions at the individual level. This capability is useful for the study of complex diseases, where individual-level variation (i.e., heterogeneity) is a significant contributing factor to complexity. Furthermore, ML methods are well suited for handling complex, high-dimensional data such as omics data through data representation, feature selection and engineering, and regularization techniques [58].

Graph representation learning

Graph representation learning (GRL) is a subset of approaches in ML focused on effectively representing and encoding graph-structured data to be used in ML algorithms [59]. Graphs are mathematical structures that use edges or connections to represent relationships between different entities. For example, social networks model the social connections between individuals. Furthermore, biological networks, such as protein-protein interaction networks, model the associations between proteins [60]. Graph neural networks (GNN) are a class of GRL methods which generalizes deep neural networks for the analysis of graph data. One popular method of GNN is graph convolutional networks (GCNs), which learn low-dimensional embedding representations of nodes and their local graph structure using an efficient localized first-order approximation of spectral graph convolutions [61].

7. Role of machine learning for modelling psychiatric disorders

ML has been extensively explored for modelling psychiatric disorders. However, we will focus on its utilization for individual-level prediction or classification purposes, primarily using omics data types, which is relevant to the context of this thesis.

Supervised ML analysis of genomic data (e.g., single nucleotide polymorphisms, copy number variations, and exome variants) has mainly been conducted to examine the diagnostic utility of ML. A systematic review showed that the area under the receiver operating characteristic curve (AUC) performance metric varied widely from 0.48 to 0.95 among psychiatric disorders including SCZ, bipolar disorder, autism, and anorexia [41]. However, a lack of common ML practices including model validation procedures, hyperparameter search, and a lack of general reporting guidelines in the field were pointed out as potential limitations for the validity of the findings.

Similarly, ML analysis of transcriptomic data in psychiatric disorders has also been conducted. Several studies have explored the classification of psychiatric disorders using transcriptomic data and ML. Most studies utilized an approach involving the identifications of differentially expressed genes, disease relevance analysis, and modelling using supervised ML for classification of psychiatric diseases [62-67]. Nearly all studies utilized differential expression signatures in blood as input features for ML. However, one study did perform a comparison of the use of messenger RNA and long non-coding RNA from brain tissue as input features for ML modelling to identify biomarkers for SCZ [68]. Reported ML performance metrics are typically moderate to high, with several studies reporting over 90% AUC metrics, however, all studies have used small datasets with at most a few hundred samples. Furthermore,

most studies used traditional ML algorithms such as support vector machines, random forests, but a few studies did make use of more recent neural network algorithms or ensemble methods.

Relatively few studies have examined the use of ML with proteomic data. ML analysis of proteomic data has been explored for diagnosing bipolar vs. unipolar depression [69]. Recent work has demonstrated the value of proteomic data as a predictor of remission in MDD [70]. Proteomic biomarkers have also been successfully used to predict the onset of psychotic disorder in clinically high-risk individuals [71]. Similarly, a few studies have explored the use of ML analysis for metabolomic data. Studies have used metabolomic profiles to distinguish depression cases from healthy controls [72, 73], and to predict depressive symptoms [74]. A study has also been conducted involving differential diagnosis of SCZ and bipolar disorder using metabolomic profiles [75].

Additionally, there have been a few examples of multi-modal approaches, such as using methylome and transcriptome data to distinguish suicide attempters, MDD cases, and healthy controls [76], and using a combination of omics and clinical data for the prediction of remission in MDD [70].

ML modelling of the pharmacogenomics aspects of psychiatric disorders has also been a research focus. The main goal of this area of research is to predict the individual-level treatment response to a medication and remission based on genomic features. For example, there have been promising findings in the prediction of response to antidepressants [77], which are a broad class of medications used to treat MDD, and antipsychotics [78], which are used to treat SCZ.

Currently, most studies have been focused on case-control designs for biomarker discovery and diagnostic purposes, with few studies on prognosis and treatment response modelling. All studies applying ML in modelling psychiatric disorders usually have a low

number of samples, while omics data are high-dimensional. In these scenarios, almost all ML algorithms can easily identify a model by overfitting to the data resulting in inflated performance results. Thus, the emphasis of ML modelling for these types of datasets needs to be on methodologies to reduce the risk of overfitting and proper model validation.

8. Hypothesis and objectives

In this thesis, our overarching hypothesis is that ML analysis of omics data can enhance our understanding of the pathophysiology of, and treatment for, complex diseases.

In Chapter 2, we utilized SCZ as our model for complex disease, and gene expression microarray data for the development of our ML methodology. Our specific hypothesis was that our ML methodology could distinguish disease cases from controls based on gene expressions from the dorsolateral prefrontal cortex (DLPFC) better than random chance and generalize to unseen data (i.e., not overfitting). We further hypothesized that genes with significant relevance to disease could be prioritized through the integration of genes identified through ML and biological gene set (pathway) analysis. Our objective was to leverage established ML methods for high-dimensional omics data and model selection procedures to reduce the risk of overfitting. More specifically, we aimed to develop ML methodologies for analyzing omics data to identify and understand the molecular basis of complex diseases, which could be valuable for the identification of treatment targets and the development of novel treatments.

In Chapter 3, we focused on MDD and improved upon our initial methodology by incorporating model evaluation on independent datasets, analyzing blood gene expression data, and utilizing covariate data for model interpretation. Our hypothesis was that patterns identified

through ML in MDD are relevant to disease and generalizable across independent datasets. Moreover, we hypothesized that ML analysis of both blood and DLPFC gene expressions can yield useful diagnosis biomarkers for MDD.

In Chapter 4, our hypothesis was that ML analysis of omics data could yield biomarkers for diagnosis, disease severity, and treatment response. Our objective was to investigate the application of ML and microRNA data for optimizing the treatment of MDD. We focused on an analysis of microRNA data from MDD patients to identify robust biomarkers for disease severity and treatment response.

Furthermore, in Chapter 5, we hypothesized that the incorporation of pharmacogenomic domain knowledge as part of ML modelling could enhance performance in the prediction of medication usage to advance precision medicine. Here, our objective was to explore the potential of ML for prioritizing medications that a patient may require based on their genetic data. We focused on larger-scale genomic data analysis using more advanced deep learning approaches and developed a novel graph-based methodology for incorporating pharmacogenomic domain knowledge as part of ML modelling.

9. Background

Omics

Genomics is the fundamental level of omics and provides insights into the potential associations of genetic variations and diseases. Genetic variations can often lead to cascading events in downstream omics levels [79]. We will explore an analysis of genetic variation data in Chapter 5. Transcriptomics focuses on the transcription of genetic information stored in DNA to RNA transcripts, which have crucial functional properties in transcript splicing, protein synthesis, and

gene regulation. For example, microRNAs (miRNA) are a class of non-coding RNAs that have been shown play roles such as suppressing the translation of messenger RNA (mRNA) into proteins, and upregulating transcription when bound to promoter DNA sequences [80]. In Chapters 2 and 3, we focus on the analysis of mRNA data, while Chapter 4 shifts the focus on to miRNA data. Proteomics focuses on the study of the set of proteins encoded through translation of mRNA, including their expression, structure, and function [81]. Lastly, metabolomics focuses on the study of a comprehensive set of metabolites including lipids, glycans, and other small molecules produced as a result of biochemical reactions [82]. Figure 1 illustrates the major omics categories and their relationships.

Omics data generation

The work in this thesis makes use of genomic and transcriptomic data based on microarray [83], and next-generation sequencing (NGS) [84], molecular biology data generation technologies, in particular expression and genotyping microarrays, and RNA sequencing (RNA-seq), which will be introduced in further detail.

Microarrays can be used to genotype multiple regions of the genome simultaneously. They can also be used for the purpose of quantifying the expression levels of genes. Typically, microarrays utilize a 2D matrix structure consisting of DNA probes fixed to beads placed on a surface (e.g., glass or silicon). The probes correspond to a predefined set of sequences of interest (e.g., sequences containing single nucleotide polymorphisms for genotyping microarrays, and gene sequences for expression microarrays).

For expression microarrays, the goal is to quantify the level of mRNA molecules corresponding to genes. Typically, RNA extraction is performed to extract total RNA from cells

or tissues. The mRNA portion of total RNA is selected based on hybridization binding of the mRNA poly(A) tail sequences to extraction beads. To be compatible with the DNA probes on the expression microarray, the mRNA is reverse transcribed into complementary DNA (cDNA) molecules with nucleotides which are fluorescently labeled. Next, the cDNA is hybridized to the probes on the microarray. Finally, a scanning process measures the fluorescence signals emitted by the hybridized cDNA and used as a proxy for the expression level of the corresponding genes.

With genotyping microarrays, a similar process occurs using DNA, however, no complementary DNA is needed, but two versions of probes corresponding to different alleles, as well as fluorescent labels of two different colors are used to differentiate the genotype of a given genomic locus.

Figure 2 shows an illustration of the data generation process for expression microarrays, starting with RNA extraction and isolation from cells or tissues, followed by reverse transcription and fluorescence labelling, and microarray hybridization and scanning of the signals.

More recently with the lowering cost of NGS technologies, complete sequencing of the genome or transcriptome through the parallelized generation of short sequence reads has been increasingly adopted. We will focus on describing RNA-seq technology in the context of transcriptomics as it is of relevance to this thesis. RNA-seq offers more comprehensive and accurate quantification of the transcriptome compared to expression microarrays by sequencing of novel and low expression transcripts without needing knowledge of the predefined set of sequences of interest. The process involves extraction of total RNA, isolation of specific RNA types (e.g., poly(A) selection for mRNA sequences, and ribosomal RNA depletion), reverse transcription into cDNA, construction of the sequencing library, PCR amplification, and

sequencing. Variations in RNA-seq such as small RNA-seq have been developed for more optimized targeting of miRNAs using size selection techniques. Furthermore, RNA-seq typically makes use of existing DNA-based NGS technologies, rather than direct sequencing of RNA, through the conversion to cDNA. The sequencing library is constructed through fragmentation of the sequences and addition of sequencing adapters. A step of polymerase chain reaction (PCR) amplification of the sequencing library is used to ensure sufficient material for sequencing.

Figure 3 illustrates the general process of RNA-seq data generation as described above, starting with RNA extraction and isolation from cells or tissues, followed by reverse transcription and library construction, PCR amplification, and finally sequencing using a NGS platform.

The majority of the work in this thesis focuses on data generated through microarray technologies, with the exception of Chapter 4, which utilized NGS for the generation of miRNA data.

Omics data preprocessing

Raw omics data requires quality control and preprocessing before they can be used in downstream analyses. For data generated through microarray technology, starting from raw fluorescence intensities, background correction is needed to separate the background noise from the true underlying biological signal. The specific methodology depends on the manufacturer of the microarray, however, the general process involves the deconvolution of observed intensities from localized background intensity around each probe, or from negative control probes intensities to obtain an estimate of the true signal intensity. After background correction, signals across multiple microarray samples need to be normalized to correct for systematic technical

sources of biases (e.g., variations in sample preparation, hybridization, scanning, etc.) which are unrelated to the underlying biological signal. For genotyping microarrays, an additional step of genotyping is performed in order to convert the intensity values into genotype calls.

In the case of RNA-seq data, after the generation of raw sequence reads, a process involving filtering of raw sequence reads for adapter sequences and read quality is performed prior to alignment of filtered reads to a reference genome using specialized alignment tools (e.g., BWA, STAR, etc.), followed by an estimation of transcript counts or abundance levels. Similar to microarray samples, normalization of RNA-seq transcript counts is necessary to ensure that different samples are comparable. However, in the case of RNA-seq, normalization methods need to account for library size (i.e., total number of reads obtained) and library composition (i.e., relative abundances of different RNA types). Multiple RNA-seq normalization methods are available each with specific assumptions by which they normalize samples [85].

Omics databases

The work in this thesis makes use of several omics data sources. Individual omics samples were obtained from three sources: Database of Genotypes and Phenotypes (dbGaP) [86], Gene Expression Omnibus (GEO) [87], and the United Kingdom Biobank (UKBB) [88]. Knowledge-based omics databases consisting of domain knowledge related to omics were also used, including The Molecular Signatures Database (MSigDB) [89], and The Pharmacogenomics Knowledgebase (PharmGKB) [90]. Chapters 2 and 3 leveraged data from dbGaP and GEO, while Chapter 5 focused on the analysis of data from the UKBB.

dbGaP is a National Institutes of Health (NIH) sponsored repository consisting of datasets with individual-level data from studies examining the relationship between genotype (as

well as other omics assay data such as gene expressions) and phenotype in humans. GEO is a similar repository, however, the primary type of data are gene expressions and other functional omics data, covers multiple organisms. The UKBB is a population-scale biobank consisting of genomic data (i.e., genotype microarray and NGS data), as well as in-depth health information from over half a million individuals.

MSigDB is a database that provides a collection of comprehensive gene sets representing various biological functions and processes for the purpose of enabling a better understanding of groups of genes. PharmGKB is a curated knowledgebase specific to pharmacogenomics including associations between drugs, genes, genetic variants, and phenotypes.

Statistical analysis methods of omics data

Numerous statistical techniques and methods have been developed for the analysis of omics data, however, we will cover the techniques and methods which are relevant to this thesis, including differential gene expression analysis, covariate adjustment, multiple testing corrections, gene set analysis, and imputation.

The goal of differential gene expression analysis is to determine which genes have significantly upregulated or downregulated levels of expression between different conditions (e.g., health vs. disease). Basic methods such as t-tests can be used to compare the means of expression levels between two groups. More advanced methods such as *limma* (linear models for microarray data) has been introduced to leverage information borrowing techniques to obtain better statistical estimates in small sample sizes [11].

Complementary to differential gene expression analysis which focuses on individual genes, gene set analysis is used to identify whether groups or pathways containing functionally

related genes (e.g., molecular functions, biological processes, or cellular components) show statistically significant differences between conditions, which could enable a better understanding of the biological mechanisms associated with conditions being studied.

Covariate adjustment is needed to control confounders when performing statistical estimation. Confounders are variables which are related to both the independent variable and the dependent variable. In the context of omics data analysis, covariate adjustment ensures that an observed association between omics variables and a dependent variable are correctly attributed and not driven by the confounders. The application of this technique is especially relevant to Chapters 2 and 3 when analyzing mRNA expression data.

Given that statistical tests are performed for tens of thousands of variables in high-dimensional omics data, the probability of observing at least one false positive result is inflated. Multiple testing correction is needed to adjust for the chance of false positives. Two commonly used correction methods are the Bonferroni correction [91], and the Benjamini-Hochberg procedure [92]. Bonferroni correction focuses on controlling the probability of making one or more false discoveries among all performed tests by adjusting the significance threshold by dividing by the number of tests performed. However, the Bonferroni method can be overly conservative, and lead to issues with inflated false negatives. The Benjamini-Hochberg procedure performs a more lenient adjustment by allowing for a reasonable proportion of false discoveries (i.e., false-discovery rate (FDR)) to be present among the significant results. We have applied multiple testing correction in several contexts with repeated hypothesis testing in Chapters 2-4.

Lastly, imputation refers to a technique used to estimate unobserved values in a dataset. In the context of genomics, imputation is often applied to genotype data to increase the

resolution of genotyping by estimating the genotypes at unobserved loci of interest, as genotyping microarrays cannot capture the full set of variations in a genome. The process leverages haplotypes, which are specific combinations of alleles often inherited together due to linkage disequilibrium (i.e., the non-random association of alleles at different genetic loci) to infer the most likely unobserved genotypes given the observed data. Imputation is a technique that has been applied to the genetic data from the UKBB analyzed in Chapter 5.

Machine learning

A brief introduction to machine learning theory, with specific on supervised learning concepts is provided to help contextualize the ML methodology used in this thesis.

In supervised learning, an algorithm is used to learn from a labelled dataset of elements consisting of features (X), and corresponding labels y (i.e., dependent variable) for each element. The elements of a dataset are typically required to be generated independently and sampled from the same underlying distribution (i.e., independent, and identically distributed – "IID" assumption).

The goal is to identify a mapping function f (i.e., model) that represents the relationship between X and y, such that $f(x_i)$ can accurately predict y_i, where (x_i, y_i) is a previously unseen pair of features and label. Two common tasks in supervised learning are regression and classification. In regression problems, the labels to be predicted are on a continuous scale, whereas the labels are discrete for classification problems. In general, a process involving data preprocessing, model development, and model evaluation are necessary for supervised learning.

Preprocessing of features is needed to prepare the data prior to analysis using supervised ML. A process of feature engineering (including feature selection and extraction) is often used to

transform or combine raw features based on domain understanding with the aim of obtaining more meaningful feature representations. Some commonly used techniques include feature scaling and normalization to standardize input features to reduce effects of outliers and noise. Alternative representations of features are sometimes useful for categorical features (e.g., creating binary features for each category) or using distributed representations [93]. Feature aggregation (e.g., combining multiple features) is also used to reduce the number of possibly redundant features. More complex methods such as principal component analysis (PCA) and genetic algorithms are utilized to discover latent variables and reduce dimensionality [94]. Feature engineering is a difficult process and typically requires in-depth domain knowledge and can lead to bias (e.g., placing higher importance on specific features) and loss of information if not properly performed. However, more recent ML methods such as deep neural networks can often automate the process of feature engineering by automatically extracting meaningful feature representations as part of the training process without the need for prior domain expertise [95].

Next, dataset splitting is performed to divide a dataset into separate training, validation, and test sets. The training set facilitates the process of creating a model that captures underlying patterns for making accurate predictions. Concurrently, the validation set is used in evaluating potential overfitting of a trained model in a method known as cross-validation, which is used to inform model selection. Of importance, overfitting refers to a scenario where a model fits the training data too closely, losing the ability to generalize to new, previously unseen data. The test set is used to evaluate the generalization performance of a model selected based on the training and validation data in previously unseen data points, thus providing a less biased assessment of the model's predictive capability.

The model development process requires selecting a ML algorithm based on the specific problem being addressed and characteristics of the dataset. ML algorithms are used to create models, which can be classified into two broad categories: parametric and non-parametric [96]. Parametric models are characterized by a fixed number of parameters. Examples of parametric models include logistic regression, support vector machines, and neural networks. Typically, training a parametric model involves finding the optimal parameters of the model which results in the optimal prediction of labels given input data. Non-parametric models are not defined by a fixed set of parameters. An example is the decision trees algorithm [97], for which the depth and complexity of a tree can vary during the model construction process.

The process of training prediction models requires defining a cost (or loss) function that encapsulates the difference between observed model predictions and the actual labels, and identifying parameters which minimize the cost function using mathematical procedures or optimization algorithms.

In addition to model parameters, most ML algorithms also have hyperparameters which need to be defined prior to the training process. Hyperparameters can have a major influence on a resulting model's performance. Hyperparameters differ from the model parameters which are updated based on the data. For example, hyperparameters are typically used to define the settings of the ML algorithm such as the model architecture (e.g., depths of the tree for a decision tree model) or the learning rate during parameter updates by the optimization algorithm. There is no general rule for choosing the set of hyperparameters for an algorithm, however, hyperparameter search such as grid search, random search, and more advanced and efficient methods such as Optuna [98], can be used to find combinations of hyperparameters that result in more optimal model performance.

After a model has been selected from the model development process, evaluation is performed using the test set which has not been previously used to obtain a less biased estimate of model performance. For regression models, metrics such as mean squared error and proportion of variance explained are often used to evaluate the performance of a model. For classification models, it is often necessary to decide upon a probability threshold to enable the classification of samples based on predicted probabilities prior to calculation of evaluation metrics. Simple accuracy is a commonly used metric under scenarios with balanced classes. In situations where class imbalance is present, accuracy has bias and metrics such as balanced accuracy is necessary. However, a more comprehensive way to understand classification performance is through a confusion (or error) matrix consisting of a quadrant which captures the True Positives (cases where the model correctly predicted the positive class; TP), True Negatives (cases where the model correctly predicted the negative class; TN), False Positives (cases where the model incorrectly predicted the positive class; FP), and False Negatives (cases where the model incorrectly predicted the negative class; FN).

Metrics including precision, recall, and F1 score can be derived from values of the confusion matrix. Precision measures the accuracy of positive predictions, and is calculated as the ratio of TP to the sum of TP and FP. Recall measures how well positive instances are identified, and is calculated as the ratio of TP to the sum of TP and FN. Precision may be prioritized in situations where misclassification of negative cases as positive (i.e., FP) incurs a high cost, whereas recall may be prioritized in situations where capturing all positive cases (i.e., minimizing FN) despite misclassifying some negative cases is important. The F1 score is the harmonic mean of the precision and recall emphasizing poor performance in either metric using a single summary score.

The area under the receiver-operator characteristics (ROC) curve (AUC) metric is another often used metric in assessing classification performance as the likelihood of positive cases ranking higher than negative cases (i.e., the ability of the model to distinguish between positive and negative cases). Unlike previously mentioned metrics, the AUC metric does not depend on a classification threshold. The AUC metric is calculated by plotting the ROC curve and taking the area under the curve. The ROC curve shows the TP rate (i.e., proportion of items in the positive class correctly predicted as positive) against the FP rate (i.e., proportion of items in the negative class falsely predicted as positive) at all possible classification thresholds (i.e., probability threshold to divide samples into two classes) thus providing a more comprehensive view of the model performance without requiring a fixed decision threshold.

Given the numerous decisions and choices to be made in model development, there has been a movement to develop AutoML (Automated Machine Learning), which are methods that simplify the model development process by framing it as a search problem for identifying optimal combinations of feature engineering, ML algorithm, and hyperparameter tuning in a fully automated manner [99].

Figure 4 illustrates the basic process of supervised learning described above starting with a labelled dataset, performing feature engineering, dataset splitting, model development and selection, to final model evaluation.

Summary of key machine learning algorithms

Several ML algorithms including logistic regression, XGBoost, neural networks, and graph convolution networks are used in this thesis. A brief overview of each algorithm will be provided.

Logistic regression is a fundamental method used for binary (i.e. positive vs. negative class) classification tasks. The model is defined by a logistic (i.e., sigmoid) function of a linear combination of predictor variables (X) and model parameters (β). The output of the model is in the range of 0 and 1, which can be interpreted as the probability of the positive class:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-z}} \text{ where } z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

The parameters of the logistic regression model can be estimated using maximum likelihood estimation, however, numerical optimization methods would be necessary in more complex models. Logistic regression has the advantage of being simplistic (e.g., no tuning of hyperparameters required) and highly interpretable, yet, due to its simplicity, it is relatively limited in capability for modelling data with complex relationships, heterogeneity, interactions, and high dimensionality.

XGBoost (Extreme Gradient Boosting), is an optimized implementation of an ensemble technique called gradient boosting [100]. Gradient boosting theorizes that weak learners, such as individual decision trees, when combined, can form a strong predictive model. In gradient boosting, each new model is trained to correct the residuals (i.e., prediction errors) made by the current ensemble of previous models. Specifically, the direction and magnitude of corrections are guided by the negative gradient of a chosen loss function in the iterative optimization process. By using the gradient in this manner, the algorithm ensures that each new model's adjustments

are oriented towards minimizing the loss function. The main hyperparameters of XGBoost are the number of estimators (i.e., the number of individual decision trees to build), max depth of each tree, gamma (i.e., a regularization parameter that controls the complexity of the trees), and learning rate, which controls the fraction of residuals (prediction errors) to correct at each boosting iteration. The use of ensembles of individual decision trees gives the algorithm the advantage of being highly flexible to model complex relationships and interactions. The optimizations of the XGBoost implementation have the additional advantage of enabling rapid construction of these models.

Neural networks refer to a class of methods composed of interconnected layers of nodes called artificial neurons. Inputs to a neural network are summarized and modelled through iterative transformations within successive layers of the neural network model. At each layer, each neuron takes a weighted sum of its inputs added to a bias term and passes the sum through a non-linear function called the activation function. The weights and biases are parameters of the model to be optimized during model training. Neural networks are flexible and can model complex interactions and non-linear relationships in data, however, they typically require large amounts of data to train, and can be less interpretable. Various classes of neural networks have been developed over the years including convolution neural networks which introduced the convolution operator to enable more effective feature extraction for image data [101], as well as recurrent neural networks for modelling sequence data [102]. The hyperparameters of neural networks often include model architecture (e.g., number of layers, number of nodes of each layer), choice of non-linear activation function, and learning rate.

Graph convolutional networks (GCNs) are a class of neural networks designed for modelling graph (network)-structured data [61]. Graphs are used to represent dependencies

between entities in data using nodes and edges between nodes in a way that is not possible with Euclidean (tabular) representations of data. For example, graphs can be used to describe regular structured data including images, language, but also more irregular structures such as social networks and molecular structure. The generalized convolution operator introduced by the GCNs enables the learning of a low-dimensional Euclidean representation of each node in a given graph that captures the features of the node as well as preserves the surrounding graph neighborhood of that node, which can be leveraged in tasks including node classification, link prediction, and graph-level prediction tasks [103].

Knowledge graphs (KGs) were originally introduced to enhance the searching of information on the web [104]. KGs can be understood as a special use case of a graph to represent multiple entities and semantic connections between them as a way to capture and store knowledge. Learning from KGs with the aim of capturing semantic information of entities within KGs has been explored in depth [105]. GCNs could be combined with KGs by capturing low-dimensional representations of concepts in the knowledge graph for use in ML modelling tasks [106]. Furthermore, biomedical KGs are a domain specific type of knowledge graph that represents biomedical relationships between entities such as genetic variants, genes, diseases, and drugs as nodes, and their relationships as the connections. Biomedical KGs can be used for a variety of precision medicine uses, including drug repurposing, comorbid risk prediction, disease risk prediction, and patient subtyping [107].

Figures

Figure 1. Major omics categories and their relationships.

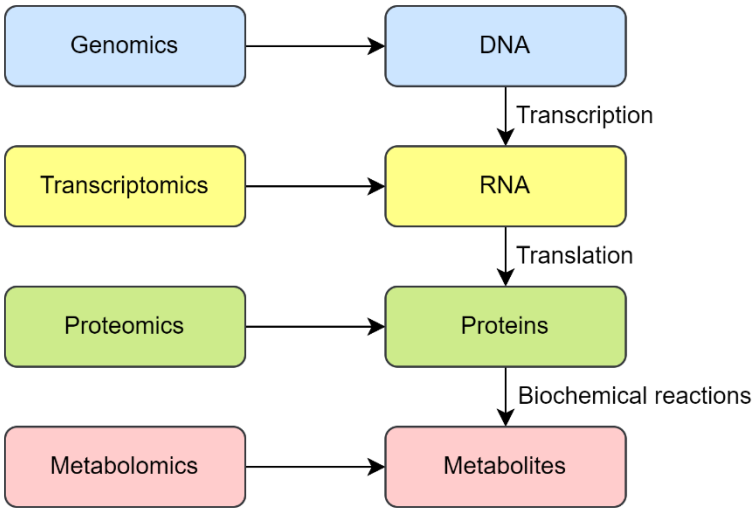


Figure 1. The relationships between the major omics categories are illustrated, starting from genomics level, which quantifies DNA data, down to the functional genomics levels including transcriptomics, which quantifies RNA data produced from the transcription of DNA, proteomics, which quantifies proteins produced from translation of RNA, and lastly, metabolites which quantifies the products of biochemical reactions.

Figure 2. Expression microarray data generation.

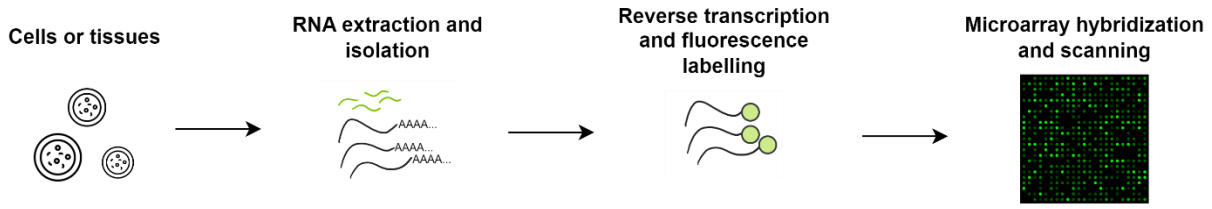


Figure 2. The process of expression microarray data generation is illustrated. The process begins with extracting RNA from cells or tissues followed by reverse transcription and fluorescence labelling, and microarray hybridization and scanning of the signals.

Figure 3. RNA-seq data generation.

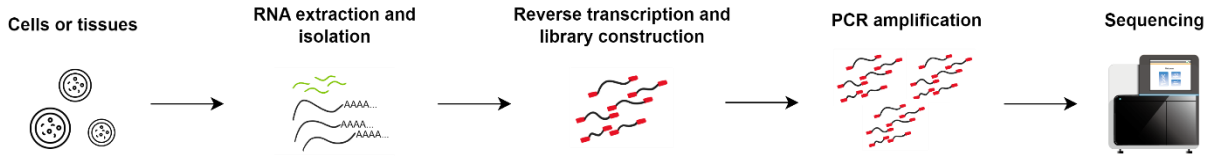


Figure 3. The process of RNA-seq data generation is illustrated. The process begins with extracting RNA from cells or tissues followed by reverse transcription and library constructions. PCR amplification of the library is performed prior to sequencing using an NGS platform.

Figure 4. Basic process of supervised learning.

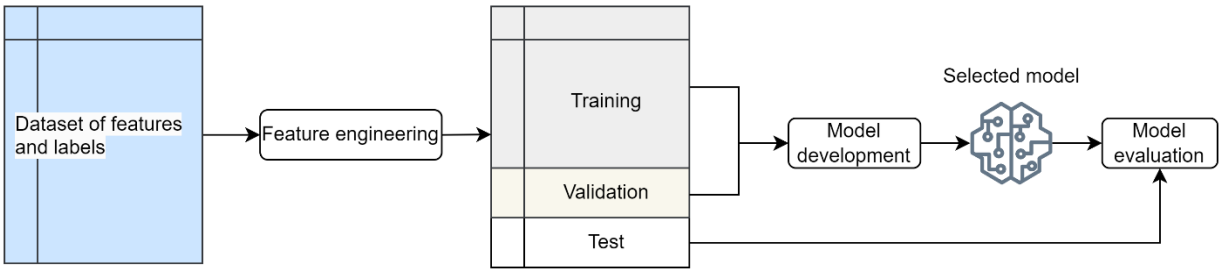


Figure 4. The basic process of supervised learning is illustrated. The basic process starts with a labelled dataset, followed by feature engineering, dataset splitting (into training, validation, and test sets), model development and selection using the training and validation sets, and final model evaluation using the holdout test set.

Chapter 2. Transcriptomics and machine learning to advance schizophrenia genetics: a case-control study using post-mortem brain data

Author names

Bill Qi[1], Sonia Boscenco[2], Janani Ramamurthy[2], Yannis J. Trakadis[1,3]

Author institutional affiliations

[1] Department of Human Genetics, McGill University, Montreal, QC, Canada

[2] Faculty of Science, McGill University, Montreal, QC, Canada

[3] Department of Medical Genetics, McGill University Health Center, Montreal, QC, Canada

Corresponding author

Yannis J. Trakadis, MD MSc FRCPC FCCMG

Medical Geneticist & Metabolics Specialist

Assistant Professor, Human Genetics

McGill University Health Centre

Room A04.3140, Montreal Children's Hospital

1001 Boul. Décarie, Montreal, Quebec, Canada, H4A 3J1

Tel: (514) 412-4427, Fax: (514) 412-4296

Email**:** yannis.trakadis@mcgill.ca

Note: *while the intention was to display figures and tables at the end of each chapter to maintain consistency with the rest of the thesis, it should be noted that Chapter 2 corresponds to a previously published manuscript, necessitating adherence to the original format.*

Abstract

Background and Objective: Alterations of the expression of a variety of genes have been reported in patients with schizophrenia (SCZ). Moreover, machine learning (ML) analysis of gene expression microarray data has shown promising preliminary results in the study of SCZ. Our objective was to evaluate the performance of ML in classifying SCZ cases and controls based on gene expression microarray data from the dorsolateral prefrontal cortex.

Methods: We apply a state-of-the-art ML algorithm (XGBoost) to train and evaluate a classification model using 201 SCZ cases and 278 controls. We utilized 10-fold cross-validation for model selection, and a held-out testing set to evaluate the model. The performance metric utilizes to evaluate classification performance was the area under the receiver-operator characteristics curve (AUC).

Results: We report an average AUC on 10-fold cross-validation of 0.76 and an AUC of 0.76 on testing data, not used during training. Analysis of the rolling balanced classification accuracy from high to low prediction confidence levels showed that the most certain subset of predictions ranged between 80-90%. The ML model utilized 182 gene expression probes. Further improvement to classification performance was observed when applying an automated ML strategy on the 182 features, which achieved an AUC of 0.79 on the same testing data. We found literature evidence linking all of the top ten ML ranked genes to SCZ. Furthermore, we leveraged information from the full set of microarray gene expressions available via univariate differential gene expression analysis. We then prioritized differentially expressed gene sets using the *piano* gene set analysis package. We augmented the ranking of the prioritized gene sets with genes from the complex multivariate ML model using hypergeometric tests to identify more robust gene sets. We identified two significant Gene Ontology molecular function gene sets:

"*oxidoreductase activity, acting on the CH-NH2 group of donors"* and "*integrin binding.*" Lastly, we present candidate treatments for SCZ based on findings from our study

Conclusions: Overall, we observed above-chance performance from ML classification of SCZ cases and controls based on brain gene expression microarray data, and found that ML analysis of gene expressions could further our understanding of the pathophysiology of SCZ and help identify novel treatments.

Keywords

Schizophrenia, Transcriptomics, Machine learning, Bioinformatics, Post-mortem

1. Introduction

The point prevalence of schizophrenia (SCZ) in Western societies is estimated to be 5 per 1000 with disablement of up to 70% [1]. In 2013, a study found that the total estimated cost of SCZ was $155.7 billion [2]. This includes costs for prescription medication, hospitalization, diagnosis, and long-term care, but also indirect costs, such as increased unemployment, decreased workplace productivity, and premature mortality.

Genetics plays a major role in the etiopathology of SCZ. The heritable component to SCZ has been estimated to be 80% from twin studies [3], however, deciphering the genetics of SCZ has been challenging. In the largest genome-wide association study on SCZ to date, 108 genetic loci were found to be associated with SCZ based on common variants [4]. In addition to common variants, rare, de novo, and structural variations have also been implicated in SCZ [5]. Overall, current findings in the literature support a complex polygenic pathophysiology for SCZ. However, polygenic risk score approaches, which are additive models, have limited effectiveness in predicting disease status [6]. Alterations of the expression of a variety of genes have been documented in SCZ, suggesting that gene expression microarrays can be useful for biomarker discovery. For example, one study analyzed gene expression in the blood of 32 untreated patients with newly diagnosed SCZ using Affymetrix microarrays [7]. Significantly altered expression of 180 genes was found when compared to healthy controls. In particular, the authors found that DAAM2 gene expression levels returned to control levels in patients who were in remission following their first episode of psychosis. This suggested that DAAM2 may be a biomarker for SCZ.

Machine learning (ML) refers to a promising collection of methods which can address the complexity of large high-dimensional data. ML comes in two varieties: supervised and

unsupervised ML. The focus in this paper is on supervised ML, which aims to make predictions of a specific outcome [8]. Labeled data is used as input, then the supervised algorithm is trained to produce outputs which accurately reflect the labels. In medicine, applications include risk estimation, radiology report classification, complex disorder prediction, and disease classification [8]. Within supervised ML, there exist both classification and regression algorithms [9]. Regression outputs are continuous, whereas classification outputs are discrete values.

Supervised ML algorithms have been used to analyze microarray gene expression data in SCZ. For example, Takahashi et al. (2010) used bioinformatics to explore if whole blood cell gene expression can be used as a biomarker for SCZ. Unpaired t-tests of gene expression data sets from 52 untreated SCZ patients and 49 normal controls identified 792 differentially expressed gene probes. After subdivision of the samples into training and testing sets, quality filtering and stepwise forward selection identified 14 probes as predictors of the diagnosis. Artificial neural networks (ANNs) were trained with the selected probes as the features. In the training set, 91.2% diagnostic accuracy was achieved, with 87.9% for the hold-out testing set [10]. Another study applied ML to gene expression microarray data from skin fibroblasts and post-mortem brain tissue samples and was able to achieve average AUC scores over 0.9 on 4-folds of cross-validation; however, the sample size was small: the skin fibroblast dataset consisted of 20 SCZ patients and 20 healthy controls, while the post-mortem brain tissue samples consisted of 23 SCZ patients and 19 healthy controls [11]. Another study collected blood-based microarray gene expression data of a total of 152 SCZ patients and 138 controls from 4 different datasets and developed a classifier with nearly 100% classification accuracy over 10-folds of

cross-validation [12]. These studies, albeit not perfect, suggest that ML analysis of microarray gene expression data can lead to clinically useful biomarkers in SCZ.

In this study, we use a methodology that aims to address the challenges identified in the previous ML studies using mRNA data. We apply a state-of-the-art supervised ML algorithm to gene expression microarray data from the dorsolateral prefrontal cortex (DLPFC) of post-mortem SCZ patients and controls. DLPFC is an area involved in executive functions. Chechko, Cieslik [13] showed differential functional connectivity patterns in regions of the DLPFC with other brain regions between SCZ patients and controls using resting-state magnetic resonance imaging. Furthermore, there is also support that the transcriptome of neurons in the DLPFC is altered in SCZ patients [14]. We contrast our methodology and findings to those of the previous studies and highlight potential biomarkers. Through bioinformatic analyses of our findings, we try to further our understanding of the pathophysiology of SCZ and identify novel candidate treatments.

2. Material and methods

2.1 Schizophrenia gene expression microarray dataset

We obtained a gene expression dataset of adult post-mortem patients with SCZ (n=201) and control subjects (n=278) from dbGaP (dbGaP Study Accession: phs000979.v1.p1). The gene expression data were obtained from the DLPFC, using the Illumina HumanHT-12 v4 Expression BeadChip platform. The dataset was then background corrected with normexp and quantile normalized using the "neqc" function from the R *limma* package (version 3.42.0).

2.2 Algorithm selection

Many powerful ML algorithms render themselves uninterpretable, making it difficult to understand their decision-making process. For our ML analysis of the data, we decided to use a state-of-the-art yet interpretable regularized gradient boosted machines (GBM) approach (XGBoost implementation, [15]). Even though regularized GBM is still a complex algorithm to interpret, it is state of the art and has been proved successful in a wide range of tasks, as illustrated in a recent study from our group (Trakadis et al., 2018). Its highly regularized built-in feature selection and reduction characteristic and ability to rank features based on their relative importance to its decision process made it a great candidate for our study. Of note, a regularized algorithm penalizes itself for complexity, and thus uses only features that are relevant and brings the most intelligence to its architecture. In our study, this means selecting only transcriptomic features that have high predictive power and discarding the less informative ones.

Specifically, XGBoost (Extreme Gradient Boosting), is a method of learning an ensemble of K classification and regression trees (CARTs), where each additional tree (f_k) is selected from the set of all possible CARTs ($\mathcal{F}$) and added to correct errors of the previous learning iteration. The input features for a sample is x_i, and the corresponding model prediction is $\hat{y}_i$,

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F}$$

CARTs differ from parametric methods such as linear and logistic regression in that it is a non-parametric method where data is recursively partitioned into smaller subsets of data to form a tree structure [16]. Using a combination of CARTs gives the ability to model non-linear combinations of data, as seen in random forest algorithms [17], which merge randomly generated

decision trees into a single "learner" [18]. The predictor space of these trees is created by division of all the possible values into distinct and non-overlapping regions.

XGBoost is similar to the random forest algorithm, however, the process of adding trees to the learner in XGBoost is an iterative process of improving the previous learner at each iteration of training, as illustrated below:

$$\hat{y}_i^{(0)} = 0$$

$$\hat{y}_i^{(1)} = \hat{y}_i^{(0)} + f_1(x_i)$$

$$\hat{y}_i^{(2)} = \hat{y}_i^{(1)} + f_2(x_i)$$

$$\dots$$

$$\hat{y}_i^{(K)} = \hat{y}_i^{(K-1)} + f_K(x_i)$$

The objective function for learning a binary classification XGBoost model is defined as the sum of the logistic loss (l), and a regularization term (Ω), with T being the number of leaves in a tree, w the leave weights, and γ and λ as hyperparameters.

$$\text{obj}(\phi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k)$$

$$\text{where } l(y_i, \hat{y}_i) = \ln\left(1 + e^{-\hat{y}_i}\right) + (1 - y_i)\ln(1 + e^{\hat{y}_i})$$

$$\text{and } \Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_t w_t^2$$

2.3 ML analysis

We randomly sampled 80% of the full dataset to be the training set, to identify the best classification model, and 20% to be the testing set. For model selection, we used 10-fold cross-validation combined with randomized hyperparameter search for 2500 iterations (i.e., training 2500 different models). The performance of each trained model is defined by the area under the receiver-operating-characteristic (ROC) curve (AUC), with cases being the positive class, averaged over all ten cross-validation folds.

We repeated the above procedure to select the best baseline model (i.e., a model trained using the same cross-validation approach but with the labels randomly permuted). Given the large number of features in the microarray dataset, it is important to control for overfitting with a baseline model. The trained model would need to perform significantly better than the baseline model in order to rule out the fact that the algorithm is overfitting to the data with random features. A one-sided Wilcoxon signed-rank test is used to compare the AUC values on each cross-validation fold between the best baseline and best trained model for significance.

Lastly, the hyperparameters of the ML algorithm (e.g., number of boosting iterations, max-depth of trees, learning rate, etc.) from the trained model and the baseline model with the best average AUC from 10-fold cross-validation are extracted. Using the extracted hyperparameters (which provided the best training conditions), the models are retrained on the full *training* set (without cross-validation to maximize sample size) to improve its performance before being evaluated on a *holdout testing* set (i.e., the data which was not used during the training phase). An overview of the complete ML analysis pipeline is shown in Figure 1.

2.4 Model evaluation

We trained binary classification models using the above-described approach to discriminate SCZ cases from controls. To assess the best trained model, we calculated and plotted the ROC curve based on the *testing* set. In order to calculate the classification accuracy, the predicted class probabilities of each testing set sample need to be converted to a discrete case or control classification. A high probability means the sample is more likely to be a "case," and a low probability means the sample is more likely to be a "control." An optimal probability cutoff threshold is needed in order to split the samples into the discrete classes. To determine the optimal cutoff threshold, we averaged the best cutoff values derived from the ROC curves from each cross-validation fold during training. The best cutoff is defined as the probability threshold dividing the cases and controls classes which maximizes the number of true positive classifications and minimizes the number of false-positive classifications (i.e., maximizing the area under the ROC curve). After a discrete class was assigned to the testing set samples, we calculated an overall balanced accuracy metric since our testing set is not balanced.

Furthermore, to provide a more detailed interpretation of the model performance, we looked at the balanced classification accuracy from high to low prediction confidence levels (i.e., deviation from the optimal cutoff threshold, where a larger deviation means higher confidence). The following technique was used to generate a plot of rolling (high to low) balanced accuracy values for the testing set. First, predictions are sorted from the highest confidence to lowest. Then, starting with a window of the top ten most confident predictions, a balanced accuracy is calculated. Then the window shifts down by one, and the process repeats until the end of the list is reached. A graph is then generated to visualize the trend of balanced accuracy from highest to lowest prediction confidence.

Lastly, we leverage recent advancements in automated ML (AutoML) to see whether the classification performance can be improved based on the top gene features prioritized from our training pipeline. The tool we choose for this analysis is Auto-Sklearn 2.0 [19], which utilizes Bayesian optimization, meta-learning, and ensemble selection to search for the best preprocessing, estimator, and hyperparameter configurations and produce a final ensemble model by combining multiple models. The total search procedure was constrained to 3600 seconds (1 hour), with the maximum run time of each ML pipeline limited to 600 seconds (10 minutes).

2.5 Gene set analysis

Gene probes mapping to the same gene were combined by averaging, and their average expression value was used as the set of gene expressions for this analysis. Differential expression statistics were obtained for each gene through the R *limma* package (version 3.42.0) using the full set of 201 SCZ cases and 278 controls. We then performed gene set analysis (GSA), which tests for altered expression for groups of genes (gene sets) between two classes (i.e., cases vs. controls). A "gene set" can represent a group of genes with a similar function or activity, or a group of genes belonging to the same biological process or pathway. We obtained the Gene Ontology (GO) molecular function gene sets [20, 21], from MSigDB (version 7.0) [22], as the source for "gene sets." The GO molecular functions gene set group genes based on related activities performed by single or multiple gene products. GSA was performed using the R *piano* package (version 2.2.0) [23]. We applied the consensus ranking method from the *piano* tool by combining gene set significance results from all available GSA methods from *piano* (section 4.3 from [24]). Lastly, any gene sets with a consensus ranking above 10 in any of the five *piano* directionality classes (i.e., five specific ways the gene sets can be significantly altered, section 4.4.2 from [24]) were selected to be important, (i.e., different between cases and controls). A

heatmap of the important gene sets, along with the median p-value from all GSA methods, was plotted.

2.6 Gene set filtering with ML genes

As described above, the *piano* GSA method A identified GO molecular functions gene sets exhibiting different expression between cases and controls. We performed a secondary enrichment analysis to augment the ranking of the gene sets derived from *piano* based on the genes utilized by our best trained ML model (ML genes) mentioned above. More specifically, a hypergeometric test was performed to determine which "gene sets" were enriched among the ML genes used to separate SCZ cases from controls. We applied the *Benjamini–Hochberg* procedure to adjust the false-discovery rate (FDR) with alpha set to 0.2. An overview of the complete gene set analysis pipeline is shown in Figure 1.

2.7 Software

The ML analyses were implemented using Python (version 3.7.1), with the *xgboost* package for training the models (version 0.81). K-fold cross-validation and hyperparameter selection during training was implemented with the *scikit-learn* package (version 0.21.2). AutoML was implemented with the *auto-sklearn* package (version 0.13.0)

Figure 1. ML analysis and gene set analysis pipeline.

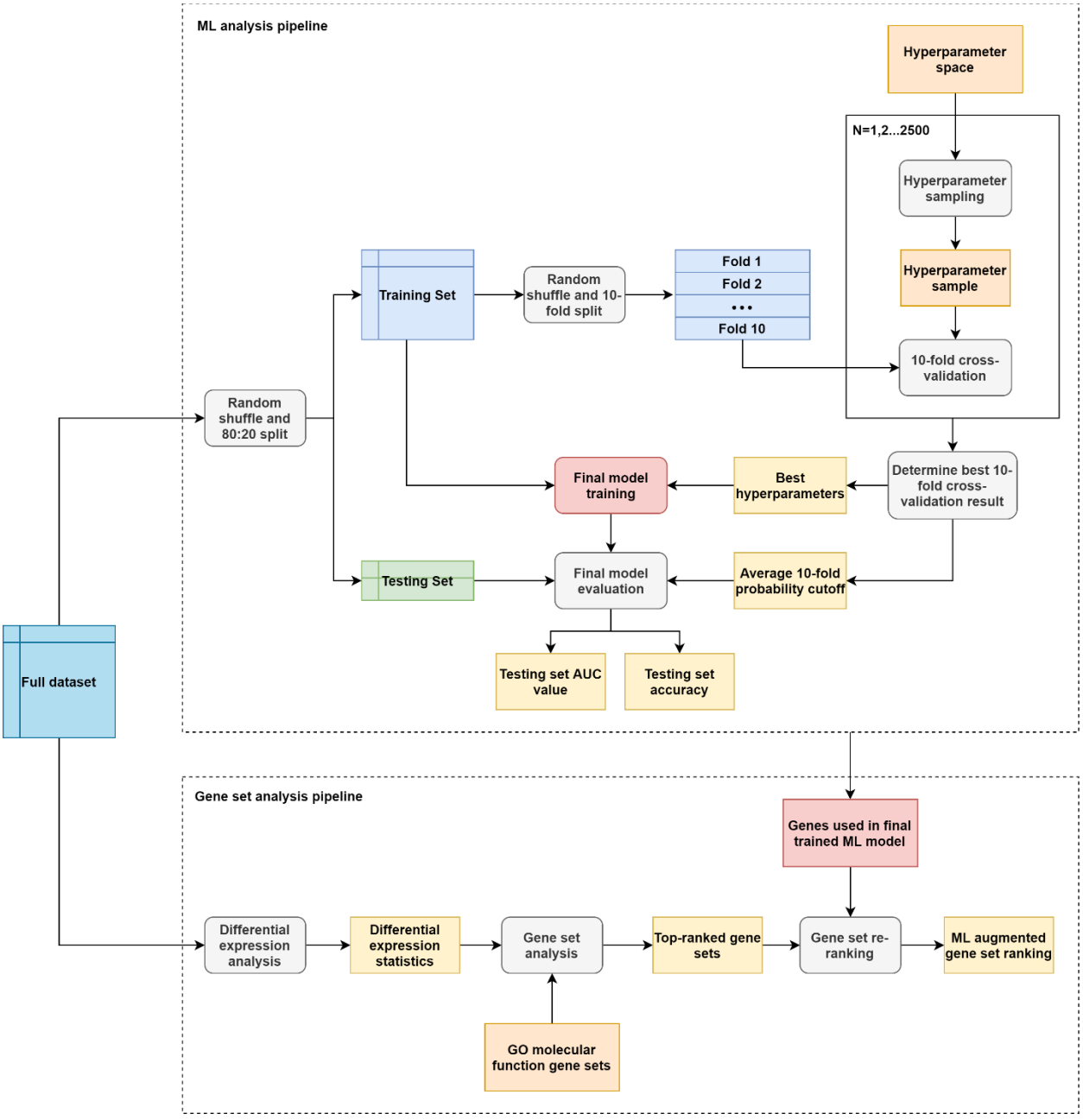


Figure 1. For the ML analysis (top section), the full original dataset was randomly shuffled and split into a training set (80%) and a testing set (20%). The training set is further shuffled and divided into 10 equal folds. For determining model hyperparameters, 2500 samples were sampled from the hyperparameter space, and 10-fold cross-validation was performed for each

sample. The best 10-fold cross-validation result was determined based on highest average validation AUC (area under the receiver operating characteristics curve) value over 10-folds. The corresponding hyperparameters (*Best hyperparameters*) were then used for final model training using the complete training set. Final model evaluation using the testing set was performed to obtain a testing set AUC value. Using a cutoff value between 0 and 1, a discrete classification is obtained by converting any predicted probability above the cutoff as a SCZ case, and any predicted probability below the cutoff as a control. The cutoff used is the *Average 10-fold probability cutoff*, which is obtained by taking the average of the probability cutoffs resulting in the optimal balanced accuracy for classification in each of the 10-folds during cross-validation of the best trained model. The final *Testing set accuracy*, is then obtained by calculating the balanced accuracy with the discretized classifications. To facilitate identification of biological pathways involved in SCZ, we performed a gene set analysis (bottom section). Differential expression analysis is first performed to obtain the *Differential expression statistics*. The statistics are then used to perform *Gene set analysis* to rank gene sets from the *GO molecular function gene sets*. Any gene set with a *piano* consensus ranking above 10 are taken as the *Top-ranked gene sets*, which are then re-ranked based on evidence of overlap with ML genes to produce the final *ML augmented gene set ranking*. A literature review was conducted for any gene sets with a false-discovery rate below 0.2.

3. Results

We obtained the best classification models for SCZ vs. controls based on our model selection procedure. The best model had an average AUC of 0.76 (SD: 0.050) over ten folds of cross-validation. In contrast, the best baseline model had an average AUC of 0.52 (SD: 0.095).

Comparison between the cross-validation AUCs of the trained and baseline models with the Wilcoxon signed-rank test showed that the performance of the trained model was significantly greater that than that of the baseline model (p-value: 0.003).

After retraining the best model based on the entire dataset, an evaluation was performed on the hold-out testing set. The retrained model achieved an AUC of 0.76 on the testing set (Figure 2A). After dividing the predictions into discrete classes based on the optimal cutoff threshold estimated from cross-validation, the actual classification results are shown in a confusion matrix (Figure 2B). After plotting the rolling balanced accuracy over testing samples ranked by prediction confidence, we observed that the most confident predictions reach a balanced accuracy of ~80-90% (Figure 2C). Overall, the best retrained classification model utilized 182 mRNA transcript features (Supplemental Table 1), with the mRNA transcripts corresponding to the *COPS3, HBB, DTNA, ITGB4, COX7A1, MAOB, SLC38A5, LBH, NODAL, GALNTL1* genes being the ten most significant features.

The 182 mRNA transcript features were selected and used in the AutoML analysis. Overall, 2402 ML pipeline runs were performed, with 2400 successful runs and 2 runs exceeding the time limit. The final ensemble model achieved an AUC of 0.79 on the testing set.

Figure 2. Testing set results for discriminating schizophrenia cases vs. controls.

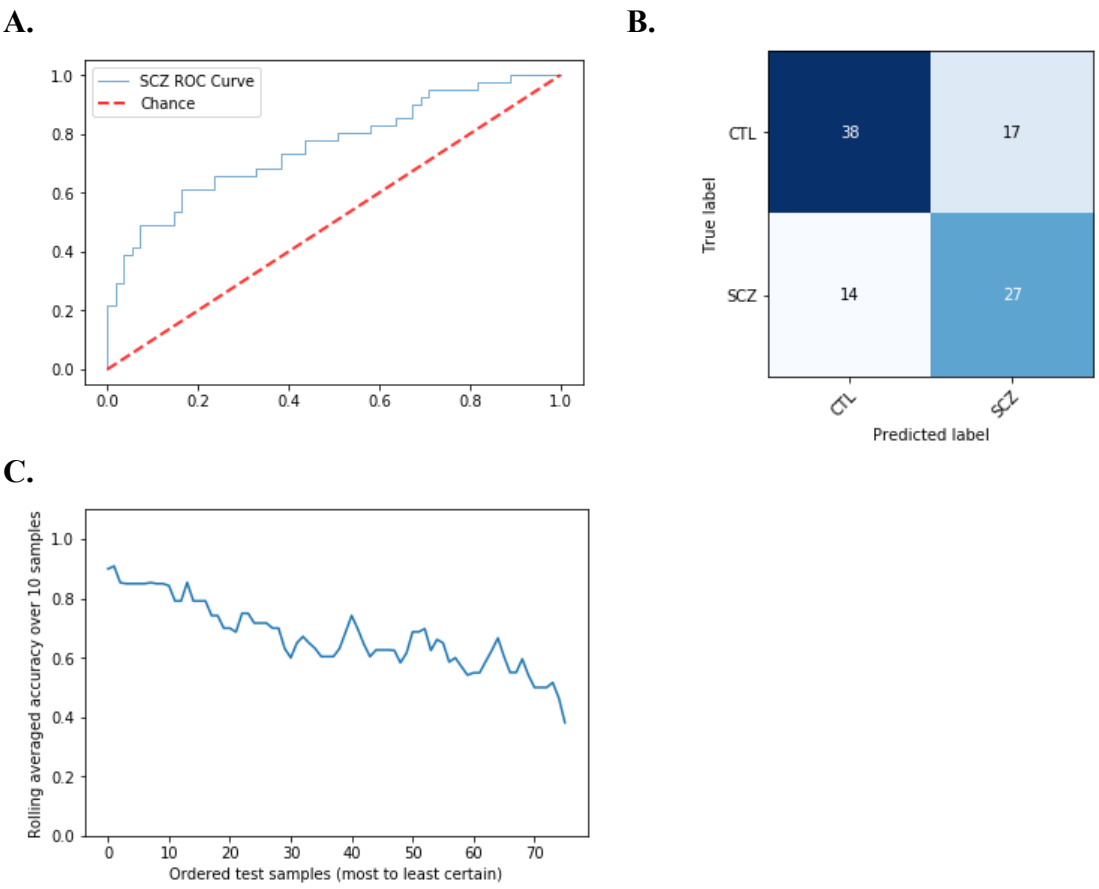


Figure 2. After obtaining the best trained classification model for discriminating between schizophrenia cases vs. controls, several metrics were used to assess the performance of the model on a previously unseen testing set (20% of the full dataset). Figure 2A shows a receiver-operating characteristics (ROC) curve for model predictions with the area under the curve (AUC) of 0.76, suggesting that the model performs well above random chance (red dashed line). The x-axis of the curve shows the false-positive rate (FPR) and the y-axis shows the true-positive rate (TPR) at each probability cutoff threshold. The ROC curve suggests that the model could distinguish cases and controls with low FPR (x-axis) and high TPR (y-axis). Figure 2B shows the actual classification results obtained based on the optimal probability cutoff threshold estimated

from 10-fold cross-validation during training. With the positive class as SCZ and negative class as CTL, the top-left, top-right, bottom-left, and bottom-right quadrants of Figure 2B show the number of true-negative (CTL predicted as CTL), false-positive (CTL predicted as SCZ), false-negative (SCZ predicted as CTL), and true-positive (SCZ predicted as SCZ) predictions, respectively. Figure 2C shows the prediction-confidence-ranked rolling balanced accuracy plot, which shows the trend of balanced classification accuracy of the most confident subset of predictions to the least confident subset of predictions (x-axis left to right) for the testing set samples. The y-axis shows the balanced accuracy value for each subset of predictions. The most-confident predictions generally have a balanced accuracy of ~80-90%.

GSA using *piano* prioritized 217 gene sets from a total of 1645 gene sets from *GO molecular functions* gene sets. We then determined which of the filtered "gene sets" were significantly enriched in our set of ML genes via the hypergeometric distribution p-values and FDR correction with alpha set to 0.2 (Supplemental Table 2). We found two significant GO terms meeting the cutoff: "*oxidoreductase activity, acting on the CH-NH2 group of donors"* and "*integrin binding*." Specifically, the genes *MAOB* and *GLDC* from the ML genes were in the "*oxidoreductase activity, acting on the CH-NH2 group of donors"* gene set, while *ITGB4*, *FERMT3*, *NISCH*, and *SRC were* ML genes part of the "*integrin binding*" gene set (Supplemental Table 2). Results from the *piano* heatmap show that the "*oxidoreductase activity, acting on the CH-NH2 group of donors*" gene set was ranked highly on the *mixed-directional up* class suggesting that a subset of genes in the gene set had significantly increased expression, while "*integrin binding*" was ranked highly in the *mixed-* and *distinct-directional up* classes, suggesting that expression of most genes in the gene set are significantly increased.

4. Discussion

Many prior studies, such as those mentioned in the introduction, have focused on the analysis of a few biomarkers or univariate analysis of multiple genes. A drawback of such strategies is that more complex connections between genes are difficult to identify. The benefit of the ML approach is that it can pick up on complex effects that might not be significant in univariate analyses. In this study, we leveraged a supervised ML algorithm in the analysis of microarray gene expression profiles from post-mortem DLPFC of SCZ cases and controls. We trained an ML classifier that can differentiate SCZ cases and controls (AUC of 0.76 over ten folds of cross-validation) significantly better than random chance (p-value: 0.003). The difference in performance for the trained and baseline model suggests that the trained ML algorithm was able to pick up on disease-related signals and not based on noise or inherent structure in the high-dimensional dataset. To this end, the ML classifier was able to generalize to a testing set split of the dataset not used during training (AUC: 0.76). Further improvement to classification performance was observed by applying AutoML to the 182 prioritized mRNA transcript features (AUC: 0.79). Additionally, the more certain model predictions were also more likely to be correct, with the balanced accuracy of the most confident predictions ranging from 80-90% (Figure 2).

Our overall results, using post-mortem brain data, are not as perfect as the ones from the prior studies by Takahashi, Hayashi [10] and the Zhang, Xie [12], showing a balanced accuracy of 87.9% and a near 100% accuracy, respectively. Obviously, one practical advantage of the prior studies is that they are focused on blood, which is readily accessible and thus a preferred sample for clinically useful biomarkers. However, only 80% of the transcriptome is shared between blood and brain tissue [25]. Hence, brain mRNA expression markers are expected to be

more accurate and more useful for furthering our understanding of the pathophysiology of SCZ and identifying novel drug targets. Moreover, in contrast to Zhang et al (2017), our approach used a separate testing set to evaluate the trained model which was not used during training. Furthermore, for the ANN results by Takahashi et al. (2010), the training set consisted of 35 cases and 33 controls while the testing set consisted of only 17 SCZ cases and 16 controls. The authors suggested that small sample size to microarray features ratio, lack in statistical power, and potential confounding due to gender differences in the two groups are limitations of their study. We believe that our brain dataset, with 201 SCZ cases and 278 controls, while still relatively small when considered for ML purposes, provide a more reliable analysis. Finally, our study addresses another challenge identified in the literature. Both the above-mentioned prior studies specifically selected for differentially expressed probes using their entire dataset, which were then used as features during the training of a classification algorithm. This likely inflated the accuracy during validation of the model. Although the test sample was not used during the training of the ML algorithm, it had been used during the feature selection process, i.e., during the determination of the differentially expressed probes that the ML algorithm ultimately focused on. In contrast, our approach did not preselect for features and used the regularization mechanism of the XGBoost algorithm to reduce the effective number of features used in the final model. Lastly, we find that AutoML strategies such as Auto-Sklearn can yield even further improvements in performance using the selected features.

To show that the genes identified in our study have relevance to SCZ, we present findings from a literature review for the top ten ML genes used by our classifier. All ten genes were found to be related to SCZ etiology. *HBB* expression has been shown to be downregulated in post-mortem analysis of brain tissue [26, 27]. Similarly, *DTNA* was found to be differentially spliced

in brain samples when compared to controls [28], and to be differentially expressed in SCZ patients in response to clozapine [29]. *ITGB4* segregated with SCZ in a study of a family conducted by O'Brien, Fiorentino [30]. The gene expression of *COX7A1* was found to be downregulated in SCZ patients versus controls [31, 32]. Many studies have suggested that the monoamine oxidase *MAOB* is a significant biomarker for SCZ [33-35]. *SLC38A5* has also been found to be significantly associated with SCZ by Guan, Cai [36], while *LBH* is known to be differentially expressed in the brains of patients with SCZ compared to healthy controls [37-41]. *GALNTL1* (also known as *GALNT16*) was also recently found to be reduced in SCZ patients [42]. Although there are no current studies linking *NODAL* to SCZ, *NODAL* is a cytokine of the transforming growth factor-β (TGF- β) superfamily, and evidence suggests that TGF-β signaling is altered in SCZ [43].

Finally, there was no direct link to SCZ for *COPS3*, the top gene from the ML model. However, there is evidence that the COP9 signalosome, which *COPS3* is a subunit of, inhibits dendritic arborization in the peripheral nervous system of a drosophila model through [44]. This is consistent with loss of dendritic spine density observed in SCZ [45]. Interestingly, a recent study found that doxycycline, a commonly used antibiotic known to inhibit COP9 activity *in vitro* [46], prevented and reversed ketamine-induced schizophrenic-like behaviors in mice [47]. The role of *COPS3,* and that of inhibitors of the COP9 signalosome in SCZ, such as doxycycline and CSN5i-3 [48], should be investigated further.

Lastly, we implemented a novel approach to highlight the GO molecular function gene sets that are important for our ML classifier, and thus for SCZ. Our approach has some notable benefits. First, we performed a univariate differential gene expression analysis contrasting SCZ cases and controls based on all genes in the dataset, and prioritized the most important gene sets

based on the robust consensus gene set analysis method from the *piano* bioinformatic tool. The differential gene expression with *piano* approach has the benefit of utilizing the full set of gene expressions available; however, this approach also results in a lot of gene sets that may not be relevant to SCZ. To address this, we applied the gene set analysis on the genes of the ML classifier using a hypergeometric test and FDR adjustment. This approach allowed us to focus on the most important gene sets which are significantly differentially expressed based on both the univariate analysis using all genes, as well as the ones significantly enriched among the ML genes. We found two significant GO terms, among the ML classifier genes, for differentiating SCZ cases and controls: "*oxidoreductase activity, acting on the CH-NH2 group of donors"* and "*integrin binding*."

With regards to genes in the *oxidoreductase activity gene set,* we already showed evidence for a link to SCZ for the *MAOB* gene above. Furthermore, *GLDC,* another member of this gene set has also been linked to SCZ in the literature [49]. Interestingly, a placebo-controlled trial with two psychosis patients with the *GLDC* copy number variant, found that glycine and D-cycloserine improved psychotic and mood symptoms [49].

Four ML genes, *ITGB4*, *FERMT3*, *NISCH*, and *SRC,* were part of the *integrin binding gene set*. There is already a link to SCZ for *ITGB4* detailed above (one of the top ten ML genes). *NISCH* (Imidazoline receptor 1)*,* which reduces GABAergic synaptic transmission [50], has also been reported to be a SCZ risk gene and a potential treatment target by Imidazoline I1 receptor agonist drugs [51]. Lastly, *SRC* (Src kinase) activity has been found to be suppressed in SCZ cases, and has also been proposed to lead to decreased NMDAR signaling [52, 53]. Several drug targets have been proposed to increase NMDAR signaling, including glycine and D-serine which target the NMDAR receptors directly [54].

5. Conclusion

To conclude, we demonstrated that supervised ML analysis of gene expression microarray post-mortem data from the DLPFC could effectively distinguish SCZ cases from controls and further our understanding of the pathophysiology of SCZ, but also the identification of novel candidate treatments. We showed a novel approach of integrating results from multivariate ML analysis with differential expression analysis to identify and prioritize and identify robust gene sets relevant to SCZ. Lastly, we demonstrate the usefulness of our approach by finding several potentially interesting treatment targets such as the COP9 signalosome, *GLDC,* *NISCH,* and *SRC*.

Acknowledgments

We thank the authors and dbGaP for access to the dataset. We would also like to thank Dr. Celia Greenwood and Dr. Jeff Xia for their suggestions and feedback on the methodology and results.

Authors' contributions

Bill Qi performed the bioinformatic and machine learning analyses and drafted the manuscript under the supervision of Yannis Trakadis who conceived and coordinated the project. Bill Qi and Yannis Trakadis designed the original methodology. Sonia Boscenco and Janani Ramamurthy performed the background literature review. All authors reviewed and provided feedback on the manuscript.

Funding

Yannis Trakadis is supported by the McGill University Health Centre Research Institute, the Canada First Research Excellence Fund (McGill University Healthy Brains for Healthy Lives Initiative), and the FRQS Chercheur Boursier Clinicien salary award.

Availability of data and materials

The SCZ data used in the preparation of this manuscript were obtained from the Database of Genotypes and Phenotypes (dbGaP) after McGill IRB approval. Raw data used is available in study phs000979.v1.p1.

References

1. Institute of Medicine Committee on Nervous System Disorders in Developing, C., in *Neurological, Psychiatric, and Developmental Disorders: Meeting the Challenge in the Developing World*. 2001, National Academies Press (US) Copyright 2001 by the National Academy of Sciences. All rights reserved.: Washington (DC).
2. Cloutier, M., et al., *The Economic Burden of Schizophrenia in the United States in 2013.* J Clin Psychiatry, 2016. **77**(6): p. 764-71.
3. Hilker, R., et al., *Heritability of Schizophrenia and Schizophrenia Spectrum Based on the Nationwide Danish Twin Register.* Biol Psychiatry, 2018. **83**(6): p. 492-498.
4. Schizophrenia Working Group of the Psychiatric Genomics, C., *Biological insights from 108 schizophrenia-associated genetic loci.* Nature, 2014. **511**(7510): p. 421-427.
5. Avramopoulos, D., *Recent Advances in the Genetics of Schizophrenia.* Mol Neuropsychiatry, 2018. **4**(1): p. 35-51.

6. Fullerton, J.M. and J.I. Nurnberger, *Polygenic risk scores in psychiatry: Will they be useful for clinicians?* F1000Research, 2019. **8**: p. F1000 Faculty Rev-1293.

7. Kuzman, M.R., et al., *Genome-wide expression analysis of peripheral blood identifies candidate biomarkers for schizophrenia.* Journal of psychiatric research, 2009. **43**(13): p. 1073-1077.

8. Deo, R.C., *Machine Learning in Medicine.* Circulation, 2015. **132**(20): p. 1920-30.

9. Zhou, L., et al., *Machine learning on big data: Opportunities and challenges.* Neurocomputing, 2017. **237**: p. 350-361.

10. Takahashi, M., et al., *Diagnostic classification of schizophrenia by neural network analysis of blood-based gene expression signatures.* Schizophr Res, 2010. **119**(1-3): p. 210-8.

11. Logotheti, M., et al., *Development and validation of a skin fibroblast biomarker profile for schizophrenic patients.* AIMS Bioengineering, 2016. **3**(4): p. 552-565.

12. Zhang, H., et al., *The Correlation-Base-Selection Algorithm for Diagnostic Schizophrenia Based on Blood-Based Gene Expression Signatures.* BioMed research international, 2017. **2017**: p. 7860506-7860506.

13. Chechko, N., et al., *Differential Resting-State Connectivity Patterns of the Right Anterior and Posterior Dorsolateral Prefrontal Cortices (DLPFC) in Schizophrenia.* Front Psychiatry, 2018. **9**: p. 211.

14. Enwright Iii, J.F., et al., *Transcriptome alterations of prefrontal cortical parvalbumin neurons in schizophrenia.* Molecular Psychiatry, 2018. **23**(7): p. 1606-1613.

15. Chen, T. and C. Guestrin. *Xgboost: A scalable tree boosting system*. ACM.

16. Jiang, T., J.L. Gradus, and A.J. Rosellini, *Supervised Machine Learning: A Brief Primer.* Behavior Therapy, 2020. **51**(5): p. 675-687.

17. Kwakernaak, S., et al., *Using machine learning to predict mental healthcare consumption in non-affective psychosis.* Schizophr Res, 2020. **218**: p. 166-172.

18. Cho, G., et al., *Review of Machine Learning Algorithms for Diagnosing Mental Illness.* Psychiatry Investig, 2019. **16**(4): p. 262-269.

19. Feurer, M., et al., *Auto-sklearn 2.0: The next generation.* arXiv preprint arXiv:2007.04074, 2020.

20. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.

21. Consortium, G.O., *The Gene Ontology resource: enriching a GOld mine.* Nucleic Acids Res, 2021. **49**(D1): p. D325-d334.

22. Subramanian, A., et al., *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.* Proceedings of the National Academy of Sciences, 2005. **102**(43): p. 15545.

23. Varemo, L., J. Nielsen, and I. Nookaew, *Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods.* Nucleic Acids Res, 2013. **41**(8): p. 4378-91.

24. Leif Varemo Wigge, I.N., *Platform for Integrative Analysis of Omics data: Vignette*. 2020.

25. Liew, C.C., et al., *The peripheral blood transcriptome dynamically reflects system wide biology: a potential diagnostic tool.* J Lab Clin Med, 2006. **147**(3): p. 126-32.

26. Martins-de-Souza, D., et al., *Alterations in oligodendrocyte proteins, calcium homeostasis and new potential markers in schizophrenia anterior temporal lobe are revealed by shotgun proteome analysis.* J Neural Transm (Vienna), 2009. **116**(3): p. 275-89.

27. Martins-de-Souza, D., et al., *Prefrontal cortex shotgun proteome analysis reveals altered calcium homeostasis and immune system imbalance in schizophrenia.* Eur Arch Psychiatry Clin Neurosci, 2009. **259**(3): p. 151-63.

28. Gandal, M.J., et al., *Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder.* Science, 2018. **362**(6420).

29. Lee, B.J., et al., *Analysis of differential gene expression mediated by clozapine in human postmortem brains.* Schizophr Res, 2017. **185**: p. 58-66.

30. O'Brien, N.L., et al., *Rare variant analysis in multiply affected families, association studies and functional analysis suggest a role for the ITGBeta4 gene in schizophrenia and bipolar disorder.* Schizophr Res, 2018. **199**: p. 181-188.

31. Arion, D., et al., *Distinctive transcriptome alterations of prefrontal pyramidal neurons in schizophrenia and schizoaffective disorder.* Mol Psychiatry, 2015. **20**(11): p. 1397-405.

32. Hjelm, B.E., et al., *Evidence of Mitochondrial Dysfunction within the Complex Genetic Etiology of Schizophrenia.* Mol Neuropsychiatry, 2015. **1**(4): p. 201-19.

33. Camarena, B., et al., *Monoamine oxidase a and B gene polymorphisms and negative and positive symptoms in schizophrenia.* ISRN Psychiatry, 2012. **2012**: p. 852949.

34. Wei, Y.L., et al., *Association study of monoamine oxidase A/B genes and schizophrenia in Han Chinese.* Behav Brain Funct, 2011. **7**: p. 42.

35. Carrera, N., et al., *Recent adaptive selection at MAOB and ancestral susceptibility to schizophrenia.* Am J Med Genet B Neuropsychiatr Genet, 2009. **150b**(3): p. 369-74.

36. Guan, J., et al., *Commonality in dysregulated expression of gene sets in cortical brains of individuals with autism, schizophrenia, and bipolar disorder.* Transl Psychiatry, 2019. **9**(1): p. 152.

37. Struyf, J., S. Dobrin, and D. Page, *Combining gene expression, demographic and clinical data in modeling disease: a case study of bipolar disorder and schizophrenia.* BMC Genomics, 2008. **9**: p. 531.

38. Narayan, S., et al., *Molecular profiles of schizophrenia in the CNS at different stages of illness.* Brain Res, 2008. **1239**: p. 235-48.

39. Zhao, Z., et al., *Transcriptome sequencing and genome-wide association analyses reveal lysosomal function and actin cytoskeleton remodeling in schizophrenia and bipolar disorder.* Mol Psychiatry, 2015. **20**(5): p. 563-572.

40. Perez-Santiago, J., et al., *A combined analysis of microarray gene expression studies of the human prefrontal cortex identifies genes implicated in schizophrenia.* J Psychiatr Res, 2012. **46**(11): p. 1464-74.

41. Manchia, M., et al., *Pattern of gene expression in different stages of schizophrenia: Down-regulation of NPTX2 gene revealed by a meta-analysis of microarray datasets.* Eur Neuropsychopharmacol, 2017. **27**(10): p. 1054-1063.

42. Mueller, T.M., N.R. Mallepalli, and J.H. Meador-Woodruff, *Altered protein expression of galactose and N-acetylgalactosamine transferases in schizophrenia superior temporal gyrus.* bioRxiv, 2019: p. 649996.

43. Nakashima, H., et al., *Canonical TGF-β Signaling Negatively Regulates Neuronal Morphogenesis through TGIF/Smad Complex-Mediated CRMP2 Suppression.* The Journal of neuroscience : the official journal of the Society for Neuroscience, 2018. **38**(20): p. 4791-4810.

44. Djagaeva, I. and S. Doronkin, *COP9 Limits Dendritic Branching via Cullin3-Dependent Degradation of the Actin-Crosslinking BTB-Domain Protein Kelch.* PLOS ONE, 2009. **4**(10): p. e7598.

45. Moyer, C.E., M.A. Shelton, and R.A. Sweet, *Dendritic spine alterations in schizophrenia.* Neuroscience letters, 2015. **601**: p. 46-53.

46. Pulvino, M., et al., *Inhibition of COP9-signalosome (CSN) deneddylating activity and tumor growth of diffuse large B-cell lymphomas by doxycycline.* Oncotarget, 2015. **6**(17): p. 14796-14813.

47. Ben-Azu, B., et al., *Doxycycline prevents and reverses schizophrenic-like behaviors induced by ketamine in mice via modulation of oxidative, nitrergic and cholinergic pathways.* Brain research bulletin, 2018. **139**: p. 114-124.

48. Schlierf, A., et al., *Targeted inhibition of the COP9 signalosome for treatment of cancer.* Nature communications, 2016. **7**: p. 13166-13166.

49. Bodkin, J.A., et al., *Targeted Treatment of Individuals With Psychosis Carrying a Copy Number Variant Containing a Genomic Triplication of the Glycine Decarboxylase Gene.* Biological psychiatry, 2019. **86**(7): p. 523-535.

50. Tanabe, M., et al., *Presynaptic I1-imidazoline receptors reduce GABAergic synaptic transmission in striatal medium spiny neurons.* The Journal of neuroscience : the official journal of the Society for Neuroscience, 2006. **26**(6): p. 1795-1802.

51. Lencz, T. and A.K. Malhotra, *Targeting the schizophrenia genome: a fast track strategy from GWAS to clinic.* Molecular psychiatry, 2015. **20**(7): p. 820-826.

52. Pitcher, G.M., et al., *Schizophrenia susceptibility pathway neuregulin 1-ErbB4 suppresses Src upregulation of NMDA receptors.* Nature medicine, 2011. **17**(4): p. 470-478.

53. Banerjee, A., et al., *Src kinase as a mediator of convergent molecular abnormalities leading to NMDAR hypoactivity in schizophrenia.* Molecular psychiatry, 2015. **20**(9): p. 1091-1100.

54. Kantrowitz, J.T. and D.C. Javitt, *N-methyl-d-aspartate (NMDA) receptor dysfunction or dysregulation: The final common pathway on the road to schizophrenia?* Brain Research Bulletin, 2010. **83**(3): p. 108-121.

Supplemental Figures and Tables

Supplemental Table 1. Most important features identified by the best machine learning model listed from most to least importance.

Feature mapped to gene names or original probe IDs (n=182)
COPS3, HBB, DTNA, ITGB4, COX7A1, MAOB, SLC38A5, LBH, NODAL, GALNTL1, NPW, AMIGO3, TRIM6-TRIM34, LOC338758, GJA3, LOC116412, HNRPR, ADAD2, PTPLA, LOC100128542, MYRIP, NUDT14, ILMN_1904618, FJX1, LOC346887, CT45-2, VPS25, ILMN_1877818, NEXN, ILMN_1917045, LOC643784, IDH1, CAMP, LOC642797, LOC644162, ABTB1, KIR2DL4, ST6GAL1, ETV5, TMTC2, MIR658, LOC100128613, ZNF180, LOC650612, CIB4, NEUROD6, OR4F16, MAPBPIP, SDCBP2, FABP3, STXBP3, TPRXL, FAM173A, AZGP1, MRPL27, XPO7, LOC653352, SDC4, MTMR8, MIR1234, LOC649896, NS3BP, C1R, CARD16, PAG1, SAC3D1, LOC653648, IQCA1, LOC100134644, C11orf70, LOC645251, LOC100128975, VSX1, LOC730226, INPP4A, KCNS3, CEL, LOC646452, LGALS3, OR11G2, LOC653424, CERK, LOC645662, SRC, TRIM14, NMUR2, CD14, SFXN4, ILMN_1866563, GLDC, SEPHS2, LOC652668, PRSS1, KCNK6, LOC649613, ILMN_1823752, RNF217, ZNF831, FASTKD5, PRDM1, LOC100131514, ILMN_1843198, LOC642147, CCNYL1, ADO, OR7E91P, LOC100130644, OR10X1, TGDS, LOC651728, LOC400807, IDI1, C11orf39, LOC100132807, SEMA6D, S100PBP, AKAP1, GPR155, LOC647163, DEPDC4, LOC728991, CCNYL1, C3orf41, C3orf25, LOC642980, AP1B1, MIR32, NISCH, LOC643959, LOC402160, EDEM3, ANXA3, HSD17B11, FOXD4L2, FERMT3, DSTN, NDUFC1, NSUN7, LOC652698, LOC390387, ALDH6A1, HIST2H3A, PLOD2, BCMO1, RGS1, ILMN_1847202, ILMN_1908780, TMEM18, TRPV6, IGFBP6, LOC390714, LOC342541, ILMN_1916146, USP53, MTNR1A, TIMP3, APOL1, AFAP1, DDX12, ILMN_1823270, LOC643213, RASAL3, LOC100128302, MCM8, FLJ36701, PPM1H, MGC10646, LOC647480, STON2, AFG3L1, FILIP1L, LOC100130344, LOC645176, EDN3, KLK11, ILMN_1852349, GOLPH4, SCARA3, OR2A25, DPP10, BCAR3, LOC642771

Supplemental Table 1. After obtaining the best trained classification model, we extracted the feature weights (i.e., number of times a feature is used in splitting the data in all XGBoost trees). The features are then ranked from highest to lowest weight. All features with a weight above zero are mapped to the corresponding gene. If a probe couldn't be mapped to a known gene, the Illumina probe ID is shown instead.

Supplemental Table 2. Hypergeometric test of enrichment of *piano* ranked gene sets in machine learning model genes.

Gene set name	p-value	FDR	# of overlap with ML genes	Gene set size	Overlapping ML genes
GO_OXIDOREDUCTASE_ACTIVITY_ACTING_ON_THE_CH_NH2_GROUP_OF_DONORS	0.0003	0.0568	2	21	GLDC, MAOB
GO_INTEGRIN_BINDING	0.001	0.1109	4	127	FERMT3, NISCH, SRC, ITGB4
GO_METALLOENDOPEPTIDASE_INHIBITOR_ACTIVITY	0.0036	0.2267	1	15	TIMP3
GO_ELECTRON_TRANSFER_ACTIVITY	0.005	0.2267	3	114	GLDC, MAOB, COX7A1
GO_OXIDOREDUCTASE_ACTIVITY	0.0052	0.2267	10	747	IDH1, COX7A1, MAOB, HBB, NDUFC1, PLOD2, HSD17B11, ADO, GLDC, ALDH6A1
GO_COENZYME_BINDING	0.0077	0.2482	5	286	IDH1, MAOB, PLOD2, GLDC, ALDH6A1
GO_ORGANIC_ACID_BINDING	0.008	0.2482	4	205	GLDC, HBB, FABP3, PLOD2
GO_COFACTOR_BINDING	0.0104	0.2489	7	498	IDH1, MAOB, HBB, PLOD2, SRC, GLDC, ALDH6A1
GO_MONOSACCHARIDE_BINDING	0.0105	0.2489	2	75	PLOD2, SCARA3
GO_FATTY_ACID_DERIVATIVE_BINDING	0.0115	0.2489	1	27	ALDH6A1
GO_FATTY_ACID_BINDING	0.0178	0.3413	1	34	FABP3
GO_PASSIVE_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.0198	0.3413	6	454	KCNS3, TRPV6, KCNK6, APOL1, NMUR2, GJA3
GO_OXIDOREDUCTASE_ACTIVITY_ACTING_ON_THE_ALDEHYDE_OR_OXO_GROUP_OF_DONORS_NAD_OR_NADP_AS_ACCEPTOR	0.0209	0.3413	1	37	ALDH6A1

GO_CHLORIDE_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.022	0.3413	2	99	NMUR2, APOL1
GO_OXIDOREDUCTASE_ACTIVITY_ACTING_ON_THE_ALDEHYDE_OR_OXO_GROUP_OF_DONORS	0.0289	0.35	1	44	ALDH6A1
GO_STRUCTURAL_CONSTITUENT_OF_MUSCLE	0.0289	0.35	1	44	NEXN
GO_BOX_H_ACA_SNORNA_BINDING	0.0298	0.35	0	5	
GO_CYCLIN_DEPENDENT_PROTEIN_SERINE_THREONINE_KINASE_REGULATOR_ACTIVITY	0.0353	0.35	1	49	CCNYL1
GO_TRANSPORTER_ACTIVITY	0.0368	0.35	12	1250	KCNS3, COX7A1, TRPV6, AP1B1, AZGP1, SFXN4, APOL1, SLC38A5, KCNK6, NMUR2, GJA3, FABP3
GO_NEUROTRANSMITTER_TRANSPORTER_ACTIVITY	0.038	0.35	1	51	SLC38A5
GO_ION_CHANNEL_BINDING	0.0391	0.35	2	124	DPP10, SRC
GO_PROTEASE_BINDING	0.0399	0.35	2	125	TIMP3, CARD16
GO_MAGNESIUM_ION_BINDING	0.04	0.35	3	213	IDH1, CERK, CIB4
GO_LEUCINE_BINDING	0.0415	0.35	0	7	
GO_OXIDOREDUCTASE_ACTIVITY_ACTING_ON_CH_OH_GROUP_OF_DONORS	0.0424	0.35	2	128	IDH1, HSD17B11
GO_HYDRO_LYASE_ACTIVITY	0.0435	0.35	1	55	TGDS
GO_NUCLEOSOMAL_DNA_BINDING	0.0435	0.35	1	55	HIST2H3A
GO_EXTRACELLULAR_MATRIX_BINDING	0.0464	0.3506	1	57	LGALS3
GO_GROWTH_FACTOR_BINDING	0.0501	0.3506	2	137	IGFBP6, ITGB4
GO_VITAMIN_BINDING	0.0501	0.3506	2	137	GLDC, PLOD2
GO_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.0501	0.3506	10	1047	KCNS3, COX7A1, TRPV6, AZGP1, SLC38A5, SFXN4, APOL1, KCNK6, NMUR2, GJA3
GO_VOLTAGE_GATED_CATION_CHANNEL_ACTIVITY	0.0537	0.3543	2	141	KCNS3, KCNK6
GO_CHANNEL_REGULATOR_ACTIVITY	0.0565	0.3543	2	144	DPP10, KCNS3
GO_RRNA_BINDING	0.0571	0.3543	1	64	FASTKD5
GO_MONOCARBOXYLIC_ACID_BINDING	0.0587	0.3543	1	65	FABP3
GO_CALCIUM_DEPENDENT_PROTEIN_KINASE_ACTIVITY	0.0588	0.3543	0	10	
GO_NUCLEOSIDE_TRIPHOSPHATE_DIPHOSPHATASE_ACTIVITY	0.0645	0.36	0	11	

GO_RAGE_RECEPTOR_BINDING	0.0645	0.36	0	11	
GO_POTASSIUM_ION_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.0715	0.36	2	159	KCNS3, KCNK6
GO_LIPID_BINDING	0.0716	0.36	7	726	ANXA3, SDCBP2, APOL1, NISCH, CAMP, CD14, FABP3
GO_CARBON_OXYGEN_LYASE_ACTIVITY	0.0719	0.36	1	73	TGDS
GO_ENZYME_INHIBITOR_ACTIVITY	0.0731	0.36	4	370	TIMP3, LGALS3, ANXA3, CARD16
GO_AMINO_ACID_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.0736	0.36	1	74	SLC38A5
GO_DOUBLE_STRANDED_RNA_BINDING	0.0736	0.36	1	74	ADAD2
GO_CELL_ADHESION_MOLECULE_BINDING	0.0747	0.36	5	488	IDH1, FERMT3, NISCH, SRC, ITGB4
GO_PHOSPHATASE_ACTIVITY	0.0798	0.3649	3	269	MTMR8, PPM1H, INPP4A
GO_FLAVIN_ADENINE_DINUCLEOTIDE_BINDING	0.0806	0.3649	1	78	MAOB
GO_ALDEHYDE_DEHYDROGENASE_NAD_ACTIVITY	0.0813	0.3649	0	14	
GO_ION_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.083	0.3649	8	879	KCNS3, COX7A1, TRPV6, KCNK6, SLC38A5, SFXN4, APOL1, NMUR2
GO_G_PROTEIN_COUPLED_RECEPTOR_BINDING	0.0841	0.3649	3	274	NPW, EDN3, ITGB4
GO_S100_PROTEIN_BINDING	0.0869	0.3665	0	15	
GO_NUCLEOSOME_BINDING	0.0878	0.3665	1	82	HIST2H3A
GO_PDZ_DOMAIN_BINDING	0.0915	0.3713	1	84	DTNA
GO_DOPAMINE_RECEPTOR_BINDING	0.0924	0.3713	0	16	
GO_ANTIOXIDANT_ACTIVITY	0.0952	0.3719	1	86	HBB
GO_LYASE_ACTIVITY	0.0965	0.3719	2	181	GLDC, TGDS
GO_PHOSPHOPROTEIN_PHOSPHATASE_ACTIVITY	0.0977	0.3719	2	182	MTMR8, PPM1H
GO_STEROID_HORMONE_RECEPTOR_BINDING	0.1028	0.3818	1	90	SRC
GO_PROTEIN_C_TERMINUS_BINDING	0.1038	0.3818	2	187	SDCBP2, SRC
GO_ACTIN_FILAMENT_BINDING	0.1076	0.387	2	190	DSTN, NEXN
GO_ACTIN_BINDING	0.1118	0.387	4	422	AFAP1, DSTN, NEXN, MYRIP
GO_PHOSPHATASE_REGULATOR_ACTIVITY	0.1124	0.387	1	95	LGALS3
GO_VOLTAGE_GATED_ION_CHANNEL_ACTIVITY	0.1152	0.387	2	196	KCNS3, KCNK6

GO_DRUG_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.1163	0.387	1	97	SLC38A5
GO_CALMODULIN_BINDING	0.1178	0.387	2	198	TRPV6, RGS1
GO_CALCIUM_RELEASE_CHANNEL_ACTIVITY	0.1195	0.387	0	21	
GO_THREONINE_TYPE_PEPTIDASE_ACTIVITY	0.1195	0.387	0	21	
GO_PROTEIN_CONTAINING_COMPLEX_BINDING	0.1255	0.3977	9	1099	LGALS3, NEXN, FERMT3, NISCH, SRC, HIST2H3A, DSTN, ITGB4, MCM8
GO_CATION_CHANNEL_ACTIVITY	0.1272	0.3977	3	319	TRPV6, KCNS3, KCNK6
GO_ANION_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.1293	0.3977	3	321	NMUR2, SLC38A5, APOL1
GO_TRANSFERASE_ACTIVITY_TRANSFERRING_NITROGENOUS_GROUPS	0.1301	0.3977	0	23	
GO_LIGAND_GATED_CATION_CHANNEL_ACTIVITY	0.1324	0.3985	1	105	KCNK6
GO_MOLECULAR_ADAPTOR_ACTIVITY	0.1366	0.3985	2	212	PAG1, SRC
GO_NUCLEAR_RECEPTOR_BINDING	0.1385	0.3985	1	108	SRC
GO_EXODEOXYRIBONUCLEASE_ACTIVITY	0.1406	0.3985	0	25	
GO_TRANSLATION_REGULATOR_ACTIVITY_NUCLEIC_ACID_BINDING	0.1427	0.3985	1	110	ABTB1
GO_PEPTIDASE_REGULATOR_ACTIVITY	0.1436	0.3985	2	217	TIMP3, CARD16
GO_KINASE_BINDING	0.1479	0.3985	6	725	CARD16, CCNYL1, AP1B1, SDC4, NISCH, SRC
GO_LIGAND_GATED_CALCIUM_CHANNEL_ACTIVITY	0.151	0.3985	0	27	
GO_MAP_KINASE_KINASE_KINASE_ACTIVITY	0.151	0.3985	0	27	
GO_OXIDOREDUCTASE_ACTIVITY_ACTING_ON_THE_CH_NH_GROUP_OF_DONORS	0.151	0.3985	0	27	
GO_GATED_CHANNEL_ACTIVITY	0.1511	0.3985	3	341	KCNK6, KCNS3, NMUR2
GO_AU_RICH_ELEMENT_BINDING	0.1561	0.3985	0	28	
GO_CALMODULIN_DEPENDENT_PROTEIN_KINASE_ACTIVITY	0.1561	0.3985	0	28	
GO_STRUCTURAL_CONSTITUENT_OF_NUCLEAR_PORE	0.1561	0.3985	0	28	
GO_ACID_THIOL_LIGASE_ACTIVITY	0.1663	0.4168	0	30	
GO_AMIDE_BINDING	0.1671	0.4168	3	355	CD14, NMUR2, ALDH6A1
GO_ALPHA_ACTININ_BINDING	0.1713	0.4177	0	31	

GO_PEPTIDE_ANTIGEN_BINDING	0.1713	0.4177	0	31	
GO_PHOSPHORIC_ESTER_HYDROLASE_ACTIVITY	0.1789	0.4224	3	365	MTMR8, PPM1H, INPP4A
GO_SULFUR_COMPOUND_BINDING	0.1799	0.4224	2	242	CEL, ALDH6A1
GO_INTRACELLULAR_LIGAND_GATED_ION_CHANNEL_ACTIVITY	0.1813	0.4224	0	33	
GO_SIGNALING_RECEPTOR_BINDING	0.1828	0.4224	12	1633	IDH1, SEMA6D, IGFBP6, LGALS3, CEL, NISCH, FERMT3, SRC, EDN3, NODAL, NPW, ITGB4
GO_PHOSPHATIDYLINOSITOL_BINDING	0.183	0.4224	2	244	SDCBP2, NISCH
GO_PROTEIN_TYROSINE_KINASE_ACTIVITY	0.1852	0.4231	1	130	SRC
GO_MOLECULAR_FUNCTION_REGULATOR	0.1903	0.4278	13	1795	BCAR3, ANXA3, KCNS3, SEMA6D, CARD16, RGS1, CCNYL1, LGALS3, RASAL3, TIMP3, EDN3, NODAL, DPP10
GO_RNA_POLYMERASE_II_REPRESSING_TRANSCRIPTION_FACTOR_BINDING	0.1912	0.4278	0	35	
GO_CATION_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.1952	0.4322	5	646	KCNS3, COX7A1, TRPV6, KCNK6, SLC38A5
GO_PROTEIN_PHOSPHATASE_BINDING	0.2005	0.4353	1	137	LGALS3
GO_MONOVALENT_INORGANIC_CATION_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.2009	0.4353	3	383	KCNS3, KCNK6, COX7A1
GO_LIGAND_GATED_ION_CHANNEL_ACTIVITY	0.2027	0.4353	1	138	KCNK6
GO_CALCIUM_ION_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.2049	0.4353	1	139	TRPV6
GO_TRANSLATION_REGULATOR_ACTIVITY	0.2071	0.4353	1	140	ABTB1
GO_ACTININ_BINDING	0.2106	0.4353	0	39	
GO_ATP_DEPENDENT_DNA_HELICASE_ACTIVITY	0.2106	0.4353	0	39	
GO_UBIQUITIN_LIKE_PROTEIN_CONJUGATING_ENZYME_ACTIVITY	0.2154	0.4409	0	40	
GO_ENDONUCLEASE_ACTIVITY_ACTIVE_WITH_EITHER_RIBO_OR_DEOXYRIBONUCLEIC_ACIDS_AND_PRODUCING_5_PHOSPHOMONOESTERS	0.2201	0.4429	0	41	
GO_LIPID_TRANSPORTER_ACTIVITY	0.2204	0.4429	1	146	FABP3
GO_GENERAL_TRANSCRIPTION_INITIATION_FACTOR_BINDING	0.2249	0.4439	0	42	

GO_CHROMATIN_BINDING	0.2251	0.4439	4	538	MCM8, HIST2H3A, VSX1, PRDM1
GO_HELICASE_ACTIVITY	0.2271	0.4439	1	149	MCM8
GO_ISOMERASE_ACTIVITY	0.2293	0.4443	1	150	IDI1
GO_NUCLEAR_HORMONE_RECEPTOR_BINDING	0.2315	0.4446	1	151	SRC
GO_CALCIUM_ION_BINDING	0.2456	0.4675	5	698	ANXA3, CIB4, C1R, EDEM3, SCARA3
GO_PROTEIN_HOMODIMERIZATION_ACTIVITY	0.2521	0.472	6	849	VPS25, IDH1, ST6GAL1, SDCBP2, MAOB, GLDC
GO_PHOSPHOLIPID_BINDING	0.2523	0.472	3	423	SDCBP2, NISCH, ANXA3
GO_DNA_HELICASE_ACTIVITY	0.2616	0.4847	0	50	
GO_PEPTIDE_BINDING	0.2636	0.4847	2	295	CD14, NMUR2
GO_HYDROLASE_ACTIVITY_ACTING_ON_ESTER_BONDS	0.2722	0.4891	5	724	MTMR8, AZGP1, CEL, PPM1H, INPP4A
GO_GTPASE_REGULATOR_ACTIVITY	0.2734	0.4891	2	301	RASAL3, RGS1
GO_METAL_ION_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.275	0.4891	3	440	TRPV6, KCNS3, KCNK6
GO_3_5_EXONUCLEASE_ACTIVITY	0.275	0.4891	0	53	
GO_ENZYME_REGULATOR_ACTIVITY	0.2777	0.49	7	1025	ANXA3, CARD16, RGS1, CCNYL1, LGALS3, RASAL3, TIMP3
GO_RAB_GTPASE_BINDING	0.2808	0.4913	1	173	MYRIP
GO_STEROL_BINDING	0.2881	0.5001	0	56	
GO_OXIDOREDUCTASE_ACTIVITY_ACTING_ON_THE_CH_CH_GROUP_OF_DONORS	0.2967	0.5069	0	58	
GO_TRNA_BINDING	0.2967	0.5069	0	58	
GO_PHOSPHATASE_BINDING	0.3009	0.5101	1	182	LGALS3
GO_HORMONE_RECEPTOR_BINDING	0.3054	0.5134	1	184	SRC
GO_STRUCTURAL_CONSTITUENT_OF_RIBOSOME	0.3076	0.5134	1	185	MRPL27
GO_CADHERIN_BINDING	0.3113	0.5157	2	324	IDH1, SRC
GO_MOLECULAR_TRANSDUCER_ACTIVITY	0.3277	0.5371	10	1544	OR2A25, KIR2DL4, MTNR1A, OR10X1, OR4F16, OR11G2, SDC4, CEL, NMUR2, CD14
GO_METHYLATED_HISTONE_BINDING	0.3301	0.5371	0	66	
GO_NUCLEASE_ACTIVITY	0.332	0.5371	1	196	AZGP1
GO_COLLAGEN_BINDING	0.3341	0.5371	0	67	

GO_DAMAGED_DNA_BINDING	0.3382	0.5377	0	68	
GO_NUCLEOSIDE_TRIPHOSPHATASE_REGULATOR_ACTIVITY	0.3395	0.5377	2	341	RASAL3, RGS1
GO_KINASE_REGULATOR_ACTIVITY	0.354	0.5567	1	206	CCNYL1
GO_PEPTIDE_N_ACETYLTRANSFERASE_ACTIVITY	0.3619	0.5609	0	74	
GO_GUANYL_NUCLEOTIDE_EXCHANGE_FACTOR_ACTIVITY	0.3649	0.5609	1	211	BCAR3
GO_TRANSMEMBRANE_SIGNALING_RECEPTOR_ACTIVITY	0.365	0.5609	8	1277	OR2A25, KIR2DL4, MTNR1A, OR10X1, OR4F16, OR11G2, NMUR2, CD14
GO_PURINE_NTP_DEPENDENT_HELICASE_ACTIVITY	0.3696	0.5609	0	76	
GO_UBIQUITIN_BINDING	0.3696	0.5609	0	76	
GO_EXONUCLEASE_ACTIVITY	0.381	0.5719	0	79	
GO_CYTOKINE_ACTIVITY	0.3822	0.5719	1	219	NODAL
GO_ATPASE_BINDING	0.3848	0.5719	0	80	
GO_UBIQUITIN_LIKE_PROTEIN_LIGASE_ACTIVITY	0.393	0.5801	1	224	RNF217
GO_MICROTUBULE_MOTOR_ACTIVITY	0.3996	0.5822	0	84	
GO_ALCOHOL_BINDING	0.4033	0.5822	0	85	
GO_GTPASE_BINDING	0.4053	0.5822	3	535	BCAR3, MYRIP, XPO7
GO_PROTEIN_DIMERIZATION_ACTIVITY	0.4074	0.5822	8	1325	VPS25, IDH1, ST6GAL1, SDCBP2, MAOB, NEUROD6, GLDC, HIST2H3A
GO_DNA_BINDING_TRANSCRIPTION_REPRESSOR_ACTIVITY_RNA_POLYMERASE_II_SPECIFIC	0.4078	0.5822	1	231	PRDM1
GO_SINGLE_STRANDED_RNA_BINDING	0.4105	0.5822	0	87	
GO_DNA_DEPENDENT_ATPASE_ACTIVITY	0.4141	0.5834	0	88	
GO_N_ACETYLTRANSFERASE_ACTIVITY	0.4176	0.5844	0	89	
GO_PROTEIN_DOMAIN_SPECIFIC_BINDING	0.4201	0.5844	4	703	DTNA, PAG1, SRC, CARD16
GO_SERINE_TYPE_ENDOPEPTIDASE_INHIBITOR_ACTIVITY	0.4316	0.5966	0	93	
GO_CYTOKINE_RECEPTOR_ACTIVITY	0.4385	0.5985	0	95	
GO_STEROID_BINDING	0.4385	0.5985	0	95	

GO_UBIQUITIN_LIKE_PROTEIN_BINDING	0.4419	0.5994	0	96	
GO_STRUCTURAL_CONSTITUENT_OF_CYTOSKELETON	0.4487	0.6048	0	98	
GO_IDENTICAL_PROTEIN_BINDING	0.4609	0.6173	10	1711	VPS25, IDH1, ST6GAL1, CARD16, SDCBP2, MAOB, SDC4, NUDT14, NISCH, GLDC
GO_CHAPERONE_BINDING	0.4653	0.6194	0	103	
GO_RNA_POLYMERASE_II_SPECIFIC_DNA_BINDING_TRANSCRIPTION_FACTOR_BINDING	0.4754	0.6288	1	264	SRC
GO_SINGLE_STRANDED_DNA_BINDING	0.4781	0.6288	0	107	
GO_ACETYLTRANSFERASE_ACTIVITY	0.4813	0.6292	0	108	
GO_KINASE_ACTIVITY	0.4879	0.634	4	760	SEPHS2, FASTKD5, CERK, SRC
GO_SMALL_GTPASE_BINDING	0.4971	0.6421	2	439	MYRIP, XPO7
GO_CYTOSKELETAL_PROTEIN_BINDING	0.5155	0.6619	5	950	NEXN, MYRIP, AFAP1, FABP3, DSTN
GO_CATALYTIC_ACTIVITY_ACTING_ON_A_TRNA	0.5238	0.6686	0	122	
GO_PRIMARY_ACTIVE_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.5296	0.672	0	124	
GO_CYTOKINE_BINDING	0.5381	0.6789	0	127	
GO_RIBONUCLEOPROTEIN_COMPLEX_BINDING	0.5437	0.6811	0	129	
GO_UNFOLDED_PROTEIN_BINDING	0.5465	0.6811	0	130	
GO_ATPASE_ACTIVITY_COUPLED_TO_MOVEMENT_OF_SUBSTANCES	0.5493	0.6811	0	131	
GO_UBIQUITIN_LIKE_PROTEIN_LIGASE_BINDING	0.5545	0.6837	1	306	SRC
GO_MOTOR_ACTIVITY	0.5655	0.6933	0	137	
GO_GTPASE_ACTIVITY	0.5807	0.7079	1	321	RGS1
GO_LIGASE_ACTIVITY	0.5888	0.7138	0	146	
GO_PROTEIN_HETERODIMERIZATION_ACTIVITY	0.5938	0.7159	2	506	SDCBP2, HIST2H3A
GO_ENZYME_ACTIVATOR_ACTIVITY	0.6072	0.7261	2	516	RASAL3, RGS1
GO_DNA_BINDING_TRANSCRIPTION_FACTOR_BINDING	0.609	0.7261	1	338	SRC
GO_S_ADENOSYLMETHIONINE_DEPENDENT_METHYLTRANSFERASE_ACTIVITY	0.6132	0.7271	0	156	

GO_CATALYTIC_ACTIVITY_ACTING_ON_RNA	0.6265	0.7389	1	349	AZGP1
GO_PROXIMAL_PROMOTER_SEQUENCE_SPECIFIC_DNA_BINDING	0.6419	0.7529	2	543	NEUROD6, PRDM1
GO_TRANSFERASE_ACTIVITY_TRANSFERRING_PHOSPHORUS_CONTAINING_GROUPS	0.6498	0.7581	4	910	SEPHS2, FASTKD5, CERK, SRC
GO_CATALYTIC_ACTIVITY_ACTING_ON_DNA	0.6681	0.7753	0	181	
GO_REGULATORY_REGION_NUCLEIC_ACID_BINDING	0.6781	0.7827	4	940	ETV5, VSX1, NEUROD6, PRDM1
GO_ANTIGEN_BINDING	0.6898	0.7858	0	192	
GO_GUANYL_NUCLEOTIDE_BINDING	0.6906	0.7858	1	393	NMUR2
GO_HISTONE_BINDING	0.6917	0.7858	0	193	
GO_PROTEIN_KINASE_ACTIVITY	0.6971	0.7879	2	590	FASTKD5, SRC
GO_UBIQUITIN_LIKE_PROTEIN_TRANSFERASE_ACTIVITY	0.7104	0.7987	1	408	RNF217
GO_ZINC_ION_BINDING	0.7271	0.8132	3	809	TRIM14, DTNA, MYRIP
GO_RNA_BINDING	0.7397	0.8216	9	1922	LGALS3, MRPL27, FASTKD5, NSUN7, ADAD2, ABTB1, AKAP1, MIR32, ALDH6A1
GO_STRUCTURAL_MOLECULE_ACTIVITY	0.7421	0.8216	3	826	VPS25, MRPL27, NEXN
GO_MICROTUBULE_BINDING	0.7662	0.844	0	238	
GO_SEQUENCE_SPECIFIC_DOUBLE_STRANDED_DNA_BINDING	0.7743	0.8477	3	865	ETV5, NEUROD6, PRDM1
GO_TRANSCRIPTION_COREPRESSOR_ACTIVITY	0.7774	0.8477	0	246	
GO_TRANSITION_METAL_ION_BINDING	0.788	0.855	4	1077	PLOD2, DTNA, MYRIP, TRIM14
GO_TRANSFERASE_ACTIVITY_TRANSFERRING_ACYL_GROUPS	0.7958	0.8592	0	260	
GO_MRNA_BINDING	0.8024	0.862	1	492	MIR32
GO_SEQUENCE_SPECIFIC_DNA_BINDING	0.816	0.8719	4	1120	ETV5, VSX1, NEUROD6, PRDM1
GO_HYDROLASE_ACTIVITY_ACTING_ON_ACID_ANHYDRIDES	0.8196	0.8719	3	928	NUDT14, RGS1, MCM8
GO_DRUG_BINDING	0.8323	0.8784	7	1725	HBB, IQCA1, CERK, SRC, GLDC, SEPHS2, MCM8
GO_ADENYL_NUCLEOTIDE_BINDING	0.8351	0.8784	6	1541	IQCA1, CERK, SRC, SEPHS2, ALDH6A1, MCM8
GO_DOUBLE_STRANDED_DNA_BINDING	0.8379	0.8784	3	957	ETV5, NEUROD6, PRDM1
GO_TRANSCRIPTION_COACTIVATOR_ACTIVITY	0.8564	0.8935	0	317	

GO_TUBULIN_BINDING	0.8608	0.8937	0	322	
GO_ATPASE_ACTIVITY_COUPLED	0.8879	0.9146	0	357	
GO_ACTIVE_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.8893	0.9146	0	359	
GO_RIBONUCLEOTIDE_BINDING	0.8987	0.9178	7	1891	IQCA1, NMUR2, CERK, SRC, SEPHS2, ALDH6A1, MCM8
GO_DNA_BINDING_TRANSCRIPTION_FACTOR_ACTIVITY	0.901	0.9178	6	1701	ZNF180, NEUROD6, VSX1, TPRXL, PRDM1, ETV5
GO_TRANSCRIPTION_FACTOR_BINDING	0.9052	0.9178	1	643	SRC
GO_PROTEIN_SERINE_THREONINE_KINASE_ACTIVITY	0.9319	0.9399	0	437	
GO_ATPASE_ACTIVITY	0.9356	0.9399	0	446	
GO_TRANSCRIPTION_COREGULATOR_ACTIVITY	0.9693	0.9693	0	564	

Supplemental Table 2. The machine learning (ML) model genes from the best trained model was extracted. A secondary enrichment analysis was performed to rank the filtered gene sets obtained from the *piano* consensus gene set analysis method. A hypergeometric test was performed to determine which of the filtered gene sets were significantly enriched for the ML genes. We then applied the *Benjamini–Hochberg procedure to calculate the false-discovery rate (FDR). The gene sets are then sorted from lowest to highest by the p-value/FDR.*

ML: Machine learning

BH: *Benjamini–Hochberg*

FDR: False-discovery rate

Bridging statement to Chapter 3

In Chapter 2, we explored the use of ML in the analysis of transcriptomic data for understanding complex diseases. We hypothesized that ML analysis of gene expression microarray data can further our understanding of the pathophysiology of complex diseases such as SCZ. We performed a supervised ML analysis of gene expression microarray data from the DLPFC of post-mortem SCZ cases and controls and found substantial literature support for the top genes contributing most to the ML classifier performance having a link with SCZ. Furthermore, we also introduced a novel method of integration of ML findings with traditional differential gene expression analysis to identify robust biological functions including *"oxidoreductase activity, acting on the CH-NH2 group of donors,"* which associates the metabolism of biological amino groups with SCZ, and *"integrin binding,"* which highlights the function of cell signaling in SCZ. Another significant contribution of our study is that we designed our ML methodology specifically to address high dimensional transcriptomics data with relatively smaller number of samples. We noted that prior ML studies often did not explicitly address the challenges associated with analyzing data of this nature. We included several mechanisms for addressing challenges related to model selection, overfitting, and evaluation, including the use of a highly regularized XGBoost algorithm, hyperparameter selection with repeated k-fold cross-validation, and model evaluation using an independent testing set.

In the following chapter, we discuss the application of ML and transcriptomics in MDD. We further validate the supervised ML and gene set analysis approaches developed in Chapter 2. One of the limitations of our study in Chapter 2 is that evaluation of model performance was based only on a single dataset source. Such an analysis could be influenced by various factors unique to the dataset and its findings may thus not be generalizable outside of that dataset. In

Chapter 3, we include an additional brain DLPFC dataset as an external evaluation dataset in the analysis of MDD cases and controls. Moreover, to explore whether our methodology could be expanded to extracting insight from blood samples, which are more widely available from living patients, we include an additional dataset of blood gene expressions for MDD cases and controls, as a comparison to brain gene expressions. Lastly, we explore the integration of covariate data for model interpretation and deriving further insights from ML models.

Chapter 2 erratum:

- On page 56, the logistic loss function contains an error. The correct form should be: $l(y_i, \hat{y}_i) = y_i \ln\left(1 + e^{-\hat{y}_i}\right) + (1 - y_i) \ln(1 + e^{\hat{y}_i})$
- On page 58, *"The best cutoff is defined as the probability threshold dividing the cases and controls classes which maximizes the number of true positive classifications and minimizes the number of false-positive classifications (i.e., maximizing the area under the ROC curve)"* is inaccurate and should be corrected. We are not maximizing the area under the ROC curve, but rather, we are finding the probability threshold for which the rectangular area under the TPR and FPR point along the curve is maximized. Another equivalently correct way for defining the best cutoff is as the probability threshold dividing the predicted samples where the average of the true-positive rate (TPR) and the inverse of the false-positive rate (FPR) (i.e., 1 - FPR) is maximized.

Chapter 3. Machine learning and bioinformatic analysis of brain and blood mRNA profiles in major depressive disorder: a case-control study

AUTHORS

Bill Qi[1], Janani Ramamurthy[2], Imane Bennani[2], Yannis J. Trakadis[1,3]

AUTHORS INSTITUTIONAL AFFILIATIONS

[1] Department of Human Genetics, McGill University, Montreal, QC, Canada

[2] Faculty of Science, McGill University, Montreal, QC, Canada

[3] Department of Medical Genetics, McGill University Health Center, Montreal, QC, Canada

Correspondence to

Yannis J. Trakadis, MD MSc FRCPC FCCMG

Medical Geneticist & Metabolics Specialist

Assistant Professor, Human Genetics

McGill University Health Centre

Room A04.3140, Montreal Children's Hospital

1001 Boul. Décarie, Montreal, Quebec, Canada, H4A 3J1

Tel: (514) 412-4427, Fax: (514) 412-4296

Email: yannis.trakadis@mcgill.ca

Abstract

This study analyzed gene expression (mRNA) data, from cases with major depression (MDD) and controls, using supervised machine learning (ML). We built on the methodology of prior studies to obtain more generalizable/reproducible results. First, we obtained a classifier trained on gene expression data from the dorsal-lateral-prefrontal-cortex (DLPFC) of post-mortem MDD cases (n=126) and controls (n=103). An average area-under-the-receiver-operating-characteristics-curve (AUC) from 10-fold cross-validation of 0.72 was noted, compared to an average AUC of 0.55 for a baseline classifier (p=0.0048). The classifier achieved an AUC of 0.76 on a previously unused testing-set. We also performed external validation using DLPFC gene expression values from an independent cohort of matched MDD cases (n=29) and controls (n=29), obtained from Affymetrix microarray (vs. Illumina microarray for the original cohort) (AUC:0.62). We highlighted gene sets differentially expressed in MDD that were enriched for genes identified by the ML algorithm. Next, we assessed the ML classification performance in *blood*-based microarray gene expression data from MDD cases (n=1581) and controls (n=369). We observed a mean AUC of 0.64 on 10-fold cross-validation, which was significantly above baseline (p=0.0020). Similar performance was observed on the testing-set (AUC:0.61). Finally, we analyzed the classification performance in covariates subgroups. We identified an interesting interaction between smoking and recall performance in MDD case prediction (58% accurate predictions in cases who are smokers vs. 43% accurate predictions in cases who are non-smokers). Overall, our results suggest that ML in combination with gene expression data and covariates could further our understanding of the pathophysiology in MDD.

Keywords

Major depression, Transcriptomics, Machine learning, Bioinformatics

Introduction

There is a lack of reliable biomarkers for Major Depressive Disorder (MDD) in clinical practice. Gene expression, measured in the form of messenger RNA (mRNA), could be useful to further our understanding of the pathophysiology of MDD and potentially lead to biomarker discovery and novel drug targets for treatment development. A given gene is transcribed to mRNA, which contains the coding instructions for the synthesis of polypeptide chains. Hence, mRNAs are the links between the genetic information stored in DNA and the encoded proteins, which are used in the different biological functions.

Several studies have focused on analyzing, using statistics, the transcriptome (i.e., the full range of mRNA) in different tissues, like the brain or peripheral blood, of MDD cases and control subjects. For example, L. Gao, Gao, Xu, and Xie (2015) performed an analysis of microarray gene expression samples obtained from three human brain tissues (hippocampus, prefrontal cortex, and striatum) of MDD patients and healthy controls to identify differentially expressed genes (DEGs). They identified 241, 218, and 327 DEGs in the MDD-hippocampus group, the MDD-prefrontal cortex group, and the MDD-striatum group, respectively. In each group, there was an enhancement of a variety of biological processes associated with the DEGs. A different study, by Woo, Lim, Myung, Kim, and Lee (2018), focused on microarray-based transcriptomic data from peripheral blood. The gene expression of 38 MDD patients and 14 healthy controls was analyzed to identify DEGs and biological mechanisms in MDD. Seven DEGs were identified in their subjects which were found to be involved in immune and inflammatory responses. Further, the authors analyzed antidepressant responders and non-responders after six weeks of treatment, which lead to the identification of additional DEGs and biological mechanisms associated with treatment response in MDD.

As exemplified by these studies, based on statistical analyses, there exist differences in gene expression patterns in MDD vs. controls, as well as between non-responders and responders to antidepressant treatment. There are a few studies which have utilized transcriptomic data in combination with ML in classifying MDD case and controls. Yi et al. (2012) applied a support vector machines (SVM) approach on blood gene expression data to classify drug-free, first-episode cases with subsyndromal symptomatic depression (SSD), MDD, and matched controls. Yu, Xue, Redei, and Bagheri (2016) applied a SVM approach to classify MDD cases and controls based on expression data of preselected blood RNA markers, measured using quantitative real-time polymerase chain reaction. Recently, a study by Bhak et al. (2019) utilized a multi-omics approach. A random forest classifier was applied on blood transcriptomic and methylomic data, combined, to distinguish between MDD cases, suicide attempters and controls. All of the above studies reported high classification accuracies ranging from 86-100%.

Additionally, machine learning (ML) methods have been used for discriminating individuals with MDD from healthy controls. A study conducted by Guilloux et al. (2015) used ML methods to analyze blood transcriptomic data. Samples were collected from 34 MDD patients with concurring anxiety at baseline and following 12 weeks of treatment with citalopram and psychotherapy. The results were compared to those of a matched control group of non-depressed patients. Using a model of 13 baseline gene expressions selected using a cross-validation procedure on the training set, they were able to predict nonremission with a corrected accuracy of 79.4% in a validation cohort. The findings of this study suggest that baseline peripheral blood-based gene expression can potentially predict nonremission following therapy with citalopram, thus optimizing therapy (precision medicine).

However, existing ML studies involving patients with MDD have relatively low sample sizes, and do not adequately address the issues of overfitting and generalization, associated with ML models. Given the promising results from ML-based approaches in analyzing complex datasets and the sparsity of studies applying ML to transcriptomics data, our paper aims to address this important gap in the medical literature. Our objective is to build on the ML methodology of prior studies in the field to generate results which are generalizable/reproducible and interpretable. We apply a state-of-the-art supervised ML algorithm to post-mortem gene expression microarray datasets from the dorsolateral prefrontal cortex (DLPFC) of patients with MDD and controls. Of note, there is evidence that there is decreased activity in the DLPFC of MDD patients from functional imaging studies, and that the DLPFC could play a role in regulating negative affect (Koenigs & Grafman, 2009). Having said this, blood is more readily accessible and offers an opportunity for quick clinical translation of any findings. Blood shares over 80% of the transcriptome with brain tissue (Liew, Ma, Tang, Zheng, & Dempsey, 2006). Moreover, shared expression trends in many biological processes have been found between whole-blood and prefrontal cortex (Sullivan, Fan, & Perou, 2006). To compare the accuracy of our ML approach using post-mortem DLPFC versus blood mRNA data, we repeated the analysis using blood transcriptomics profiles from a separate case-control dataset.

Methods

Major depression gene expression microarray datasets

First, our analyses focused on gene expression data from the DLPFC of adult patients with MDD and controls. More specifically, we obtained a gene expression dataset of post-mortem patients with mental disorders and control subjects from dbGaP (dbGaP Study

Accession: phs000979.v1.p1). The gene expression data were obtained through the Illumina HumanHT-12 v4 Expression BeadChip platform. For preprocessing, the dataset was background corrected with the *normexp* and quantile normalized using the *neqc* function from the *limma* R package (version 3.42.0) (Smyth, 2005). To convert the probe expression values to gene expression values, we mapped the probes to genes based on the mappings provided in the *illuminaHumanv4.db* R package. For genes with multiple mapped probes, we took the average between the probe expression values as the final gene expression value. We limited the analysis to adults (≥ 18 years old), and Caucasian subjects, due to significant imbalances in race between MDD cases and controls potentially confounding downstream analyses. Further, we checked for outlier subjects based on inspection of subject-wise boxplots and a covariate-annotated plot of principal components 1 and 2. The final processed dataset consists of 126 MDD cases and 103 controls (total of 229). Lastly, we explored the differences in covariates between MDD cases and controls. This dataset will be referred to as the *brain mRNA* dataset in subsequent sections.

We obtained two external replication datasets from the Gene Expression Omnibus (GEO) repository (GEO accession: GSE54567 and GSE54568). These datasets also contain gene expressions from the DLPFC from MDD cases and controls. The GSE54567 dataset consists of 14 male MDD cases and 14 matched controls, while the GSE54568 dataset consists of 15 female MDD cases and 15 matched controls. Both cohorts were originally described by Chang et al. (2014) and the datasets were obtained using the Affymetrix Human Genome U133 Plus 2.0 array platform. We thus combined the two datasets, and quantile normalized them together using the *normalizeBetweenArrays* function from the *limma* R package. The probe expression values were mapped to gene expressions using the probe set annotations from GEO (GEO accession:

GPL570). The final dataset consists of 29 MDD cases, and 29 controls (total of 58). This dataset will be referred to as the *external brain mRNA* dataset in subsequent sections.

We also obtained another MDD dataset from dbGaP (dbGaP Study Accession: phs000486.v1.p1) consisting of blood-based gene expression data (RNA quantified by the Affymetrix U219 expression assays platform) from the blood cells of patients with major depression and healthy controls. The dataset consists of two different cohorts of patients and controls (Netherland Twin Register and Netherlands Study of Depression and Anxiety (NESDA)). Due to possible confounding from structural differences between the two cohorts, lack of MDD cases in the NTR cohort, as well as differences in sample preparation, we performed our analysis using patients and controls from only the NESDA cohort. The NESDA cohort consists of 1581 MDD cases and 369 healthy controls aged 18 and above (total of 1950). The dataset was normalized using the *normalizeBetweenArrays* function from the *limma* R package. Probe expression values were mapped to gene expressions based on the probe set annotations from GEO (GEO accession: GPL13667). Further, we checked for outlier subjects based on inspection of subject-wise boxplots and a covariate-annotated plot of principal components 1 and 2. Lastly, we explored the differences in covariates between MDD cases and controls. This dataset will be referred to as the *blood mRNA* dataset in subsequent sections.

ML algorithm selection

Many powerful ML algorithms render themselves uninterpretable, making it difficult to understand their decision-making process. We used a state-of-the-art yet interpretable regularized gradient boosted machines (GBM) approach, XGBoost implementation, (Chen & Guestrin), which has been proved successful in a wide range of tasks, as illustrated in a recent

study from our group (Trakadis et al., 2018). Its highly regularized built-in feature selection and reduction characteristic and ability to rank features based on their relative importance to its decision process made it a great candidate for our study. Of note, a regularized algorithm penalizes itself for complexity, and thus uses only features that are relevant and brings the most intelligence to its architecture. In our study, this means selecting only transcriptomic features that have high predictive power and discarding the less informative ones.

Machine learning analysis

For each ML analysis performed, we randomly sampled 80% of the full dataset to be the *training* set, to identify the best classification model, and 20% to be the *testing* set for independent evaluation. For the *brain mRNA* dataset, the *training* set consisted of 103 MDD cases and 80 controls. The *testing* set consisted of 23 MDD cases and 23 controls. For the *blood mRNA* dataset, the *training* set consisted of 1267 MDD cases and 293 controls. The *testing* set consisted of 314 MDD cases and 76 controls. To ensure the training and testing datasets are adequately similar, we performed a comparison of the subject covariates in the *training* vs. *testing* sets.

For model selection, we used a 10-fold cross-validation combined with randomized hyperparameters approach repeated for 2500 iterations (i.e., training 2500 different models). The performance of each trained model is defined by the area under the receiver-operating-characteristic (ROC) curve (AUC), with cases being the positive class, averaged over all 10 cross-validation folds. We then repeated the above procedure to select the best baseline model (i.e., a model trained using the same cross-validation approach but with randomly permuted labels). Given the large number of models being trained (2500) and the complexity of the models,

it may be possible to pick an overfit model by chance. The baseline model, therefore, is used as a comparison of how well the model selection procedure performed by chance based on random permuted labels. To assess whether performance of the trained model is higher than chance, a one-sided Wilcoxon signed-rank test is used to compare the AUC values on each cross-validation fold between the best baseline and best trained model for significance. In order to apply the Wilcoxon signed-rank tests, the equivalence of the cross-validation folds is maintained between the trained vs. baseline model by utilizing consecutive splitting of the same already shuffled training set into 10-folds.

Lastly, the hyperparameters of the ML algorithm (e.g., number of boosting iterations, max-depth of trees, learning rate, etc.) from the best trained model with the best average AUC from 10-fold cross-validation are extracted. Using these extracted hyperparameters, the model is retrained on the full *training* set (without cross-validation to maximize sample size) to improve its performance before being evaluated on a holdout *testing* set (i.e., the data which was not used during the training phase).

Classification of cases and controls based on gene expression data

We trained binary classification models using the above-described approach to distinguish MDD cases from controls for both the *brain mRNA* and the *blood mRNA* datasets. To assess the best trained model, we calculated and plotted the ROC curve based on the *testing* set. In order to calculate the classification accuracy, the predicted class probabilities of each testing set sample need to be converted to a discrete case or control classification. A high probability means the sample is more likely to be a "case," and a low probability means the sample is more likely to be a "control." An optimal probability cutoff threshold is needed in order to split the

samples into the discrete classes. To determine the optimal cutoff threshold, we averaged the best cutoff values derived from the ROC curves from each cross-validation fold during training. The best cutoff is defined as the probability threshold dividing the cases and controls classes which maximized the number of true positive classifications and minimized the number of false-positive classifications (i.e., maximizing the area under the ROC curve). After a discrete class was assigned to the testing set samples, we calculated an overall balanced accuracy metric. The balanced accuracy adjusts for imbalanced classes and is defined as the arithmetic mean of the sensitivity and specificity of a classifier. The balanced accuracy is equivalent to regular accuracy in the case of balanced class sizes (i.e., total number of correct predictions divided by total number of predictions). Furthermore, we looked at the balanced classification accuracy from high to low prediction confidence levels (i.e., deviation from the optimal cutoff threshold, where a larger deviation means higher confidence). The following technique was used to generate a plot of rolling (high to low) balanced accuracy values for the testing set. First, predictions are sorted from the highest confidence to the lowest. Then, starting with a window of the top n (where n determines the size of the window) most confident predictions, a balanced accuracy is calculated. The size of the window n is chosen based on the size of the testing set, with testing sets with a higher number of samples or more class imbalance having a larger window size, in order to adequately capture samples from both classes for balanced accuracy calculation. The window shifts down by one each time, and the balanced accuracy calculation process repeats until the end of the confidence-ordered testing set is reached. A graph is then generated to visualize the rolling trend of balanced accuracy from highest to lowest prediction confidence subsets.

Lastly, to provide a more detailed interpretation of covariate subgroup performance of the ML model, we summarized the number of 1) correctly classified MDD cases, 2) incorrectly

classified MDD cases, 3) correctly classified controls, and 4) incorrectly classified controls under each covariate subgroup (i.e., male vs. female, smoker vs. non-smoker etc.). This was done to assess for the consistency of the classification performance and identify any interactions between subgroups of covariates and classification performance.

Classification of cases and controls based on gene expression and covariates

Using the approaches described in the previous section, we trained an ML model based on only covariates (sex, smoker status, alcohol status, weight, and height) from the *brain mRNA* dataset to distinguish MDD cases from controls. We also repeated the analysis by training an ML model based on both covariates and gene expression data. With regards to training/cross-validation and testing set samples the same approach was used, as described before. We compared the 10-fold cross-validation performance from the best performing model trained on 1) *gene expressions only (from the above section)*, 2) *covariates only*, and 3) *covariates and gene expressions,* using a two-sided Wilcoxon signed-rank test.

External validation of ML classifier

Although a separate testing set and subgroup performance analysis is used for evaluation of the ML models from the above sections, we obtained the *external brain mRNA* dataset to perform external validation of our ML results from the *brain mRNA* dataset and support their generalizability.

The *external brain mRNA* dataset consists of an equal number of matched cases and controls. However, it is not directly comparable with the original *brain mRNA* dataset since they are obtained from different array platforms. Thus, to be able to perform external validation of the

findings, we trained a separate logistic regression model based on the gene features selected by the XGBoost algorithm (*XGBoost genes*). Specifically, both *the brain mRNA* and *external brain mRNA* datasets were filtered for the common set of *XGBoost genes*. The gene feature columns in both filtered datasets were *separately* standardized to have zero mean and unit variance for compatibility. Next, the *brain mRNA* dataset was used for training the logistic regression model. Here, 10-fold cross-validation along with an exhaustive grid-search for hyperparameters (regularization method and weight) was performed to select the model with the highest mean AUC on the 10-fold cross-validation. The resulting hyperparameters were then used to refit the model on the whole *brain mRNA* dataset to produce the final logistic regression classifier. Lastly, the classifier was evaluated on the *external brain mRNA* dataset.

Gene set analyses

Differential expression statistics were obtained for each gene through the R *limma* package (version 3.42.0) using the full set of cases and controls adjusted for covariates (age, sex, smoker status, alcohol status, postmortem interval, and pH). To investigate the underlying pathophysiology of MDD, we performed *gene set* analyses (GSA) using all genes available with the cases against controls from the *brain mRNA* dataset. GSA tests for altered expression for groups of genes (*gene sets*) between two classes (i.e., cases vs. controls). A *gene se*t can represent a group of genes with a similar function or activity, or a group of genes belonging to the same biological process or pathway. We obtained the Gene Ontology (GO) molecular function *gene sets* from MSigDB (Liberzon et al., 2011) as the source for *gene sets*. The GO molecular functions *gene sets* group genes based on related activities performed by single or multiple gene products. GSA was performed using the R *piano* package (version 2.6.0) (Varemo,

Nielsen, & Nookaew, 2013). We applied the consensus ranking method from the *piano* tool by combining *gene set* significance results from all available GSA methods from *piano* (section 4.3 from Wigge and Nookaew (2020)). Any *gene sets* with a consensus ranking above 10 in any of the five *piano* directionality classes (i.e., five specific ways the *gene sets* can be significantly altered; section 4.4.2 from Wigge and Nookaew (2020)) were selected to be important, (i.e., different between cases and controls).

Gene set re-ranking with ML genes

As described above, the *piano* GSA method identified GO molecular functions *gene sets* exhibiting different expression between cases and controls. We performed a secondary enrichment analysis to augment the ranking of the *gene sets* derived from *piano* based on the genes utilized by the XGBoost model trained on the *brain mRNA* dataset (*XGBoost genes*) mentioned above. More specifically, a hypergeometric test was performed to determine which *gene sets* were enriched among the ML genes used to separate MDD cases from controls. We applied the *Benjamini–Hochberg* procedure to adjust the false-discovery rate (FDR) with alpha set to 0.1.

Results

For the *brain mRNA* dataset, the *training* set consisted of 103 MDD cases and 80 controls. The *testing* set consisted of 23 MDD cases and 23 controls. For the *blood mRNA* dataset, the *training* set consisted of 1267 MDD cases and 293 controls. The *testing* set consisted of 314 MDD cases and 76 controls. We performed a comparison of the subject covariates in the *training*

vs. *testing* sets and confirmed that the training and testing datasets are adequately similar (Supplemental Tables 1 and 2 for the *brain mRNA* and *blood mRNA* datasets, respectively).

Brain mRNA dataset

We obtained the best classification models based on our model selection procedure. For the *brain mRNA* dataset, the best model trained using only *gene expression data* for discriminating MDD cases from controls had an average AUC of 0.72 (standard deviation (SD): 0.10) over 10 cross-validation folds. In contrast, the best baseline model trained using the same dataset but with randomly permuted labels had an average AUC of 0.55 (SD: 0.12). Comparison between the 10-fold cross-validation AUCs of the trained and baseline models with the Wilcoxon signed-rank test showed that the performance of the trained model was significantly greater than that of the baseline model (p-value: 0.0048), suggesting that the performance was greater than expected by chance. After refitting the best model based on the entire dataset, a total of 62 genes were utilized in the final model (Supplemental Table 3). The final model achieved an AUC of 0.76 on the testing set (Figure 1A). After dividing the predictions into binary classes based on the optimal ROC cutoff estimated from cross-validation, the actual classification results are shown in a confusion matrix (Figure 1B). The overall balanced accuracy of all testing samples was 67%, with the most confident predictions having a balanced accuracy of around 85% (Figure 1C). The training AUC, baseline AUC, p-value, and testing set AUC of the *external brain mRNA* dataset are shown in Table 1.

The differences in covariates between MDD cases and controls are summarized in Supplemental Table 4. We did not identify any major differences between correctly and incorrectly classified MDD cases and controls in each of the subgroups (Supplemental Table 5).

The training AUC, baseline AUC, p-value, and testing set AUC of the *brain mRNA* dataset are shown in Table 1.

Next, we compared a classification model trained using covariates only. The *covariates only* model had an average 10-fold cross-validation AUC of 0.83 (SD: 0.075). This is consistent with the significant differences noted between MDD subjects and controls, in terms of their covariates (Supplemental Table 4). However, when we trained a model using *covariates and gene expressions* we observed an average 10-fold cross-validation AUC of 0.71 (SD: 0.087). We compared the AUC obtained on each cross-validation fold to determine whether the AUCs obtained from each of the models and found that the *covariates only* model performed higher than the *gene expressions only* and *covariates and gene expressions* models (p-values of 0.049 and 0.0039, respectively). Of note, adding covariates did not have a significant impact on classification performance. We observed no difference between the *gene expressions only* versus *covariates and gene expressions* models (p-value of 0.77).

External brain mRNA dataset

We evaluated the ML model on the *external brain mRNA* dataset, after training a logistic regression model on the *brain mRNA* dataset, using only the *genes* of the *XGBoost classifier*. Only 49 of the genes in the *external brain mRNA* dataset overlapped with the original *XGBoost genes* (i.e. 79% of the *XGBoost classifier genes*). The best cross-validated logistic regression model based on the *brain mRNA* dataset and these 49 genes had an average 10-fold cross-validation AUC of 0.91 (SD: 0.049). After refitting on the whole *brain mRNA* dataset, the AUC performance of the classifier on the *external brain mRNA* dataset was found to be 0.62 (Figure 2A). The overall balanced accuracy value was 62% (Figure 2B), with more confident predictions

having a balanced accuracy of around 72% (Figure 2C). To understand the directional effect of each gene, we plotted the logistic regression coefficients of each gene in the model (Figure 3). The training AUC, baseline AUC, p-value, and testing set AUC of the *external brain mRNA* dataset are shown in Table 1.

Gene set analysis

Lastly, we applied the GSA method to find which GO molecular functions *gene sets* exhibit different expression patterns for the *brain mRNA* cases vs. controls. The *piano* consensus GSA was used to prioritize GO molecular functions *gene sets* (n=1645). Forty *gene sets* were prioritized using the *brain mRNA* data. We then determined which of the prioritized *gene sets* were significantly enriched in our set of ML genes via the hypergeometric distribution p-values and FDR correction with alpha set to 0.1. *Metalloaminopeptidase activity* (mixed-directional up-regulated in MDD; FDR: 0.019)*, oxidoreductase activity acting on a heme group of donors* (distinct directional down-regulated in MDD; FDR: 0.019)*,* and *aminopeptidase activity* (mixed-directional up-regulated in MDD; FDR: 0.029) were found to be significant *gene sets*. The ML genes that overlap with these *gene sets* are *ENPEP*, *COX6A1* (Supplemental Table 6). The dysregulation directionalities of the *gene sets* are also consistent with the directionalities of the gene coefficients for *ENPEP* (positive coefficient) and *COX6A1* (negative coefficient) (Figure 3).

Blood mRNA dataset

We then repeated the gene expression classification analyses with the *blood mRNA* dataset for discriminating all 1581 MDD cases from 369 controls. We found that the best trained model achieved an average cross-validation AUC of 0.64 (SD: 0.041). In comparison, the best

baseline model achieved an AUC of 0.56 (SD: 0.030). The Wilcoxon signed-rank test showed that the performance of the best trained model was significantly better than the baseline model (p-value: 0.0020). After refitting on the whole training set, the model utilized a total of 1376 genes (Supplemental Table 7) and achieved an AUC of 0.61 (Figure 4A). The overall balanced accuracy was 56% on the testing set (Figure 4B), with the most confident predictions having a balanced accuracy of around 60% (Figure 4C). We summarized the differences in covariates between MDD cases and controls in Supplemental Table 8. When performing the covariate subgroups analysis, we observed a significant interaction between classification performance and subjects' *smoker status* (p-value: 0.0024; Supplemental Table 9), in which a correct prediction was more likely given an MDD case was a smoker (58% (79/136) correct predictions conditioned on MDD case being a smoker, vs. 43% (77/178) correct conditioned on MDD case being a non-smoker). Furthermore, no major difference was observed for controls (60% correct for smokers vs. 64% correct for non-smokers). The values for the training AUC, baseline AUC, p-value, and testing set AUC of the *blood mRNA* dataset are shown in Table 1.

Discussion

ML classification performance

We were able to successfully discriminate between MDD cases and controls using post-mortem *brain mRNA* data (AUC:0.72). Furthermore, based on the gene features identified, we performed external validation using DLPFC gene expression values from an independent cohort of matched MDD cases (n=29) and controls (n=29) (AUC:0.62). We observed a lower performance on the *external brain mRNA* dataset. This should be expected due to the differences in array platforms (Illumina vs. Affymetrix) used to acquire the gene expression data between

the datasets, as well as differences in sample preparation. Of note, although the overall balanced accuracy value for the *external brain mRNA* dataset was 62% (Figure 2B), more confident predictions had a balanced accuracy of around 72% (Figure 2C).

Overall, our model evaluations provide a more reliable and conservative measure of ML performance compared with prior studies (Bhak et al., 2019; Guilloux et al., 2015; Khodayari-Rostamabad, Reilly, Hasey, Debruin, & Maccrimmon, 2010; Yi et al., 2012; Yu et al., 2016). Specifically, previously published studies perform feature selection, often through DEG analysis, and training of the ML classifier based on the set of preselected features, on the same dataset (Bhak et al., 2019; Yi et al., 2012). Further, evaluation of ML performance is through only a k-fold cross-validations step, without a final held-out testing set (which was not used during model selection), or an independent cohort (Bhak et al., 2019; Khodayari-Rostamabad et al., 2010; Yi et al., 2012; Yu et al., 2016). Although the performance measures from cross-validations on the same dataset reflect internal validation for a classifier, they do not reflect the generalizability or replicability of the model, which requires external validation with fully independent data (Steyerberg & Harrell, 2016). An independent validation MDD cohort was previously only included in the study by Guilloux et al. (2015). However, even in this study, the reported prediction performance was based on cross-validation in the validation cohort (i.e., training a model and evaluating on the same validation cohort), and not truly external validation, which requires that the data not be used during the model training process.

In our study, not only did we calculate the AUC of the classifier on a previously unused testing-set of our dataset (i.e. a subset of data which was used during any training or cross-validation process), but we also reported the model performance based on external validation on an independent cohort. This suggests that our findings represent a more accurate reflection of

model generalization, as compared to past studies. Moreover, our study utilized datasets with relatively larger sample sizes (n=229 for *brain mRNA*, n=58 for *external brain mRNA*, and n=1950 for *blood mRNA*) compared with the similar studies (n=182 for Bhak et al. (2019); n=64 for Yu et al. (2016); n=24 for Yi et al. (2012)) previously published. Sample size is important in ML, as it can help prevent overfitting and improve the generalizability of the findings. Another point that sets our study apart is that when classifying MDD subjects and controls, we explored the use of covariate data, in addition to gene expression data. Using only the covariate data, the ML classification performance was higher when compared with using only gene expression data. This comes as no surprise given the significant differences between MDD subjects and controls on most of the covariates (Supplemental Table 4). However, we did not observe an increase in classification performance when covariates were added to gene expressions, as compared to using only gene expressions. This could be due to the fact that the XGBoost algorithm does not perform an exhaustive search over all possible splits when selecting features in constructing a tree. Rather, it uses an *approximate* algorithm for split finding during tree learning, which attempts to find an approximate split for each feature (Chen & Guestrin). Thus, it is possible that the approximate splits for the covariates were less optimal, when compared with the much larger number of gene features with more optimal approximations, and were thus not selected to be incorporated into the model.

However, our covariate subgroups analysis using the *blood mRNA* dataset, revealed a significant interaction between *smoker status* and a higher recall performance for MDD (58% conditioned on smoker cases vs. 43% conditioned on non-smoker cases). There is a known positive association between smoking and depression (Fluharty, Taylor, Grabski, & Munafò, 2017), with recent Mendelian randomization studies supporting a causal effect of smoking on

depression (Wootton et al., 2020; Yuan, Yao, & Larsson, 2020). Smoking is also known to influence the expression of several genes (Kopa & Pawliczak, 2018). By conditioning on the *smoker status* variable, we have eliminated/controlled for the confounding effects of *smoker status*. Thus, combined with current literature, our finding of higher recall performance in the smoker subgroup of MDD cases suggest that there may be specific differences in gene expressions in MDD cases who are smokers, as compared with the ones who are not. However, since *smoker status* is controlled for, these differences are *not* due to the effect of smoking. They may characterize a subtype of MDD patients with distinct (shared) pathophysiology.

Lessons on pathophysiology

To advance our knowledge of genes which are important in the pathophysiology of MDD, we implemented a novel approach using the *brain mRNA* dataset. We highlighted the *GO molecular function gene sets* that are important based on the differentially expressed genes, as well as, based on the genes identified through the ML classifier. First, we performed a univariate differential gene expression analysis contrasting MDD cases and controls based on all genes in the dataset, while adjusting for covariates. We prioritized the ten top-ranked *gene sets* in each directionality category based on the robust consensus *gene set* analysis method from the *piano* bioinformatic tool. The differential gene expression with *piano* approach has the benefit of utilizing the full set of gene expressions available; however, this approach also results in a lot of *gene sets* that may not be relevant to MDD. To address this, we re-ranked the *gene sets* using the genes identified through the ML classifier, based on a hypergeometric test and FDR adjustment. This approach allowed us to focus on the most important gene sets which are significantly differentially expressed based on both the univariate analysis using all genes, as well as the ones significantly enriched among the ML genes. Such robust genes can be important in the

pathophysiology of MDD and the identification of novel candidate treatments. Based on this approach, we highlight the role of over-expression of the *metalloaminopeptidase activity, oxidoreductase activity acting on a heme group of donors,* and *aminopeptidase activity* in MDD pathophysiology. The specific genes from the ML model that overlap with these *gene sets* are *ENPEP* and *COX6A1*.

Our review of the literature revealed prior evidence for a link of these two key genes to MDD, supporting that the molecular processes mediated by these genes may be relevant for MDD. The *ENPEP* gene encodes glutamyl aminopeptidase, which converts angiotensin II to angiotensin III for up-regulating blood pressure as a major part of the renin-angiotensin system (RAS) (Holmes, Spradling-Reeves, & Cox, 2017). Interestingly, the *RAS* has been proposed to be a potential drug target in depression, with several studies finding angiotensin-converting enzyme inhibitors and angiotensin receptor blockers to be effective in depression (Vian et al., 2017). This is consistent with our findings of an up-regulation of *ENPEP* in MDD cases. The other gene that was highlighted with our novel approach, namely *COX6A1*, encodes a subunit of the cytochrome *c* oxidase (COX). COX is involved in the oxidative phosphorylation (OXPHOS) process in ATP production. Lowered activity of COX and other defects in the OXPHOS process leading to lowered ATP production have been reported in both patients with depression and relevant animal models (Allen, Romay-Tallon, Brymer, Caruncho, & Kalynchuk, 2018).

We also performed a literature review on the top 20 genes, which were identified based on the *brain mRNA* dataset (Supplemental Table 3). Overall, we have identified links to MDD in literature for 11 genes (*CX3CR1, TMEM245, COL4A1, PRAMEF1, TMEM52, A2M, DDC-AS1, GRP88, GALR3, VPS53, CRYBA1*), in addition to *ENPEP* and *COX6A1* mentioned above. *CX3CR1* and *A2M* had the most literature reports linked to MDD. *CX3CR1* encodes the C-X3-C

Motif Chemokine Receptor 1 and has been reported to be up-regulated in microglia cells of MDD patients (Snijders et al., 2020), as well as, in single-cell analysis of CD11b cells of MDD patients (Böttcher et al., 2020). In a mouse model with lipopolysaccharide- induced depression, increased depression-like behavior was observed in *CX3CR1* knockout mice vs. control mice (Corona et al., 2010). The finding from this study is consistent with the down-regulation of *CX3CR1* in the DLPFC of MDD patients noted in our study (Figure 3). Our study also shows a decrease in A2M expression in MDD patients. *A2M* encodes Alpha-2-Macroglobulin and has been previously identified as a candidate MDD susceptibility gene. Genetic polymorphisms within the gene were found to be significantly associated with MDD (Zhao et al., 2020). Similarly, there are reports of increased A2M expression in whole blood of MDD patients (Cattaneo et al., 2020). Other genes worth highlighting include *CRYBA1, GALR3*, *GPR88,* and *DDC-AS1*. The largest GWAS study conducted for MDD has identified a significant variant located within the *CRYBA1* gene (Wray et al., 2018). *GALR3* encodes for Galanin Receptor Type 3, and this pathway has been linked to depression (Kuteeva et al. (2008). *GPR88* (Del Zompo et al. (2014); Logue et al. (2009) and *DDC-AS1*, a long non-coding RNA in the antisense direction of the DDC gene (Giardina et al., 2011) (Børglum et al., 1999) have been linked to mood disorders and to the metabolism of different neurotransmitters (dopamine and serotonin).

Conclusion and Future Directions

In conclusion, we have shown that ML analysis of gene expression data could effectively distinguish MDD cases from controls and further our understanding of the pathophysiology of MDD. Our results support that the genes identified based on the *brain mRNA* dataset are important for MDD. However, we need larger sample sizes to account for the heterogeneity of

MDD and allow for precision medicine. In support of this statement, although the overall balanced accuracy value for both the *brain mRNA* dataset and the *external brain mRNA* dataset are not ideal, both ML algorithm performance results were significant, and more confident predictions had a higher balanced accuracy. The findings from *brain mRNA* data may be more useful for the development of new treatment options, than those from *blood mRNA* data, given their relevance to MDD pathophysiology. Although, our results using *blood mRNA* data were less good, the trained model performed significantly better than the baseline model. Moreover, our covariate subgroups analysis using the *blood mRNA* dataset, revealed a significant interaction between *smoker status* and a higher recall performance for MDD. The differences at the transcriptomic level noted in this subgroup of patients are *not* due to the effect of smoking, given *smoker status* was controlled for in this analysis. Our results, along with the results from Guilloux et al. (2015), suggest that *blood mRNA*-based ML models are also very promising, especially when analyzed along with covariate data. They can potentially serve as a valuable tool for precision medicine in MDD with regards to identifying subtypes of patients with unique pathophysiology, and for informing diagnosis, prognosis, treatment selection and response monitoring.

With regards to future directions, it would be interesting to perform an ML analysis using transcriptomics, along with other laboratory data. Several ML studies in MDD have used magnetic resonance imaging (MRI) based datatypes (S. Gao, Calhoun, & Sui, 2018). A study in 2010 proposed an ML method which used successfully the patient's pre-treatment electroencephalogram (EEG) to predict the individual's response to selective serotonin reuptake inhibitors (SSRIs) (Khodayari-Rostamabad et al., 2010). Bhak et al. (2019) utilized a multi-omics approach based on ML applied on blood transcriptomic and methylomic data, combined,

to distinguish between MDD cases, suicide attempters and controls. Combining multi-omics data with brain MRI and EEG data in a future ML study could lead to clinically useful results.

Acknowledgments

We thank the authors and dbGaP, as well as the GEO repository, for access to the dataset. We would also like to thank Dr. Celia Greenwood and Dr. Jeff Xia for their suggestions and feedback on the methodology and results.

Contributors

Bill Qi performed the bioinformatic and machine learning analyses and drafted the manuscript under the supervision of Yannis Trakadis who conceived and coordinated the project. Bill Qi and Yannis Trakadis designed the original methodology. Imane Bennani and Janani Ramamurthy performed the background literature review. All authors reviewed and provided feedback on the manuscript.

Availability of data and materials

The datasets used in the preparation of this manuscript were obtained from the Database of Genotypes and Phenotypes (dbGaP) after McGill IRB approval. Raw dbGaP data used is available in the studies phs000979.v1.p1 and phs000486.v1.p1. The Gene Expression Omnibus (GEO) datasets are publicly available in the repositories GSE54567 and GSE54568.

Role of funding source

Yannis Trakadis is supported by the McGill University Health Centre Research Institute and the Canada First Research Excellence Fund (McGill University Healthy Brains for Healthy Lives Initiative). The funding source *had no further role in study design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the paper for publication.*

Conflict of interest

None

References

Allen, J., Romay-Tallon, R., Brymer, K. J., Caruncho, H. J., & Kalynchuk, L. E. (2018). Mitochondria and Mood: Mitochondrial Dysfunction as a Key Player in the Manifestation of Depression. *Frontiers in Neuroscience, 12*, 386. doi:10.3389/fnins.2018.00386

Bhak, Y., Jeong, H.-o., Cho, Y. S., Jeon, S., Cho, J., Gim, J.-A., . . . Lee, S. (2019). Depression and suicide risk prediction models using blood-derived multi-omics data. *Translational Psychiatry, 9*(1), 262. doi:10.1038/s41398-019-0595-2

Børglum, A. D., Bruun, T. G., Kjeldsen, T. E., Ewald, H., Mors, O., Kirov, G., . . . Kruse, T. A. (1999). Two novel variants in the DOPA decarboxylase gene: association with bipolar affective disorder. *Molecular Psychiatry, 4*(6), 545-551. doi:10.1038/sj.mp.4000559

Böttcher, C., Fernández-Zapata, C., Snijders, G. J. L., Schlickeiser, S., Sneeboer, M. A. M., Kunkel, D., . . . Priller, J. (2020). Single-cell mass cytometry of microglia in major

depressive disorder reveals a non-inflammatory phenotype with increased homeostatic marker expression. *Translational Psychiatry, 10*(1), 310. doi:10.1038/s41398-020-00992-2

Cattaneo, A., Ferrari, C., Turner, L., Mariani, N., Enache, D., Hastings, C., . . . Pariante, C. M. (2020). Whole-blood expression of inflammasome- and glucocorticoid-related mRNAs correctly separates treatment-resistant depressed patients from drug-free and responsive patients in the BIODEP study. *Transl Psychiatry, 10*(1), 232. doi:10.1038/s41398-020-00874-7

Chang, L. C., Jamain, S., Lin, C. W., Rujescu, D., Tseng, G. C., & Sibille, E. (2014). A conserved BDNF, glutamate- and GABA-enriched gene module related to human depression identified by coexpression meta-analysis and DNA variant genome-wide association studies. *PloS One, 9*(3), e90980. doi:10.1371/journal.pone.0090980

Chen, T., & Guestrin, C. (2016). *Xgboost: A scalable tree boosting system*.

Corona, A. W., Huang, Y., O'Connor, J. C., Dantzer, R., Kelley, K. W., Popovich, P. G., & Godbout, J. P. (2010). Fractalkine receptor (CX3CR1) deficiency sensitizes mice to the behavioral changes induced by lipopolysaccharide. *Journal of Neuroinflammation, 7*, 93. doi:10.1186/1742-2094-7-93

Del Zompo, M., Deleuze, J. F., Chillotti, C., Cousin, E., Niehaus, D., Ebstein, R. P., . . . Meloni, R. (2014). Association study in three different populations between the GPR88 gene and major psychoses. *Mol Genet Genomic Med, 2*(2), 152-159. doi:10.1002/mgg3.54

Fluharty, M., Taylor, A. E., Grabski, M., & Munafò, M. R. (2017). The Association of Cigarette Smoking With Depression and Anxiety: A Systematic Review. *Nicotine Tob Res, 19*(1), 3-13. doi:10.1093/ntr/ntw140

Gao, L., Gao, Y., Xu, E., & Xie, J. (2015). Microarray Analysis of the Major Depressive Disorder mRNA Profile Data. *Psychiatry Investigation, 12*(3), 388-396. doi:10.4306/pi.2015.12.3.388

Gao, S., Calhoun, V. D., & Sui, J. (2018). Machine learning in major depression: From classification to treatment outcome prediction. *CNS Neuroscience & Therapeutics, 24*(11), 1037-1052. doi:10.1111/cns.13048

Giardina, G., Montioli, R., Gianni, S., Cellini, B., Paiardini, A., Voltattorni, C. B., & Cutruzzolà, F. (2011). Open conformation of human DOPA decarboxylase reveals the mechanism of PLP addition to Group II decarboxylases. *Proceedings of the National Academy of Sciences of the United States of America, 108*(51), 20514-20519. doi:10.1073/pnas.1111456108

Guilloux, J. P., Bassi, S., Ding, Y., Walsh, C., Turecki, G., Tseng, G., . . . Sibille, E. (2015). Testing the predictive value of peripheral gene expression for nonremission following citalopram treatment for major depression. *Neuropsychopharmacology, 40*(3), 701-710. doi:10.1038/npp.2014.226

Holmes, R. S., Spradling-Reeves, K. D., & Cox, L. A. (2017). Mammalian Glutamyl Aminopeptidase Genes (ENPEP) and Proteins: Comparative Studies of a Major Contributor to Arterial Hypertension. *J Data Mining Genomics Proteomics, 8*(2). doi:10.4172/2153-0602.1000211

Khodayari-Rostamabad, A., Reilly, J. P., Hasey, G., Debruin, H., & Maccrimmon, D. (2010). Using pre-treatment EEG data to predict response to SSRI treatment for MDD. *Conference Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2010*, 6103-6106. doi:10.1109/iembs.2010.5627823

Koenigs, M., & Grafman, J. (2009). The functional neuroanatomy of depression: distinct roles for ventromedial and dorsolateral prefrontal cortex. *Behavioural Brain Research, 201*(2), 239-243. doi:10.1016/j.bbr.2009.03.004

Kopa, P. N., & Pawliczak, R. (2018). Effect of smoking on gene expression profile - overall mechanism, impact on respiratory system function, and reference to electronic cigarettes. *Toxicol Mech Methods, 28*(6), 397-409. doi:10.1080/15376516.2018.1461289

Kuteeva, E., Wardi, T., Lundström, L., Sollenberg, U., Langel, U., Hökfelt, T., & Ogren, S. O. (2008). Differential role of galanin receptors in the regulation of depression-like behavior and monoamine/stress-related genes at the cell body level. *Neuropsychopharmacology, 33*(11), 2573-2585. doi:10.1038/sj.npp.1301660

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

Liew, C. C., Ma, J., Tang, H. C., Zheng, R., & Dempsey, A. A. (2006). The peripheral blood transcriptome dynamically reflects system wide biology: a potential diagnostic tool. *Journal of Laboratory and Clinical Medicine, 147*(3), 126-132. doi:10.1016/j.lab.2005.10.005

Logue, S. F., Grauer, S. M., Paulsen, J., Graf, R., Taylor, N., Sung, M. A., . . . Pausch, M. (2009). The orphan GPCR, GPR88, modulates function of the striatal dopamine system: a possible therapeutic target for psychiatric disorders? *Mol Cell Neurosci, 42*(4), 438-447. doi:10.1016/j.mcn.2009.09.007

Smyth, G. K. (2005). Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor* (pp. 397-420): Springer.

Snijders, G., Sneeboer, M. A. M., Fernández-Andreu, A., Udine, E., Boks, M. P., Ormel, P. R., . . . de Witte, L. D. (2020). Distinct non-inflammatory signature of microglia in post-mortem brain tissue of patients with major depressive disorder. *Molecular Psychiatry*. doi:10.1038/s41380-020-00896-z

Steyerberg, E. W., & Harrell, F. E., Jr. (2016). Prediction models need appropriate internal, internal-external, and external validation. *Journal of Clinical Epidemiology, 69*, 245-247. doi:10.1016/j.jclinepi.2015.04.005

Sullivan, P. F., Fan, C., & Perou, C. M. (2006). Evaluating the comparability of gene expression in blood and brain. *American Journal of Medical Genetics. Part B: Neuropsychiatric Genetics, 141b*(3), 261-268. doi:10.1002/ajmg.b.30272

Trakadis, Y. J., Sardaar, S., Chen, A., Fulginiti, V., & Krishnan, A. (2018). Machine learning in schizophrenia genomics, a case-control study using 5,090 exomes. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 180*(2), 103-112.

Varemo, L., Nielsen, J., & Nookaew, I. (2013). Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Research, 41*(8), 4378-4391. doi:10.1093/nar/gkt111

Vian, J., Pereira, C., Chavarria, V., Köhler, C., Stubbs, B., Quevedo, J., . . . Fernandes, B. S. (2017). The renin-angiotensin system: a possible new target for depression. *BMC Medicine, 15*(1), 144. doi:10.1186/s12916-017-0916-3

Wigge, L. V., & Nookaew, I. (2020). Platform for Integrative Analysis of Omics data. Retrieved from https://bioconductor.org/packages/release/bioc/vignettes/piano/inst/doc/piano-vignette.pdf

Woo, H. I., Lim, S. W., Myung, W., Kim, D. K., & Lee, S. Y. (2018). Differentially expressed genes related to major depressive disorder and antidepressant response: genome-wide gene expression analysis. *Experimental and Molecular Medicine, 50*(8), 92. doi:10.1038/s12276-018-0123-0

Wootton, R. E., Richmond, R. C., Stuijfzand, B. G., Lawn, R. B., Sallis, H. M., Taylor, G. M. J., . . . Munafò, M. R. (2020). Evidence for causal effects of lifetime smoking on risk for depression and schizophrenia: a Mendelian randomisation study. *Psychological Medicine, 50*(14), 2435-2443. doi:10.1017/s0033291719002678

Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., . . . Sullivan, P. F. (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature Genetics, 50*(5), 668-681. doi:10.1038/s41588-018-0090-3

Yi, Z., Li, Z., Yu, S., Yuan, C., Hong, W., Wang, Z., . . . Fang, Y. (2012). Blood-based gene expression profiles models for classification of subsyndromal symptomatic depression and major depressive disorder. *PloS One, 7*(2), e31283. doi:10.1371/journal.pone.0031283

Yu, J. S., Xue, A. Y., Redei, E. E., & Bagheri, N. (2016). A support vector machine model provides an accurate transcript-level-based diagnostic for major depressive disorder. *Transl Psychiatry, 6*(10), e931. doi:10.1038/tp.2016.198

Yuan, S., Yao, H., & Larsson, S. C. (2020). Associations of cigarette smoking with psychiatric disorders: evidence from a two-sample Mendelian randomization study. *Scientific Reports, 10*(1), 13807. doi:10.1038/s41598-020-70458-4

Zhao, M., Chen, L., Qiao, Z., Zhou, J., Zhang, T., Zhang, W., . . . Yang, X. (2020). Association Between FoxO1, A2M, and TGF-β1, Environmental Factors, and Major Depressive Disorder. *Front Psychiatry, 11*, 675. doi:10.3389/fpsyt.2020.00675

Figures and Tables

Figure 1. Brain mRNA testing set results for discriminating MDD cases vs. controls.

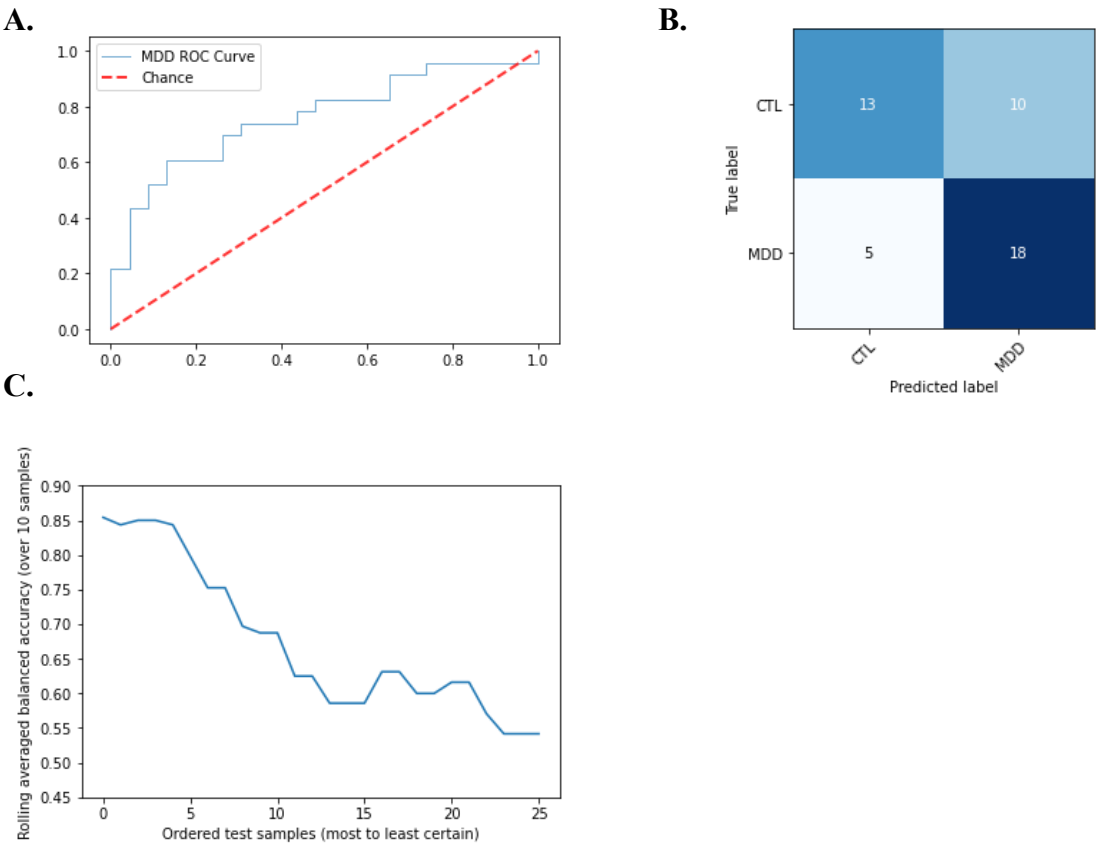


Figure 1. After obtaining the best trained classification model for discriminating between MDD cases vs. controls from the *brain mRNA* dataset, several metrics were used to assess the performance of the model on a previously unseen testing set (20% of the full dataset). Figure 1A shows the ROC curve for model predictions. The x-axis of the curve shows the false-positive rate (FPR) and the y-axis shows the TPR for a given probability cutoff threshold. The area under the

ROC is 0.76, suggesting that the model performs above random chance (red dashed line). Figure 1B shows the actual classifications based on the best positive-negative class cutoff threshold estimated from cross-validation during training. The overall balanced accuracy was 67%. Figure 1C shows the prediction-confidence-ranked rolling balanced accuracy plot, which shows the trend of balanced classification accuracy (i.e., average of sensitivity and specificity) of the most confident subset of predictions to the least confident subset of predictions (x-axis left to right) for the testing set samples. The rolling window size *n* is set to 10 for calculating the balanced accuracies. The y-axis shows the balanced accuracy value for each subset of predictions. The most-confident predictions generally have a balanced accuracy of ~85%.

CTL: Control; FPR: false-positive rate; MDD: Major depression;

ROC: receiver-operating characteristics curve; TPR: true-positive rate

Figure 2. External brain mRNA testing set results for discriminating MDD cases vs. controls.

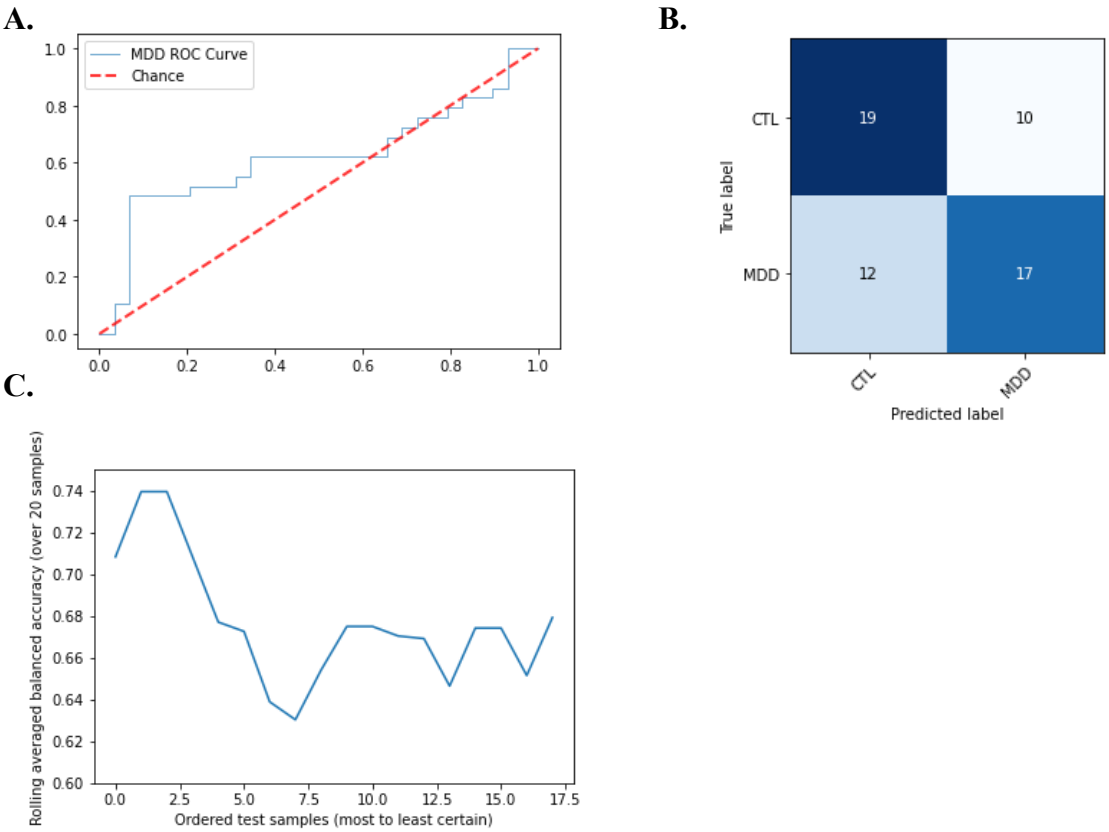


Figure 2. After training a logistic regression model for discriminating between MDD cases vs. controls using the *brain mRNA* dataset, several metrics were used to assess the performance of the model on the *external brain mRNA* dataset testing set (n=58). Figure 2A shows the ROC curve for model predictions. The x-axis of the curve shows the FPR and the y-axis shows the TPR for a given probability cutoff threshold. The area under the ROC is 0.62, suggesting that the model performs above random chance (red dashed line). Figure 2B shows the actual classifications based on the best positive-negative class cutoff threshold estimated from cross-validation during training. The overall balanced accuracy was 62%. Figure 2C shows the prediction-confidence-ranked rolling balanced accuracy plot, which shows the trend of balanced

classification accuracy (i.e., average of sensitivity and specificity) of the most confident subset of predictions to the least confident subset of predictions (x-axis left to right) for the testing set samples. The rolling window size *n* is set to 20 for calculating the balanced accuracies. The y-axis shows the balanced accuracy value for each subset of predictions. The most-confident predictions generally have a balanced accuracy of ~72%.

CTL: Control; FPR: false-positive rate; MDD: Major depression;

ROC: receiver-operating characteristics curve; TPR: true-positive rate

Figure 3. Logistic regression coefficients for gene features.

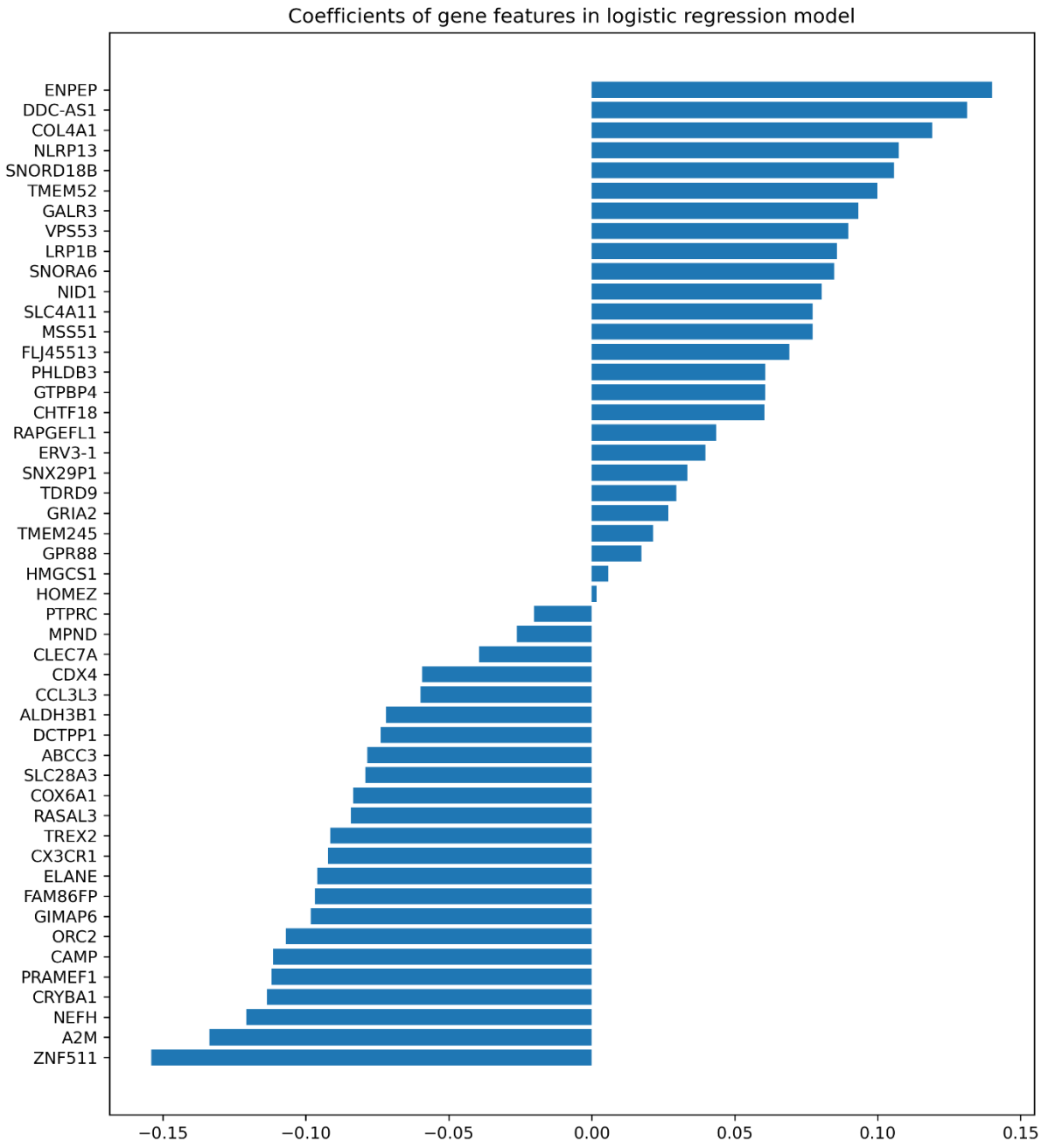


Figure 3. A logistic regression model was trained based on the XGBoost prioritized genes using the *brain mRNA* dataset. Evaluation of the model was performed using the *external brain mRNA* dataset. The coefficients of each gene features in the final trained model were extracted and plotted. A positive coefficient indicates that an increase in the gene expression will increase the

likelihood for MDD. A negative coefficient indicates that a decrease in the gene expression will increase the likelihood for MDD.

Figure 4. Blood mRNA testing set results for discriminating MDD cases vs. controls.

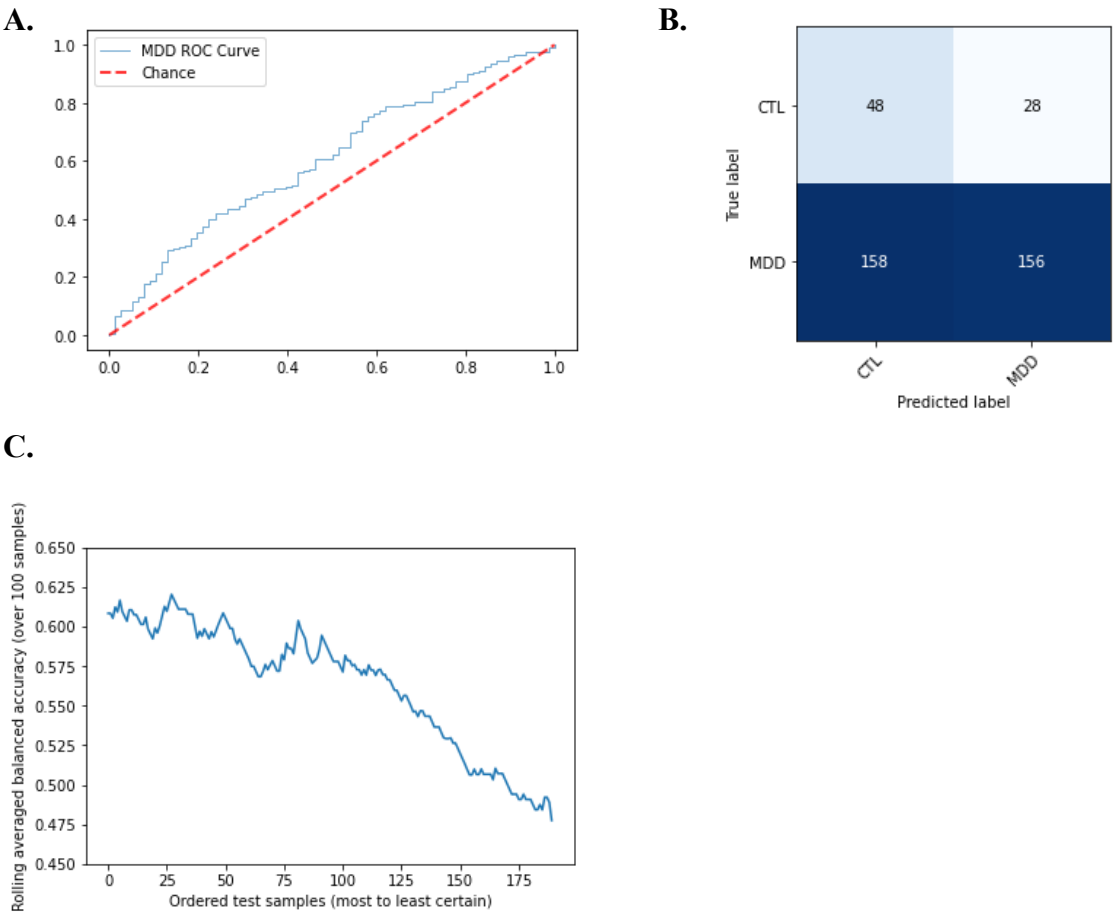


Figure 4. After obtaining the best trained classification model for discriminating between MDD cases vs. controls from the *blood mRNA* dataset, several metrics were used to assess the performance of the model on a previously unseen testing set (20% of the full dataset). Figure 4A shows the ROC curve for model predictions. The x-axis of the curve shows the FPR and the y-axis shows the TPR for a given probability cutoff threshold. The area under the ROC is 0.61, suggesting that the model performs above random chance (red dashed line). Figure 4B shows the

actual classifications based on the best positive-negative class cutoff threshold estimated from cross-validation during training. The overall balanced accuracy was 56%. Figure 4C shows the prediction-confidence-ranked rolling balanced accuracy plot, which shows the trend of balanced classification accuracy (i.e., average of sensitivity and specificity) of the most confident subset of predictions to the least confident subset of predictions (x-axis left to right) for the testing set samples. The rolling window size *n* is set to 100 for calculating the balanced accuracies. The y-axis shows the balanced accuracy value for each subset of predictions. The most-confident predictions generally have a balanced accuracy of ~60%.

CTL: Control; FPR: false-positive rate; MDD: Major depression;

ROC: receiver-operating characteristics curve; TPR: true-positive rate

Table 1. Model cross-validation and testing set AUC scores (gene expression data models).

Classification task	**Mean AUC (SD) of trained model from cross-validation**	**Mean AUC (SD) of baseline model from cross-validation**	**Wilcoxon signed-rank test p-values**	**Testing set AUC for retrained model**
Brain mRNA – MDD vs. controls	0.72 (0.10)	0.55 (0.12)	0.0048	0.76
Blood mRNA – All MDD vs. controls	0.64 (0.041)	0.56 (0.030)	0.0020	0.61

Table 1. This table summarizes the average area under the receiver-operating characteristics curve (AUC) from 10-fold cross-validation for the best trained models and the best baseline models for each of the classification analyses. The Wilcoxon signed-rank test was used to compare the AUC from the trained model against the baseline model. P-values from these comparisons are shown. After retraining the best model on the full training set, an evaluation was conducted on the testing set and the resulting AUCs are listed.

Supplemental Figures and Tables

Supplemental Table 1. Summary of the covariates between the training and testing sets from the *brain mRNA* dataset.

	Training	Testing	Test statistic	P-value
Diagnosis			0.36	0.55
Diagnosis (Control)	80	23		
Diagnosis (MDD)	103	23		
Sex			0.092	0.76
Sex (F)	53	15		
Sex (M)	130	31		
Smoker status			0.19	0.91
Smoker status (No)	106	25		
Smoker status (Unknown)	18	5		
Smoker status (Yes)	59	16		
Alcohol status			0.0068	0.93
Alcohol status (Negative)	126	33		
Alcohol status (Positive)	21	6		
Antipsychotic status			1.38	0.50
Antipsychotic status (Negative)	115	29		
Antipsychotic status (Not Tested)	24	9		
Antipsychotic status (Positive)	8	1		
Antidepressant status			1.27	0.53
Antidepressant status (Negative)	92	23		
Antidepressant status (Not Tested)	23	9		
Antidepressant status (Positive)	32	7		
Mood stabilizer status			1.29	0.52
Mood stabilizer status (Negative)	110	29		
Mood stabilizer status (Not Tested)	27	9		
Mood stabilizer status (Positive)	10	1		
Benzodiazepine status			3.60	0.17
Benzodiazepine status (Negative)	124	30		
Benzodiazepine status (Not Tested)	9	6		
Benzodiazepine status (Positive)	14	3		
Nicotine/cotinine status			2.33	0.31
Nicotine/cotinine status (Negative)	95	20		
Nicotine/cotinine status (Not Tested)	19	7		

Nicotine/cotinine status (Positive)	33	12		
THC status			1.04	0.59
THC status (Negative)	118	31		
THC status (Not Tested)	28	7		
THC status (Positive)	1	1		
Cocaine status			1.09	0.58
Cocaine status (Negative)	136	37		
Cocaine status (Not Tested)	4	0		
Cocaine status (Positive)	7	2		
Opiates status			1.89	0.39
Opiates status (Negative)	124	36		
Opiates status (Not Tested)	3	0		
Opiates status (Positive)	20	3		
Continuous variables				
Age	44.71	43.65	0.46	0.64
pH	6.46	6.44	0.50	0.62
Postmortem interval	32.68	29.98	0.91	0.36
Height	68.63	68.19	0.65	0.52
Weight	194.84	186.29	0.96	0.34

Supplemental Table 1. A summary of the covariate differences between the training and testing set subjects is presented. For the categorical variables, a count of subjects within each sub-category by diagnosis status is performed, and a chi-squared test is used to obtain a p-value for the contingency table for each categorical variable. For the continuous variables, the group means for the training and testing set subjects are recorded. An independent t-test is then performed to obtain p-values for the difference between the groups.

Supplemental Table 2. Summary of the covariates between the training and testing sets from the *blood mRNA* dataset.

	Training	Testing	Test statistic	P-value
Diagnosis			0.060	0.81
Diagnosis (Control)	293	76		
Diagnosis (MDD)	1267	314		
Sex			0.16	0.69
Sex (Female)	1037	264		
Sex (Male)	523	126		
Smoker status			0.059	0.81
Smoker status (No)	923	234		
Smoker status (Yes)	637	156		
Alcohol status			3.39	0.18
Alcohol status (No)	219	55		
Alcohol status (Unknown)	3	3		
Alcohol status (Yes)	1338	332		
Menostats			5.16	0.40
Menostats ((Recent) pregnancy)	69	13		
Menostats (Male/NA)	523	126		
Menostats (Menopause, natural)	7	0		
Menostats (Menopause, operation or disease)	288	69		
Menostats (Not in menopause)	656	180		
Menostats (Unknown)	17	2		
Education			1.75	0.46
Education (Basic)	107	23		
Education (High)	520	143		
Education (Intermediate)	933	224		
Continuous variables				
Age	42.26	41.44	1.13	0.26
BMI	25.80	25.51	1.01	0.31

Supplemental Table 2. A summary of the covariate differences between the training and testing set subjects is presented. For the categorical variables, a count of subjects within each sub-category by diagnosis status is performed, and a chi-squared test is used to obtain a p-value for the contingency table for each categorical variable. For the continuous variables, the group means for the training and testing set

subjects are recorded. An independent t-test is then performed to obtain p-values for the difference between the groups.

Supplemental Table 3. List of all 62 genes selected by XGBoost algorithm in the construction of the classifier for distinguishing MDD cases from controls for the *brain mRNA* dataset, ranked from most important to least important.

CX3CR1, TMEM245, SNORD103C, COL4A1, PRAMEF1, COX6A1, ENPEP, MSS51, LRP1B, TMEM52, A2M, DDC-AS1, TDRD9, GPR88, CCL3L3, GALR3, PHLDB3, FAM86FP, VPS53, CRYBA1, ERV3-1, RASAL3, GTPBP4, ZNF511, MPND, SNORA6, NEFH, SLC28A3, DCTPP1, RAPGEFL1, CLEC7A, MIR616, MIR519E, SLC4A11, FLJ45513, GIMAP6, LY6E-DT, DEFB133, CDX4, CAMP, SLC10A6, VENTXP7, ORC2, GRIA2, ELANE, HOMEZ, CHTF18, DEFB136, PTPRC, SNORD18B, MIR599, NLRP13, EP400P1, NID1, ALDH3B1, LIPN, SNX29P1, HMGCS1, SNORD114-17, TREX2, ABCC3, MIR1289-2

Supplemental Table 4. Summary of the covariates between MDD cases and controls from the *brain mRNA* dataset.

	MDD	**Control**	**Test statistic**	**P-value**
Sex			12.34	0.00044
Sex (F)	50	18		
Sex (M)	76	85		
Smoker status			33.34	5.74E-08
Smoker status (No)	51	80		
Smoker status (Unknown)	20	3		
Smoker status (Yes)	55	20		
Alcohol status			22.99	1.63E-06
Alcohol status (Negative)	59	100		
Alcohol status (Positive)	24	3		
Antipsychotic status			23.40	8.29E-06
Antipsychotic status (Negative)	69	75		
Antipsychotic status (Not Tested)	5	28		
Antipsychotic status (Positive)	9	0		
Antidepressant status			64.63	9.26E-15
Antidepressant status (Negative)	39	76		
Antidepressant status (Not Tested)	5	27		
Antidepressant status (Positive)	39	0		
Mood stabilizer status			24.34	5.19E-06
Mood stabilizer status (Negative)	68	71		
Mood stabilizer status (Not Tested)	5	31		
Mood stabilizer status (Positive)	10	1		
Benzodiazepine status			18.81	8.22E-05
Benzodiazepine status (Negative)	62	92		

Benzodiazepine status (Not Tested)	5	10		
Benzodiazepine status (Positive)	16	1		
Nicotine/cotinine status			38.07	5.40E-09
Nicotine/cotinine status (Negative)	35	80		
Nicotine/cotinine status (Not Tested)	25	1		
Nicotine/cotinine status (Positive)	23	22		
THC status			46.70	7.24E-11
THC status (Negative)	48	101		
THC status (Not Tested)	33	2		
THC status (Positive)	2	0		
Cocaine status			13.56	0.0011
Cocaine status (Negative)	71	102		
Cocaine status (Not Tested)	3	1		
Cocaine status (Positive)	9	0		
Opiates status			24.16	5.68E-06
Opiates status (Negative)	60	100		
Opiates status (Not Tested)	2	1		
Opiates status (Positive)	21	2		
Continuous variables				
Age	43.42	45.82	-1.30	0.19
pH	6.38	6.55	-4.56	8.31E-06
Postmortem interval	34.10	29.75	1.83	0.068
Height	67.78	69.17	-2.48	0.014
Weight	181.81	202.02	-2.78	0.0061

Supplemental Table 4. A summary of the covariate differences between MDD cases and controls is presented. For the categorical variables, a count of subjects within each sub-category by diagnosis status is performed, and a chi-squared test is used to obtain a p-value for the contingency table for each categorical variable. For the continuous variables, the group means for the MDD cases and controls are recorded. An independent t-test is then performed to obtain p-values for the difference between the groups.

Supplemental Table 5. Summary of the covariates between the correctly and incorrectly classified subjects from the *brain mRNA* testing set.

	Correct-MDD	Incorrect-MDD	Correct-control	Incorrect-control	Test statistic	P-value	BH-FDR
Sex					5.08	0.17	0.24
Sex (F)	9	2	2	2			
Sex (M)	9	3	11	8			
Smoker status					10.93	0.091	0.19
Smoker status (No)	6	2	11	6			
Smoker status (Unknown)	4	1	0	0			
Smoker status (Yes)	8	2	2	4			
Alcohol status					7.12	0.068	0.18
Alcohol status (Negative)	9	2	13	9			
Alcohol status (Positive)	3	2	0	1			
Antipsychotic status					8.61	0.20	0.24
Antipsychotic status (Negative)	10	4	7	8			
Antipsychotic status (Not Tested)	1	0	6	2			
Antipsychotic status (Positive)	1	0	0	0			
Antidepressant status					16.30	0.012	0.098
Antidepressant status (Negative)	6	2	7	8			
Antidepressant status (Not Tested)	1	0	6	2			
Antidepressant status (Positive)	5	2	0	0			
Mood stabilizer status					8.61	0.20	0.24
Mood stabilizer status (Negative)	10	4	7	8			
Mood stabilizer status (Not Tested)	1	0	6	2			
Mood stabilizer status (Positive)	1	0	0	0			
Benzodiazepine status					8.11	0.23	0.26
Benzodiazepine status (Negative)	9	3	9	9			
Benzodiazepine status (Not Tested)	1	0	4	1			
Benzodiazepine status (Positive)	2	1	0	0			
Nicotine/cotinine status					19.20	0.0038	0.061
Nicotine/cotinine status (Negative)	2	1	11	6			
Nicotine/cotinine status (Not Tested)	6	1	0	0			
Nicotine/cotinine status (Positive)	4	2	2	4			
THC status					10.80	0.095	0.19
THC status (Negative)	6	3	12	10			
THC status (Not Tested)	5	1	1	0			
THC status (Positive)	1	0	0	0			
Cocaine status					4.74	0.19	0.24
Cocaine status (Negative)	10	4	13	10			

Cocaine status (Positive)	2	0	0	0			
Opiates status					7.31	0.063	0.18
Opiates status (Negative)	9	4	13	10			
Opiates status (Positive)	3	0	0	0			
Continuous variables							
Age	39.72	40.60	44.69	50.90	1.82	0.16	0.24
pH	6.31	6.51	6.46	6.60	2.64	0.062	0.18
Postmortem interval	30.47	29.60	28.85	30.75	0.059	0.98	0.98
Height	67.02	65.88	70.46	67.80	2.68	0.063	0.18
Weight	173.50	187.00	207.73	173.50	1.11	0.36	0.38

Supplemental Table 5. A summary of the covariate differences between the correctly classified MDD cases and controls, and the incorrectly classified MDD cases and controls is presented. For the categorical variables, a count of subjects within each sub-category by diagnosis status is performed, and a chi-squared test is used to obtain a p-value for the contingency table for each categorical variable. For the continuous variables, the group means for the training and testing set subjects are recorded. A one-way ANOVA is then performed to obtain p-values for a difference between the groups. We then applied the *Benjamini–Hochberg procedure to calculate the false-discovery rate (FDR) based on the p-values.*

BH-FDR: *Benjamini–Hochberg false-discovery rate*

Supplemental Table 6. Hypergeometric test of enrichment of *piano* ranked *gene sets* in machine learning model genes from *brain mRNA* dataset.

Gene sets	p-value	BH-FDR	# overlap	Gene set size	Overlapping ML genes
GO_METALLOAMINOPEPTIDASE_ACTIVITY	0.001819	0.018576	1	22	ENPEP
GO_OXIDOREDUCTASE_ACTIVITY_ACTING_ON_A_HEME_GROUP_OF_DONORS	0.003377	0.018576	1	30	COX6A1
GO_AMINOPEPTIDASE_ACTIVITY	0.007813	0.028647	1	46	ENPEP
GO_ELECTRON_TRANSFER_ACTIVITY	0.043164	0.118702	1	114	COX6A1
GO_STRUCTURAL_MOLECULE_ACTIVITY	0.085157	0.187345	4	826	COL4A1, NEFH, NID1, CRYBA1
GO_TRANSMEMBRANE_SIGNALING_RECEPTOR_ACTIVITY	0.155809	0.28565	5	1277	CX3CR1, GALR3, GRIA2, PTPRC, GPR88
GO_G_PROTEIN_COUPLED_RECEPTOR_ACTIVITY	0.242579	0.381195	3	866	CX3CR1, GALR3, GPR88
GO_RIBONUCLEOTIDE_BINDING	0.305626	0.420236	6	1891	TDRD9, ABCC3, GIMAP6, GTPBP4, CHTF18, NLRP13
GO_DRUG_BINDING	0.392566	0.479803	5	1725	HMGCS1, TDRD9, ABCC3, CHTF18, NLRP13
GO_ADENYL_NUCLEOTIDE_BINDING	0.481346	0.52948	4	1541	TDRD9, CHTF18, NLRP13, ABCC3
GO_DNA_BINDING_TRANSCRIPTION_FACTOR_ACTIVITY	0.769717	0.769717	3	1701	CDX4, HOMEZ, ZNF511

Supplemental Table 6. The machine learning (ML) model genes from the best trained model, using the *brain mRNA* dataset, were extracted. A secondary enrichment analysis was performed to rank the filtered gene sets obtained from the *piano* consensus gene set analysis method. A hypergeometric test was performed to determine which of the filtered gene sets were significantly enriched for the ML genes. We then applied the *Benjamini–Hochberg procedure to calculate the false-discovery rate (FDR). The gene sets are then sorted from lowest to highest by the p-value.*

ML: Machine learning

BH-FDR: *Benjamini–Hochberg false-discovery rate*

Supplemental Table 7. List of all 1376 genes selected by XGBoost algorithm in the construction of the classifier for distinguishing MDD cases from controls for the *blood mRNA* dataset, ranked from most important to least important.

ZMAT1, GPR4, RAB5C, FUNDC2, TMEM71, TNNC1, FASN, RPS15A, SLC39A4, TOR1A, FAM129A, C1orf74, LCLAT 1, C6orf25, PPP1R3A, SAR1B, SERPINA3, DTX3, LRRC25, OR2AK2, FNDC7, IRF7, CLDN18, MLLT3, RNASE11, SPA NXN5, GOLT1B, ATF2, SCN3B, C17orf58, GPR64, LRRC14B, CAMK1D, ABCC13, COX4NB, PRUNE2, PPP1R2P9, CN GA1, OR5B21, PRSS23, SFRP2, LPXN, ACTN1, AP3B1, ISG20L2, PHOSPHO2, ENOX1, SCG5, TUBA3D, IL3, APPL2, C17orf95, NKAIN1, OTUB2, PSIP1, IGL@, DENND4C, NPM3, FAM185A, KCNJ15, BCL2L10, PEA15, MTMR6, PAX1, SERPINA10, ARHGEF17, LIN7C, CXCL11, MIXL1, KIAA1486, MAGI2, CHST10, C6orf1, FAM200A, AKAP12, RPL24, LOC100294376, AMICA1, TRPC4, HYAL4, TCEAL5, GP2, CMAS, TAS2R5, MIOS, LY75, MYEOV, TBC1D9, SPDYE2, PVALB, NEFM, TGFBR3, NUP85, SLC25A32, RALYL, PIK3R6, SKIV2L, LOC647979, PXDN, PKP4, PTPN5, NRAP, P ECAM1, PSMA6, ASB8, CCDC40, C14orf138, ENPP2, RNASEH1, RPL9, SENP8, KLRB1, SIRPG, TMED3, SLC25A38, VPS4B, CNTN4, KPNA3, GALM, CDK9, ELAVL4, BCL6B, ATG10, CDKN1C, C21orf88, TLL2, TERT, RND2, PIK3C2 G, PFKL, ZNF560, C10orf11, MAP1D, LEKR1, PLIN3, SLFNL1, COBL, NKAIN2, TCF23, NRXN1, TBC1D2, CCDC144 C, SOCS7, DOK5, ARHGEF37, NCKAP1, HTR1D, PLD3, ABCB7, CXCL6, C15orf54, CNO, CKAP2, DAD1, TAF10, RA B6B, QTRTD1, ISL1, ZNF487, ZNF559, ANGPTL2, LRRC17, LOC284009, SLK, CNBP, ZNF229, FBXO36, IFRD2, GNG 3, DCAF11, KRTAP5-4, UBE2O, TPRKB, FAM127B, OR51B6, PRICKLE1, ZHX3, ATXN7L3B, ATF6, MED22, PPIH, P RX, HCFC1R1, AGAP3, GKN1, DHX36, GPR128, CRTAP, SNORA84, PDCD1, C3orf36, GSTM3, POT1, C17orf42, ENT PD5, SLC25A12, KRTAP20-2, CENPJ, ARF4, GPR133, PSMB4, ZNF43, ProSAPiP1, WDR11, ARHGEF33, C8orf4, PCDH B17, LEPREL1, WIPI2, SLC38A8, LOC729815, ALDH3A1, ZNF479, MESDC1, SNORA58, MPPED1, AKR1CL1, CDK11 A, GCDH, FAM20A, ZNF718, TRAF7, C9orf167, NMT1, CXorf56, SCFD1, KRT9, KDELR3, SDPR, TMEM56, MTMR2, LCA5L, PKD1L3, RAB27B, DLG5, C11orf45, ST6GALNAC3, TRIM62, HOOK1, ARHGAP27, TAL1, ATP5G3, OSBPL1 0, CHRNB2, BAHD1, CISD2, RGMB, SEC11C, PHF6, BZW1L1, POGZ, ARFRP1, HIPK3, SRR, ZNF726, ADARB2, CO MTD1, TPTE, DHRS11, FLYWCH2, IL22, OR4A15, FBN1, NETO2, MTA3, ZIM3, C19orf55, DKK4, CXXC5, FAM131A, PRSS48, ERVFRDE1, GNAZ, LOC100289511, PDZD3, ARL13B, NTM, BAG5, PITPNA, POU4F2, THRA, SFPQ, PRKA B2, PPP2CB, C12orf69, CARNS1, CCRL2, SPAG11A, DBX1, ERP27, OR13F1, UQCRB, CNTN6, F9, PDE5A, SYN1, NT 5M, VPS11, RYR3, OR2Z1, ADAM17, OR52B6, TRIM23, FLJ41423, NUFIP1, PNN, ZNF572, Septin 4, BHLHE41, DBF4 B, PRSS42, CCDC157, CCDC144NL, TCF25, FLJ43950, CARHSP1, ZIC5, ELN, KIAA1875, HIST1H2BF, ODF4, ART4, FAM86C, CNTD2, CMTM8, FBXO45, PTPN4, RSL1D1, RTL1, SUN3, GLIPR1L2, RAD21L1, PMS2CL, NSUN5, HOXC1 3, MFHAS1, GRM5, KCNK16, ASH1L, COL7A1, FAH, MGC39372, FCGR3B, NIT2, PARVG, PIGB, ASB13, MFSD2B, RIMS4, THEM4, MESTIT1, C5orf54, VGLL3, GPR37L1, NHEDC2, NPBWR2, ANKZF1, SRGAP3, CCNE2, POC5, AKA P5, AIM1, FRK, CLIP4, TRAK1, C1orf26, TSKS, NIPAL2, MAS1L, GAD2, ZKSCAN1, ING3, ARV1, CTDSP2, PRRG2, HECTD2, TJP1, TTLL9, ENDOG, OR2J2, RC3H2, SNX21, LOC348120, PABPC5, CIDEB, ZBTB7B, COLEC11, ZNF358, RIN3, WBP4, MAP3K14, IL2RB, IL10RA, BBX, GATA4, C14orf177, TOP2A, ILF2, FCGR3A, MEIS1, OCEL1, CCDC67, MEX3A, ZNF534, CBX6, CCDC33, FIGLA, CLRN3, WDR78, NXNL1, OR2L13, P2RY2, MAP1B, FER, APOA1BP, BCL 6, TFB1M, NUTF2, GRIK4, KIF5A, PPP1R1C, CLDN19, TSC22D1, PCDHGB1, ARHGEF11, DNAJC8, MYO1H, NUDT 4, LHCGR, EIF2AK3, RPS24, SNORA78, DENND4A, VDR, RNF44, NAPG, SMAD1, ADIPOR2, SAMD12, GPR81, ZNH IT1, FAT2, PKD1, PTPRF, C3orf57, FKBP1A, ANKMY2, MAN1C1, TAC1, TDGF3, ODF3B, SNORD94, SGCG, DSC2, MYT1L, RNF151, H2AFZ, TUBB2A, TBX2, LUC7L2, RP1-177G6.2, IFI30, GHRL, PAOX, GART, TTBK1, KRTAP5-7, DPT, GLIS2, MMD2, MED7, INSIG2, CCDC89, KIAA0146, AMZ2, REG1P, AEN, CHST2, VEGFC, GPR15, C2CD3, CO MMD6, SNX22, FLAD1, OSGIN1, BIN2, COPB2, FLRT2, CHCHD4, CLDND1, PCTP, TES, CRB2, C12orf59, TRIM25, C CNB1, PEX6, HPD, ASCL4, C22orf29, EGFL8, CHPF2, C10orf79, SSR4, ZNF22, GNG5, AP2B1, KLK14, PSPH, NCRNA 00205, PLEKHA8, TMEM90B, KCNV2, DVL3, C6orf118, MED13, FAM65B, P2RY12, ZNF704, C10orf91, OR3A1, DNAJ A2, ZCCHC3, EIF3A, PRF1, SMN2, FBXW2, ACAD9, ADAMTS13, KPRP, PIP5KL1, SLCO1B1, EID3, OR51M1, FLJ375 43, ACTRT2, CTPS, DYNC1H1, GLT6D1, C8orf17, ZNF362, C16orf13, PIGP, PNPLA3, C18orf1, CARS, SCNN1B, KCN E4, FAM71C, WDR89, CAT, RBCK1, ZNF568, MTMR12, KRTAP13-1, RPS8, EHBP1, LOC100291462, CRLF2, GANC, LOC440313, OR5K3, EXOSC9, CYB561, HOXC4, KLF3, SPINK2, IRS4, GNG2, CHCHD3, LOC100289409, UST, PAPL, SEC23B, C12orf39, NEU4, ZNF423, DECR1, MND1, NANP, CCDC93, TBC1D8B, UPF2, LRRIQ1, YIPF1, SOX6, PMS2L 11, KCNK3, ENC1, CCDC130, HARS, UGT2B11, RHBDL1, TYK2, CNOT6, EXPH5, C20orf72, C13orf38, DLEU1, ERCC 1, DRG1, NANOS1, HOXB1, IGDCC3, STX5, AGTR1, SNCA, LOC402644, KIAA0317, PLCZ1, INSC, HOXD11, RBM2 3, ATOH8, ZNF736, SLC25A18, CLLU1OS, GPR141, LMTK2, MIA3, DYTN, CMBL, CCDC84, NUDT13, LOC339803, K LF7, NCRNA00085, KCNA1, CRISP2, CARTPT, SLC16A4, COQ9, LOC645961, PATE3, MBLAC1, PKN1, CNOT6L, FB XO4, C2CD2, C17orf63, SAP30, EMD, MFN2, KCNJ10, HLX, C11orf21, SOAT2, ZNF587, CALCB, NR2E3, KCP, MDN 1, SLC41A2, SLC4A10, FCER1G, ZNF10, HMGB4, PROCR, TIMM50, MAP7, RMND5A, SERPINB8, ELF5, TMEM39A, SLC6A15, RG9MTD3, FSTL4, ESR2, ZDHHC16, S100A3, GPR20, KRTAP25-1, TSPAN6, C6orf106, DGCR8, OR56A3, F BXL6, LOC642587, GDAP1, ADCY2, WNT2, ABT1, ZNF57, TSTD1, FNBP1L, WDR45L, LASS5, FAM166B, P2RY1, C RADD, ANKRD30A, LSM5, NS3BP, RBM10, RQCD1, PTF1A, SATL1, FKBP6, FERMT1, CNR1, PARP10, ARHGAP29,

MBD5, NTS, UBD, SLC25A41, ZDHHC21, CXorf40A, CUL4B, DACT2, NR1I3, GOLGA2B, XIRP2, TSPYL4, NCRNA00207, LIPH, GGH, IL20RB, PHF20, MT2A, C21orf125, TRPC7, HOXB13, CYP11A1, BMP10, ZNF345, BEST1, UPF3A, C9orf24, CNIH2, IAH1, THBD, RAP2C, SEC31B, L1CAM, FAM113A, TSPAN17, NEFL, SHISA2, TMEM97, FLNC, PCBD2, ERBB3, JAM3, DPF1, SEC16A, NME1, RPL10L, MYO1C, C1orf125, SLBP, CYP8B1, FLJ32063, KIAA0174, RPS19BP1, IGH@, STRADA, POLR2J, TAF1, SEMA6A, PRR7, ST3GAL5, SMEK1, FREM2, SNX27, TBP, MC4R, LOC100129500, SLC25A44, PCGF3, NUDT21, PRSS53, KLHL15, FAM169B, ORM2, DNAJB5, COL4A4, NDUFA9, ACADSB, PCDH1, ZNF713, CLDN5, GPRC5A, CPSF1, CLSTN2, VPS33B, MRPL1, MTMR11, REEP5, CLIC3, FAM54B, C17orf104, PIGA, ILDR2, ZFAT, C20orf173, NARG2, RBP1, OSTCL, NTRK2, CDH17, CMPK1, SPIRE1, CAPN9, RGS17, TBX21, ZNF391, CLEC3A, KRT17, LMAN2L, LOC100130428, ZNF770, ANKRD52, ZNF735, PLA2G7, SLC25A43, SDHB, PRKG1, CCR2, VAPB, MMD, TNNC2, ANTXR1, GPAM, MDK, ENO2, KCNK6, ELMOD2, KRTAP4-8, SPTLC2, PIK3R3, CLNK, HMGCL, LIPE, CDC123, OR4M1, ZNF441, LOC100134391, LOC220930, PSORS1C3, OR4C12, C10orf53, PTAR1, JUND, WDSUB1, ZNF628, SRP14, MCTP2, PHKA2, KDELC1, NKIRAS2, TOMM34, PARP16, NDUFA10, TYW1, LUC7L3, CEBPE, PIWIL2, KLHL21, LIN7A, DYNLRB1, CREB3L1, LOC100287301, ZBTB24, BEX1, NXT1, VAPA, IGF1, ZNF41, TRAF3IP1, DNAJC1, OR1L1, PIN4, OBFC1, PLEKHA9, PDE3A, CYP27A1, KCND1, RNF34, CD247, LIMS2, LOC729678, JAG2, DCLRE1B, RBM41, FGD5, C11orf74, WIF1, SLC16A1, IFNA7, SURF1, KDM4B, POLR2H, PARD3B, ITFG2, MT1H, PATE2, SHE, SDHAF1, ZNF821, LOC401093, FAM45B, C6orf141, CCDC155, ZAK, PKM2, EPHA8, GBP1, ANAPC2, GNA12, SECTM1, C7orf13, COCH, DEFB1, LNX1, C1orf87, NELL1, ATP13A1, LRRC20, OAF, IGSF22, SLC2A3, CHST15, PARP6, FKBP10, APOL3, CDH18, SLC5A12, OR10A5, MIER3, GABRR3, ATXN3L, SP100, BLVRA, FGF20, CCDC55, FAT3, TMBIM1, CALHM3, C7orf23, FAM70A, HOXA6, BTNL9, GRIN2D, PMS2L5, IDH3B, ADAMTS20, CDC42SE2, NUAK2, PEX2, HAS2, SLC41A1, OSBPL7, CALU, XAGE3, OR51B4, ZNF579, SPSB1, IL2, SBSN, NHLRC4, CLEC1A, MAGEE1, C12orf75, IQUB, LDLR, OR6K3, PLAC1L, HTN3, LOC653888, KRTAP4-5, TBCCD1, UBE4A, PLEKHM1, C1orf213, SLC29A4, FAM195A, F11, UBA52, MST1P9, PANK2, GUCY1A3, LHX9, PPBPL1, MAB21L2, KRT6B, EXOSC5, March 9, MGC131512, ZPLD1, NHP2L1, RBM3, PCNA, COPS8, C4orf23, SNORA37, CAPN2, GEMIN5, NUDT1, C1orf130, NDUFB1, TTLL1, SNRPA1, CNPY4, DPY19L2, FZD6, MTSS1, SLC22A9, TMEM74, ABCC10, STAT1, ITGA8, LOC727726, PTGES3, C20orf43, ATP5G1, CHST12, SCTR, ZNF185, ZNF569, KRTAP21-2, ANKRD36, MAML2, SAMD11, PSMD4, TMEM101, SPINK14, SMCR7L, KLK3, GABRA6, CALCOCO2, NCRNA00219, OR5AU1, NDNL2, PPP1R14B, KBTBD13, SLC39A5, IMPDH1, ZNF681, SERPINB1, TCEA3, ITGAE, HIST1H2BE, CCDC74B, TTPA, RBM8A, LCE2D, TF, FAM19A4, TMEFF2, ZBTB48, FAM184A, ZCCHC8, C6orf182, KAT2A, GABARAPL3, PIP5K1B, KIAA2026, MBLAC2, ZNF193, FLJ44082, LAD1, DNAJC6, ZSCAN4, TEX19, GTSE1, COQ6, SLC7A4, TNFSF4, CPXCR1, HOOK2, KANK2, FAM3A, IMPG1, NBN, GBP3, FUCA2, HCP5, SAMD13, C5, FAM75B, INSR, PPFIBP2, C3orf71, PPWD1, OTC, KIAA0100, HRK, HSPA9, DDX11, UQCRHL, C2orf82, PPAN-P2RY11, LMO4, SCD, LOC100287290, GPR25, UBE2G2, CCL16, NUDT8, COMMD9, TMEM231, TRAJ17, SESN1, FAM196A, SHROOM3, RP1, HPSE, MPZL1, KLRD1, BPHL, LOC100129503, LILRA4, FBXL8, SIN3B, SDR42E1, TRPV1, KCNQ3, NLRP12, VIT, FAM64A, OR11H6, ABHD13, HES7, ELAVL1, UGT3A1, LOC283867, FUCA1, CLYBL, HSD17B7P2, ROR1, IRX4, C9orf64, C6orf89, SLC25A31, SERPINB4, CUZD1, IFT81, ISCA2, OR10A2, TRBV9, ACTRT1, DDO, RBKS, KIAA0141, DOM3Z, SASH3, DPP6, POFUT1, PSMC4, NOP14, DCXR, F11R, USE1, CD200R1L, CMYA5, DDAH1, DENND5B, PRDM8, ZSCAN5A, ANKRD16, TFAP2C, ZNF433, DCAF5, PRSS50, HMX2, MUC13, ALDH4A1, ELOVL4, CHAD, ZNF614, CYP4A22, NOL3, SAP130, REEP2, CYC1, C19orf50, C3orf22, PSMD6, NTN1, Septin 5, GRASP, LPAR6, MLF1, NCAPG, THPO, BLOC1S2, LOC100130539, ACBD7, MXRA5, CD163L1, C1orf103, CYP19A1, OR4D2, RPS4Y1, SRPR, GEMIN4, LOC220729, C8orf49, FAM9C, CRABP2, GRM4, DKFZP586I1420, PSMB2, TXNDC6, WIT1, SNTN, CDKN2B, DCTN2, PROZ, FOLR2, APBB1, C11orf59, TCTEX1D1, DSCR3, IGFBP6, BAGE, RBP4, CACNG3, LGALS2, CORO2B, ALDH18A1, MLH3, C15orf55, CABC1, RNF121, TMEM143, PPP1CA, C1orf173, GALNTL2, SERP2, ARMC10, DOK2, CA7, ZNF600, FANCG, CACNB3, FDFT1, OR10H3, FAM171A1, C12orf36, LIPK, PPM1L, VPS26A, PRDX4, GMCL1L, ADAM29, CASS4, CLIC6, GCNT4, KCNK10, DEFB134, MTHFR, FAM84B, EXOC3, CHCHD2, PYGM, C4orf49, RSPH3, RNF17, GOLGA6L7P, GLRA2, CRYGA, HIST1H1T, CHKA, SCARNA14, LUZP1, PCDHB1, BDKRB1, LZTFL1, RDH5, C17orf105, TMC1, MBOAT4, KCNC3, DPY19L3, SI, CDC42EP3, LY6G5C, MAN1A1, SLITRK1, SLC28A1, TLR6, LGALS9B, IRF2BP1, NR5A2, RPS28, GCSH, KCNC1, HIST1H2BA, NCAM1, GPR176, OSBPL3, CYP1B1, HSBP1L1, HTRA3, LCE2C, C20orf69, IFLTD1, SNRPD2, BLK, KDELC2, FABP9, C10orf108, NEK9, TMEM158, SGSM3, ZBTB10, NT5C3L, C9orf47, C22orf42, LYRM5, POLR1C, NCOA4, CNGA4, FKBP14, CCDC115, SLAMF9, ELK4, DNAI2, TGFB1I1, C11orf61, C9, PCDHGB7, TMEM192, PPFIBP1, BRP44, TBC1D30, C1orf95, ERAS, KILLIN, PRDX2, ZNF483, AMELY, VSTM2A, SERINC5, CBFB, C18orf23, TTC1, PCDP1, SC4MOL, MRPS9, STAB2, GCG, ANKRD50, MPI, TREX1, PPP1R2P1, TUBGCP4, LRRIQ4, FRAT2, WDR82, SNRPA, ESD, MME, MIER2, RSPO2, GALNT6, CXCR6, LOC100128108, MAPK14, KCNA4, POLG, DPP8, EIF2AK4, SFRS14, LIMCH1, KRTAP12-1, TSGA10, ZNF28, ILVBL, SAPS2, KTI12, PCDHGA8, NOS1, LEO1

Supplemental Table 8. Summary of the covariates between MDD cases and controls from the *blood mRNA* dataset.

	MDD	Control	Test statistic	P-value
Sex			9.18	0.0025
Sex (Female)	1080	221		
Sex (Male)	501	148		
Smoker status			31.33	2.17E-08
Smoker status (No)	890	267		
Smoker status (Yes)	691	102		
Alcohol status			10.95	0.0042
Alcohol status (No)	242	32		
Alcohol status (Unknown)	5	1		
Alcohol status (Yes)	1334	336		
Menostats			16.89	0.0047
Menostats ((Recent) pregnancy)	74	8		
Menostats (Male/NA)	501	148		
Menostats (Menopause, natural)	5	2		
Menostats (Menopause, operation or disease)	285	72		
Menostats (Not in menopause)	702	134		
Menostats (Unknown)	14	5		
Education			27.42	1.11E-06
Education (Basic)	118	12		
Education (High)	498	165		
Education (Intermediate)	965	192		
Continuous variables				
Age	41.92	42.81	-1.20	0.23
BMI	25.83	25.36	1.57	0.12

Supplemental Table 8. A summary of the covariate differences between MDD cases and controls is presented. For the categorical variables, a count of subjects within each sub-category by diagnosis status is performed, and a chi-squared test is used to obtain a p-value for the contingency table for each categorical variable. For the continuous variables, the group means for the MDD cases and controls are recorded. An independent t-test is then performed to obtain p-values for the difference between the groups.

Supplemental Table 9. Summary of the covariates between the correctly and incorrectly classified subjects from the *blood mRNA* testing set.

	Correct-MDD	Incorrect-MDD	Correct-control	Incorrect-control	Test statistic	P-value	BH-FDR
Sex					4.62	0.20	0.36
Sex (Female)	109	110	26	19			
Sex (Male)	47	48	22	9			
Smoker status					14.40	0.0024	0.017
Smoker status (No)	77	101	36	20			
Smoker status (Yes)	79	57	12	8			
Alcohol status					8.50	0.20	0.36
Alcohol status (No)	19	27	4	5			
Alcohol status (Unknown)	0	2	0	1			
Alcohol status (Yes)	137	129	44	22			
Menostats					10.42	0.58	0.77
Menostats ((Recent) pregnancy)	6	6	1	0			
Menostats (Male/NA)	47	48	22	9			
Menostats (Menopause, operation or disease)	24	33	7	5			
Menostats (Not in menopause)	77	71	18	14			
Menostats (Unknown)	2	0	0	0			
Education					11.57	0.072	0.25
Education (Basic)	12	8	3	0			
Education (High)	48	59	26	10			
Education (Intermediate)	96	91	19	18			
Continuous variables							
Age	41.17	41.28	43.06	41.00	0.29	0.83	0.83
BMI	25.55	25.66	24.54	25.17	0.54	0.66	0.77

Supplemental Table 9. A summary of the covariate differences between the correctly classified MDD cases and controls, and the incorrectly classified MDD cases and controls is presented. For the categorical variables, a count of subjects within each sub-category by diagnosis status is performed, and a chi-squared test is used to obtain a p-value for the contingency table for each categorical variable. For the continuous variables, the group means for the training and testing set subjects are recorded. A one-way ANOVA is then performed to obtain p-values for a difference between the groups. We then applied the *Benjamini–Hochberg procedure to calculate the false-discovery rate (FDR) based on the p-values.*

BH-FDR: *Benjamini–Hochberg false-discovery rate*

Bridging statement to Chapter 4

The studies presented in Chapters 2 and 3 demonstrated the potential of ML and transcriptomic data in advancing our understanding of SCZ and MDD, supporting the usefulness of ML analysis of gene expression microarray data in furthering our understanding of the pathophysiology of complex diseases. Our study in Chapter 3 applied supervised ML analysis to both brain and blood gene expression microarray data from MDD cases and controls, providing valuable insights into the disorder's underlying pathophysiology. Furthermore, the methodological improvements implemented in this study, such as the analysis of an external brain gene expression dataset for additional model validation, addressed the limitations associated with model evaluation using a single data source. The findings from our improved methodology further support the confidence in the relevance of the genes identified for a role in MDD.

The ML methodologies developed in Chapters 2 and 3, as part of our first objective, are important for identifying and understanding the molecular basis of complex diseases. As we illustrated in our publications, this work could contribute to the identification of treatment targets and the development of novel treatments.

In Chapter 4, we advance our second objective towards precision medicine, investigating the application of ML for optimizing the treatment of complex diseases, with an analysis of microRNA data for disease diagnosis, severity prediction, and treatment response prediction in the context of MDD. MicroRNAs play a crucial role in the regulation of gene expression [108], and have been implicated in the pathophysiology of various diseases [109], including MDD [110]. Studying microRNA expression profiles could provide valuable insights into the molecular mechanisms underlying MDD and potentially lead to the identification of more stable and reliable biomarkers for the disorder. Furthermore, the transition from mRNA to microRNA

data offers a novel and more regulatory-level perspective on the molecular mechanisms underlying MDD. In addition, the incorporation of clinical measures such as the Montgomery-Asberg Depression Rating Scale (MADRS) scores and treatment response data enable a more comprehensive evaluation of the utility of microRNA-based ML models in the clinical management of MDD patients. Our approach aims to further advance our understanding of MDD but also to improve patient outcomes by leveraging omics and ML for informing treatment decisions and better outcomes for patients through precision medicine.

Chapter 3 erratum:

- On page 100, *"The best cutoff is defined as the probability threshold dividing the cases and controls classes which maximized the number of true positive classifications and minimized the number of false-positive classifications (i.e., maximizing the area under the ROC curve)"* is inaccurate and should be corrected. We are not maximizing the area under the ROC curve, but rather, we are finding the probability threshold for which the rectangular area under the TPR and FPR point along the curve is maximized. Another equivalently correct way for defining the best cutoff is as the probability threshold dividing the predicted samples where the average of the true-positive rate (TPR) and the inverse of the false-positive rate (FPR) (i.e., 1 - FPR) is maximized.

Chapter 4. Machine learning analysis of blood microRNA data in major depression: a case-control study for biomarker discovery

AUTHORS

Bill Qi[1], MSc (cand.); Laura M. Fiori[2], PhD; Gustavo Turecki[2], MD PhD; Yannis J. Trakadis[1,3], MD MSc FRCPC FCCMG

AUTHORS INSTITUTIONAL AFFILIATIONS

[1] Department of Human Genetics, McGill University, Montreal, QC, Canada

[2] Department of Psychiatry, McGill Group for Suicide Studies, Douglas Mental Health University Institute, McGill University, Montreal, Quebec, Canada

[3] Department of Medical Genetics, McGill University Health Center, Montreal, QC, Canada

Correspondence to

Yannis J. Trakadis, MD MSc FRCPC FCCMG

Medical Geneticist & Metabolics Specialist

Assistant Professor, Human Genetics

McGill University Health Centre

Room A04.3140, Montreal Children's Hospital

1001 Boul. Décarie, Montreal, Quebec, Canada, H4A 3J1

Tel: (514) 412-4427, Fax: (514) 412-4296

Email: yannis.trakadis@mcgill.ca

Abstract

Background: There is a lack of reliable biomarkers for Major depressive disorder (MDD) in clinical practice. However, several studies have shown an association between alterations in microRNA levels and MDD, albeit none of them has taken advantage of machine learning (ML)

Method: Supervised and unsupervised ML were applied to blood microRNA expression profiles from a MDD case-control dataset (n=168) to distinguish between 1) case vs. control status, 2) MDD severity levels defined based on the Montgomery-Asberg Depression Rating Scale (MADRS) and 3) antidepressant responders vs. non-responders.

Results: MDD cases were distinguishable from healthy controls with an area-under-the receiver-operating characteristic curve (AUC) of 0.97 on testing data. High vs. low severity cases were distinguishable with an AUC of 0.63. Unsupervised clustering of patients, before supervised ML analysis of each cluster for MDD severity, improved the performance of the classifiers (AUC of 0.70 for cluster 1 and 0.76 for cluster 2). Antidepressant responders could not be successfully separated from non-responders, even after patient stratification by unsupervised clustering. However, permutation testing of the top microRNA, identified by the ML model trained to distinguish responders vs. non-responders in each of the two clusters, showed an association with antidepressant response. Each of these microRNA markers was only significant when comparing responders vs. non-responders of the corresponding cluster, but not using the heterogeneous un-clustered patient set.

Conclusions: Supervised and unsupervised ML analysis of microRNA may lead to robust biomarkers for monitoring clinical evolution and for more timely assessment of treatment in MDD patients.

Introduction

In the United States, the lifetime prevalence for MDD is 20.6% among individuals aged 18 years or older. Almost half (49%) of the cases have severe and 39.7% moderate depression (Hasin et al., 2018). Without early treatment, there can be permanent consequences on the patient's brain function which increase their risk of experiencing additional depressive episodes (Moylan et al., 2013). Overall, the economic burden of MDD is more than 170 billion per year, and appears to be increasing over time (Greenberg et al. (2015). However, there is still a lack of reliable biomarkers that can guide patient monitoring and timely assessment of treatment efficacy.

Increasing evidence suggests that molecular signaling for depression is linked with microRNA expression and that the dysregulation of microRNA signaling can initiate or exacerbate depressive pathophysiology (Hansen and Obrietan, 2013). MicroRNAs are small noncoding RNA molecules that play a role in the regulation of gene expression and neuronal physiology. Smalheiser et al. (2012) found that the expression of several microRNAs was significantly down-regulated in the prefrontal cortex of depressed suicide individuals compared to matched psychiatric control subjects. Bocchio-Chiavetto et al. (2013) measured the expression of microRNA in 10 depressed individuals before and after treatment with antidepressants. After the treatment with antidepressants, two microRNAs were significantly down-regulated and 28 were up-regulated. In a recent randomized placebo-controlled trial, we identified several microRNA markers of duloxetine treatment response which were replicated in two independent clinical trials, an animal model, and post-mortem brain samples (Lopez et al., 2017). The findings suggest that there is a strong possibility that microRNAs are involved in the pathophysiology of depression and affect the mechanism of action of antidepressants.

Machine learning (ML) algorithms have been created for analyzing complex multivariate data with a focus on empirical predictive power and generalizability. ML has demonstrated success in clinical psychiatry in terms of diagnosis, prognosis, treatment decisions, and biomarker detection (Dwyer et al., 2018). A review of the literature on ML and MDD shows a shortage of studies that apply ML methods to analyze microRNA data (Gao et al., 2018). However, a recent paper has demonstrated the effective use of ML in identifying a serum microRNA signature for Alzheimer's disease that could predict disease status with 85.7% accuracy (Zhao et al., 2019).

Given the important role of microRNAs in MDD, and the effectiveness of ML in taking advantage of complex data, we aimed to explore whether ML analysis of blood microRNA profiles can serve as a new approach for biomarker discovery in MDD.

Methods

Participant recruitment

The study protocol was approved by the Research Ethics Board of the Douglas Mental Health University Institute (DMHUI). Informed written consent was obtained from all participants. All participants were recruited from an outpatient clinic at the Douglas Mental Health University Institute in Montréal, Canada, and assessed by an experienced psychiatrist using the SCID-I (First et al., 2012) following DSM-IV criteria. Patients were all suffering from a current major depressive episode as part of a major depressive disorder (MDD). Exclusion criteria included comorbidity with other major psychiatric disorders, bipolar disorder, alcohol or substance abuse over the last 6 months, or a severe medical condition. None of the participants

were medicated at baseline. None had received fluoxetine or lithium over the last month or any psychotropic medication over the last week. Depression severity was determined using the Montgomery-Asberg Depression Rating Scale (MADRS). MADRS measures are based on 10 different symptoms including 1) apparent sadness, 2) reported sadness, 3) inner tension, 4) reduced sleep, 5) reduced appetite, 6) concentration difficulties, 7) lassitude, 8) inability to feel, 9) pessimistic thoughts, and 10) suicidal thoughts. The MADRS scores were collected at baseline and again after 8 weeks of antidepressant treatment.

Sample Processing

Peripheral blood samples were collected at baseline and after 8 weeks, and tubes were frozen using a sequential freezing process. Whole blood for RNA was collected in EDTA tubes and filtered using LeukoLOCK filters (Life Technologies). Total RNA was extracted using a modified version of the LeukoLOCK Total RNA Isolation System protocol and included DNase treatment to remove genomic DNA. The RNA quality was assessed using the Agilent 2200 Tapestation, and only samples with RNA Integrity Number (RIN) $\geq$ 6.0 were used.

Small RNA-seq

All libraries were prepared using the Illumina TruSeq small RNA Library preparation protocol following the manufacturer's instructions. Samples were sequenced at the McGill University and Genome Quebec Innovation Centre (Montreal, Canada) using the Illumina HiSeq2000 with 50nt single-end reads. All sequencing data were processed using CASAVA 1.8+ (Illumina) and extracted from FASTQ files. The Fastx_toolkit was used to trim the Illumina adapter sequences. Additional filtering based on defined cutoffs was applied, including 1) Phred quality (Q) mean scores higher than 30, 2) reads between 15-40nt in length, 3) adapter detection based on perfect-10nt match, and 4) removal of reads without a detected adapter. Additionally,

we used Bowtie (Song et al., 2014) to align reads to the human genome (GRCh37) and ncPRO-seq (Chen et al., 2012) in combination with miRBase (V20) (Kozomara and Griffiths-Jones, 2013) to match them to known microRNA sequences. Furthermore, all sequencing data was normalized with the Bioconductor – DESeq2 package (Love, Huber, & Anders, 2014), using a detection threshold of 10 counts per miRNA.

The number of microRNA samples included for subsequent analyses includes the baseline (T0) and week 8 (T8) of 140 MDD cases and 28 healthy controls. The total number of microRNA features is 285.

ML analysis

Many powerful ML algorithms render themselves uninterpretable, making it difficult to understand their decision-making process. For our machine learning analysis of the data, we decided to use a state-of-the-art yet interpretable regularized gradient boosted machines (GBM) approach (XGBoost implementation, (Chen and Guestrin)), which we also demonstrated as an effective algorithm in our previous study of schizophrenia (Trakadis et al., 2019).

Datasets are split into 70% and 30% for training and testing. A model selection procedure based on 5-fold cross-validation with 2500 iterations of parameter search is used to obtain the best training parameters using only the training dataset (n=122). After obtaining the best training parameters, we retrained the model without 5-fold cross-validation, i.e. using at once the entire training set, and evaluated the model on the testing set. The model performance metric we used is the area under the receiver-operating characteristics curve (AUC).

Classification analyses

With regards to discriminating cases from healthy controls, we trained the ML model using only the T0 microRNA data, to ensure that the medication effect would not act as a confounder in the analysis.

For the severity class classification, we used the MADRS cutoff scores suggested by Snaith et al. (1986). Individuals' MADRS scores were classified as "normal-mild" (MADRS scores from 0-19) or "moderate-severe" (MADRS scores 20 and above). Using the class-labeled dataset, we identified the best classification model for classifying samples into these two MDD grades. However, for this analysis, we used the T8 microRNA data and MADRS scores, because at T0 all but two cases had MADRS scores ranging from 0-19 (thus, using T0, almost all samples would be labeled as "moderate-severe").

We then repeated severity class classification after unsupervised clustering of the T8 microRNA data, which was done to factor in the heterogeneity of MDD. More specifically, 500 iterations of a consensus k-means clustering method (Monti et al., 2003) were applied to the entire case-control dataset (n=174). The model selection and evaluation procedure were then performed separately for each cluster, under the assumption that the patients in each cluster are less heterogeneous at the pathophysiology and microRNA level. If our assumption is correct, training the ML algorithm to identify signatures specific for the "normal-mild" versus "moderate-severe" class would be more efficient after unsupervised clustering, and thus, the classification based on supervised ML analysis of the microRNA data would improve with this approach.

Lastly, we explored the relationship between microRNA and antidepressant response among MDD patients, using the difference between the T8 and T0 microRNA values (n= 138, because two MDD cases were missing MADRS scores at T0). With regards to severity levels, the scores used were based on previous definitions: “normal” (0-6), “mild” (7-19), “moderate” (20-34), and “severe” (> 34) (Snaith et al., 1986). Antidepressant response in our study was defined as a decrease of two severity levels when comparing the patient's T8 and T0 MADRS scores. For example, a patient with a change from “severe” to “mild” or a change from “moderate” to “normal” would be labeled as a responder (RES). We obtained a split of 46 RES and 92 non-responders (NRES). We then repeated the same ML procedure described above to obtain a classifier for RES vs. NRES.

Since each patient was taking a mixture of multiple antidepressants, to address this heterogeneity and improve the performance of the classifier, we performed unsupervised clustering on the “T8-T0” dataset using the consensus k-means algorithm described above, in the MDD severity classification section. Samples were split into two clusters. We then performed ML classification analysis for antidepressant response separately in each individual cluster. To explore if the top microRNA for each cluster (i.e., the microRNA with the maximum importance in the ML classification model) was associated with antidepressant response, a permutation test was performed. Specifically, 500000 iterations were performed to derive the empirical p-value of a difference in mean between responders and non-responders. The significance threshold was set at 0.05. In the case of multiple top microRNAs (multiple top microRNAs having equal maximum importance), the p-values were adjusted using the Bonferroni correction method. To explore if the top microRNA(s) identified in each cluster were specific to that cluster, we performed permutation testing for microRNAs extracted from the first cluster using samples from the

second cluster, and vice versa. We also performed permutation tests for the top microRNAs identified from the clusters using all MDD samples (un-clustered). Finally, we extracted the top microRNA from the antidepressant classification model trained on all samples and performed permutation testing to explore if ML analysis of the data before stratification was helpful in the identification of a marker for treatment response.

Clinical history analysis

To explore how clinical history factors into antidepressant response, we examined whether patients with a prior history of treatment with antidepressants responded differently compared to antidepressant-naïve patients. We also examined whether patients who present with their first major depressive episodes (MDE) (i.e., no prior episodes besides the current one) responded differently compared to patients with recurring MDEs (collected from SCID-IA (DSM-IV), question A29). Antidepressant response here is defined as a ratio of the T8 to T0 MADRS score (T8/T0) in order to capture more precise differences in antidepressant response between groups using the permutation tests. Permutation tests were performed for 500000 iterations to derive the p-value for a significant difference in antidepressant response between the groups compared. Multiple testing was adjusted using the Bonferroni correction method.

Bioinformatic analysis

We extracted the microRNA features used by the best *case-control* classification model and performed pathway analysis using the DIANA-miRPath v3.0 pathway analysis web-server (Vlachos et al., 2015) to obtain KEGG pathway terms significantly related to the set of microRNA features. Pathways with a false-discovery rate (FDR) less than 0.05 were selected.

Software

The ML model was implemented using the Python (v.3.7.1) programming language (https://www.python.org/) with the 'xgboost' (v.0.81) library (https://xgboost.readthedocs.io/). The consensus clustering procedure was implemented using the 'scikit-learn' (v.0.21.2) (https://scikit-learn.org/) and 'scipy' (v.1.3.0) (https://www.scipy.org/) Python libraries.

Results

The demographics of patients and controls are summarized in Supplemental Table 1. 65% of MDD cases, and 46% of controls, were female. Moreover, 80% of MDD cases, and 82% of controls, were Caucasians. The mean MADRS score at T0 was 33 (SD: 6.2) for cases, and 0.6 (SD: 1.1) for controls. The mean MADRS score at T8 was 17.4 (SD: 10.9) for cases, and 1.1 (SD: 1.7) for controls. 56% (n=79) of patients reported presenting with their first MDE. 14% (n=19) MDD patients were antidepressant naïve prior to current treatment.

As summarized in Table 1, for classification of *cases* and *controls*, the best trained model achieved an average cross-validation AUC of 0.93 (std. 0.06), and testing set AUC of 0.97. The best trained model trained to distinguish cases from controls utilized 33 out of 285 total microRNAs measured (Table 2). Pathway analysis for the 33 microRNAs found the following significantly enriched pathways with FDR < 0.05: 1) Prion diseases, 2) TGF-beta signaling pathway, 3) Morphine addiction, 4) Signaling pathways regulating pluripotency of stem cells, 5) Mucin type O-Glycan biosynthesis, and 6) Proteoglycans in cancer.

Classification of individuals as *normal-mild* vs. *moderate-severe* MADRS grades using their microRNA data based on best trained model showed an average cross-validation AUC of

0.76 (std. 0.11). After retraining the best model on the full dataset and evaluating on the testing set, we obtained an AUC of 0.63.

For the clustering approach, we obtained two clusters (cluster 1: 89 subjects, cluster 2: 79 subjects) of similar sample size, which did not show differences in terms of MDD severity. The best model for cluster 1 samples achieved an average cross-validation AUC of 0.75 (std. 0.18), while the best model for cluster 2 samples achieved an average cross-validation AUC of 0.72 (std. 0.15). When evaluated on the testing sets, the cluster 1 model achieves an AUC of 0.76, while the cluster 2 model achieves an AUC of 0.70. Table 1 summarizes the results for each of the analyses.

For antidepressant response classification, we obtained an average cross-validation AUC of 0.62 (SD: 0.13), and an AUC of 0.57 on the testing set. After clustering, we again obtained two balanced clusters (cluster 1: 69 subjects, cluster 2: 69 subjects). We did not notice a separation of responders from non-responders based on clustering. The best model for cluster 1 samples achieved an average cross-validation AUC of 0.65 (SD: 0.085), while for cluster 2 the average cross-validation AUC was 0.67 (SD: 0.16). On testing set evaluation, the cluster 1 model achieves an AUC of 0.54, while the cluster 2 model achieves an AUC of 0.49. For cluster 1, after supervised ML for classification of treatment response, the top and only microRNA utilized by the ML model was *hsa-miR-5701*. Following permutation testing, this microRNA was found to be significantly different between responders and non-responders in cluster 1 (p=0.021), but not in cluster 2, nor the original (un-clustered) dataset. For cluster 2, there were 4 microRNAs, all with equal importance, including: *hsa-let-7b-3p*, *hsa-let-7g-5p*, *hsa-miR-130b-3p*, and *hsa-miR-30d-3p*. Following permutation testing, the only nominally significant microRNAs were *hsa-let-7b-3p* (p=0.021) and *hsa-miR-130b-3p* (p=0.045), albeit neither were significant after Bonferroni

correction (p=0.082 and p=0.18, respectively). Neither of these four markers was significantly different between responders and non-responders in cluster 1 or the original (un-clustered) dataset. Finally, when extracting the top microRNA from the antidepressant classification model trained on all samples, the top microRNA was not found to be associated with treatment response following permutation testing (p=0.12).

Of note, we observed that antidepressant-naïve patients responded significantly better than those with who have taken antidepressants in the past (p=0.00058, with Bonferroni correction), but did not observe a significant difference in response between patients who present with their first vs. recurring MDEs (p= 0.59, with Bonferroni correction).

Discussion

In this paper, we demonstrate how ML analysis of blood microRNA data could lead to biomarkers with potential clinical utility. Our assumption was that if this is true, ML analysis of microRNA data should not only lead to the successful classification of cases from controls, but also to the efficient separation of individuals with mild vs. severe depression.

First, we showed that microRNA data could be used to discriminate baseline medication-free MDD cases from controls (AUC: 0.97 using the test dataset). Of note, this result is not expected to be confounded by medication effects since we used only the T0 pre-treatment trial microRNA data. To show that the microRNA signals are relevant to MDD, we conducted a pathway analysis using the microRNAs identified by the ML model (FDR < 0.05). We identified six pathways and highlighted the evidence in the literature for a link with MDD.

For example, there is evidence that *endogenous prion protein (PrP(C))* is associated with MDD. PrP(C) were reduced in the white matter (Weis et al., 2008), and Brodmann's (BA) areas 6 and 10 (Dean et al., 2019) in patients with MDD. PrP(C) has also been shown to modulate depressive-like behavior in mice (Gadotti et al., 2012).

TGF-beta family of cytokines may also play a role in MDD. TGF-beta has been observed to be significantly elevated in the peripheral blood of MDD patients (Davami et al., 2016). Furthermore, a study found a significant decrease in TGF-beta1 in MDD patients after 6 weeks of treatment with an antidepressant (Kim et al., 2007).

The dopamine and reward systems are major parts of the *morphine addiction pathway* (Kim et al., 2016), and there is a link between dopamine neurons and depression (Knowland and Lim, 2018).

There is also evidence linking *stem cell and cell renewal capacity* to MDD. In mice with interferon-α induced depression, neural stem cell proliferation was found to be suppressed (Zheng et al., 2014). Furthermore, shorter telomere length (Verhoeven et al., 2014) is also associated with a higher severity of depression.

Although no direct link exists between *mucin type O-glycans* and MDD, a study showed that the p75 neurotrophin receptor (p75NTR), a heavily glycosylated protein, had a polymorphism, Ser205Leu, for a predicted O-glycosylation site which had a protective effect for MDD (Fujii et al., 2011).

Next, we showed that microRNAs could be leveraged to distinguish subjects with normal-mild from moderate-severe MDD (AUC: 0.63). We also demonstrated that the use of unsupervised clustering, aimed at reducing MDD heterogeneity, can improve model performance

in our MDD grade classification task (AUC of 0.76 for cluster 2 and 0.70 for cluster 1). This supported our assumption that after unsupervised clustering the individuals in each cluster were less heterogeneous. This lead to a more efficient training of the ML algorithm to identify signatures specific for the "normal-mild" versus "moderate-severe" class. The sample size of our dataset is relatively small. However, given our results, we expect that performance estimates would improve and become more precise with ML models trained on larger samples.

We found that the differences between the T8 and T0 microRNAs were not strongly predictive of response status (AUC: 0.57 on the testing set). This came as no surprise, given that the patients were undergoing treatment with different antidepressants, thus leading to heterogeneity negatively impacting the performance of the ML model. Patient stratification partially addressed this, as we saw a slight boost to the 5-fold cross-validation performance for each cluster, compared to the un-clustered analysis. However, we did not see any improvement in classifying response status on the testing set. We believe that the poor performance of the ML models on the testing set is likely due to the small sample size of each cluster, but that there may still be intelligence derived from the approach. This is supported by the significant association of the top microRNA within each cluster with antidepressant response, which was specific (i.e., was not observed when analyzing the data of the other cluster or the un-clustered data). Furthermore, no marker was found for antidepressant response when extracting the top microRNA from the antidepressant classification model trained on all samples (i.e., un-clustered dataset). Putting everything together, we take this as evidence that a clustering approach, combined with supervised ML, could be useful to identify biomarkers in subgroups of patients which would otherwise be missed when analyzing heterogeneous populations.

Of note, our approach, using regularized ML with empirical cross-validation and testing as a method to prioritize features, rather than multiple univariate testing, facilitates finding relevant biomarkers with minor effects or complex interactions that would otherwise be filtered out by multiple testing correction. This is very important given the complex relationships of different factors contributing to MDD, such as duration of the depressive episode, duration of illness, and recurrence. For example, we found that patients who have no history of taking antidepressants responded significantly more to treatment compared to those with past history. At first sight, this should come as no surprise, since the usage of more antidepressant in the past indicates that the patient did not respond to the previous antidepressants, and thus that they are harder to treat. However, we did not observe a difference between patients experiencing their first MDE vs. patients with recurring MDEs, which contradicts this line of thinking and underlines the complexity of the different factors and their interaction in MDD.

Changes at the microRNA level are downstream to the different contributing clinical factors, thus explaining why we were able to distinguish cases from controls (AUC: 0.97) successfully. However, in order to better understand the contribution of each factor in MDD, further studies with larger sample size and more optimal patient stratification, with the inclusion of genetic, functional genetic, and detailed clinical data, would be recommended. Moreover, future studies should not be focused on examining changes between binary time points for antidepressant treatment, but rather serial (i.e., at multiple time points) MADRS evaluation and collection of microRNA data. With this design, we could explore if early changes at the microRNA level after treatment initiation could predict treatment response at a later point, which would have major clinical implications in treatment optimization.

Conclusion

Our manuscript provides preliminary evidence that ML analysis of blood microRNA profiles may constitute a reliable approach for biomarker discovery for MDD (affected vs. unaffected) clinical status, but also for clinical evolution (severity and treatment response), thus facilitating a more personalized approach in treating patients with MDD.

Authors' contributions

Bill Qi performed the bioinformatic and machine learning analyses and drafted the manuscript under the supervision of Yannis Trakadis who conceived and coordinated the project. Bill Qi and Yannis Trakadis designed the original methodology. Gustavo Turecki coordinated patient recruitment and data collection. Laura Fiori oversaw the production of the microRNA data and put together the Sample Processing and Small RNA-seq sections of the methodology section. All authors reviewed and provided feedback on the manuscript.

Funding

This work was supported by the Canadian Institutes of Health Research (CIHR) (grant number FRN/MOP#111260), as well as Janssen Research & Development. Gustavo Turecki holds a Canada Research Chair (Tier 1) and a NARSAD Distinguished Investigator Award. He is supported by grants from the CIHR (FDN148374 and EGM141899), and by the *Fonds de recherche du Québec -Santé* through the Quebec Network on Suicide, Mood Disorders and Related Disorders. Yannis Trakadis is supported by the McGill University Health Centre Research Institute and the Canada First Research Excellence Fund (McGill University Healthy Brains for Healthy Lives Initiative).

Acknowledgments

We acknowledge and thank Imane Bennani and Felicia Russo for helping with the review of the literature.

Ethical Statement

This study was approved by the Internal Review Board at the Douglas Mental Health University Institute.

Statement of Interest

None

Tables

Table 1. Model cross-validation and testing set AUC scores.

Analysis	**Mean AUC (SD) of trained model from cross-validation**	**Testing set AUC for final retrained model**
Classification of cases and controls	0.93 (0.06)	0.97
Classification of MDD severity grades	0.76 (0.11)	0.63
Classification of MDD severity grades – cluster 1	0.75 (0.18)	0.76
Classification of MDD severity grades – cluster 2	0.72 (0.15)	0.70
Classification of antidepressant response	0.622 (0.13)	0.57
Classification of antidepressant response – cluster 1	0.652 (0.085)	0.54
Classification of antidepressant response – cluster 2	0.670 (0.16)	0.49

Table 1. Model selection and evaluation were performed for each of the analyses listed in the table. The mean AUC across 5-folds of cross-validation during model training for the best model is presented, as well as the AUC from the evaluation on the testing set for the final retrained model.

AUC: Area under the receiver-operating characteristics curve

Table 2. Most important microRNA features used by the case-control classification model.

MicroRNA features ordered by decreasing importance (n=33)
hsa-miR-27a-3p, hsa-miR-197-3p, hsa-miR-22-5p, hsa-miR-221-3p, hsa-miR-126-3p, hsa-miR-128-1-5p, hsa-miR-30b-5p, hsa-miR-339-3p, hsa-miR-301a-3p, hsa-miR-345-5p, hsa-miR-505-3p, hsa-miR-1249, hsa-miR-132-3p, hsa-miR-550a-5p, hsa-miR-589-5p, hsa-miR-769-5p, hsa-miR-10b-5p, hsa-miR-210-3p, hsa-miR-628-3p, hsa-let-7d-3p, hsa-miR-148a-5p, hsa-miR-155-5p, hsa-miR-140-3p, hsa-miR-150-3p, hsa-miR-181a-5p, hsa-miR-24-3p, hsa-miR-629-5p, hsa-let-7a-3p, hsa-miR-194-5p, hsa-miR-28-3p, hsa-miR-378a-3p, hsa-miR-6852-5p, hsa-miR-7706

Table 2. The most important features used by the best performing machine learning model from the classification of cases and controls analysis, which could distinguish between cases and controls with an AUC of 0.97, were extracted and listed in order of decreasing importance.

AUC: Area under the receiver-operating characteristics curve

References

Bocchio-Chiavetto L, Maffioletti E, Bettinsoli P, Giovannini C, Bignotti S, Tardito D, Corrada D, Milanesi L, Gennarelli M (2013) Blood microRNA changes in depressed patients during antidepressant treatment. Eur Neuropsychopharmacol 23:602-611.

Chen C-J, Servant N, Toedling J, Sarazin A, Marchais A, Duvernois-Berthet E, Cognat V, Colot V, Voinnet O, Heard E (2012) ncPRO-seq: a tool for annotation and profiling of ncRNAs in sRNA-seq data. Bioinformatics 28:3147-3149.

Chen T, Guestrin C Xgboost: A scalable tree boosting system. In, pp 785-794: ACM.

Davami MH, Baharlou R, Ahmadi Vasmehjani A, Ghanizadeh A, Keshtkar M, Dezhkam I, Atashzar MR (2016) Elevated IL-17 and TGF-β Serum Levels: A Positive Correlation between T-helper 17 Cell-Related Pro-Inflammatory Responses with Major Depressive Disorder. Basic Clin Neurosci 7:137-142.

Dean B, Tsatsanis A, Lam LQ, Scarr E, Duce JA (2019) Changes in cortical protein markers of iron transport with gender, major depressive disorder and suicide. World J Biol Psychiatry:1-8.

Dwyer DB, Falkai P, Koutsouleris N (2018) Machine Learning Approaches for Clinical Psychology and Psychiatry. Ann Rev Clin Psych 14:91-118.

First MB, Spitzer RL, Gibbon M, Williams JBW (2012) Structured Clinical Interview for DSM-IV® Axis I Disorders (SCID-I), Clinician Version, Administration Booklet: American Psychiatric Association Publishing.

Fujii T, Yamamoto N, Hori H, Hattori K, Sasayama D, Teraishi T, Hashikura M, Tatsumi M, Okamoto N, Higuchi T, Kunugi H (2011) Support for association between the Ser205Leu polymorphism of p75NTR and major depressive disorder. J Hum Genet 56:806-809.

Gadotti VM, Bonfield SP, Zamponi GW (2012) Depressive-like behaviour of mice lacking cellular prion protein. Behav Brain Res 227:319-323.

Gao S, Calhoun VD, Sui J (2018) Machine learning in major depression: From classification to treatment outcome prediction. CNS Neurosci Ther 24:1037-1052.

Greenberg PE, Fournier A-A, Sisitsky T, Pike CT, Kessler RC (2015) The economic burden of adults with major depressive disorder in the United States (2005 and 2010). J Clin Psychiatry 76:155-162.

Hansen KF, Obrietan K (2013) MicroRNA as therapeutic targets for treatment of depression. Neuropsychiatr Dis Treat 9:1011.

Hasin DS, Sarvet AL, Meyers JL, Saha TD, Ruan WJ, Stohl M, Grant BF (2018) Epidemiology of adult DSM-5 major depressive disorder and its specifiers in the United States. JAMA Psychiatry 75:336-346.

Kim J, Ham S, Hong H, Moon C, Im HI (2016) Brain Reward Circuits in Morphine Addiction. Mol Cells 39:645-653.

Kim YK, Na KS, Shin KH, Jung HY, Choi SH, Kim JB (2007) Cytokine imbalance in the pathophysiology of major depressive disorder. Prog Neuropsychopharmacol Biol Psychiatry 31:1044-1053.

Knowland D, Lim BK (2018) Circuit-based frameworks of depressive behaviors: The role of reward circuitry and beyond. Pharmacol Biochem Behav 174:42-52.

Kozomara A, Griffiths-Jones S (2013) miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res 42:D68-D73.

Lopez JP et al. (2017) MicroRNAs 146a/b-5 and 425-3p and 24-3p are markers of antidepressant response and regulate MAPK/Wnt-system genes. Nat Commun 8:15497-15497.

Monti S, Tamayo P, Mesirov J, Golub T (2003) Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. Machine Learning 52:91-118.

Moylan S, Maes M, Wray NR, Berk M (2013) The neuroprogressive nature of major depressive disorder: pathways to disease evolution and resistance, and therapeutic implications. Mol Psychiatry 18:595-606.

Smalheiser NR, Lugli G, Rizavi HS, Torvik VI, Turecki G, Dwivedi Y (2012) MicroRNA expression is down-regulated and reorganized in prefrontal cortex of depressed suicide subjects. PLoS One 7:e33201.

Snaith RP, Harrop FM, Newby tDA, Teale C (1986) Grade Scores of the Montgomery—Åsberg Depression and the Clinical Anxiety Scales. Br J Psychiatry 148:599-601.

Song L, Florea L, Langmead B (2014) Lighter: fast and memory-efficient sequencing error correction without counting. Genome Biol 15:509.

Trakadis YJ, Sardaar S, Chen A, Fulginiti V, Krishnan A (2019) Machine learning in schizophrenia genomics, a case-control study using 5,090 exomes. Am J Med Genet B Neuropsychiatr Genet 180:103-112.

Verhoeven JE, Revesz D, Epel ES, Lin J, Wolkowitz OM, Penninx BW (2014) Major depressive disorder and accelerated cellular aging: results from a large psychiatric cohort study. Mol Psychiatry 19:895-901.

Vlachos IS, Zagganas K, Paraskevopoulou MD, Georgakilas G, Karagkouni D, Vergoulis T, Dalamagas T, Hatzigeorgiou AG (2015) DIANA-miRPath v3.0: deciphering microRNA function with experimental support. Nucleic Acids Res 43:W460-W466.

Weis S, Haybaeck J, Dulay JR, Llenos IC (2008) Expression of cellular prion protein (PrP(c)) in schizophrenia, bipolar disorder, and depression. J Neural Transm (Vienna) 115:761-771.

Zhao X, Kang J, Svetnik V, Warden D, Wilcock G, David Smith A, Savage MJ, Laterza OF (2019) A Machine Learning Approach to Identify a Circulating MicroRNA Signature for Alzheimer Disease. The Journal of Applied Laboratory Medicine 5:15-28.

Zheng LS, Hitoshi S, Kaneko N, Takao K, Miyakawa T, Tanaka Y, Xia H, Kalinke U, Kudo K, Kanba S, Ikenaka K, Sawamoto K (2014) Mechanisms for interferon-α-induced depression and neural stem cell dysfunction. Stem Cell Rep 3:73-84.

Bridging statement to Chapter 5

In Chapters 2 and 3, we developed ML methodologies for analyzing omics data to identify and understand the molecular basis of complex diseases.

In Chapter 4, as part of our second objective to investigate the application of ML for optimizing the treatment of complex diseases in terms of advancing precision medicine, we demonstrated the use of ML analysis of microRNA expression profiles in MDD for diagnosis, prognosis, and treatment response prediction. Our study provided valuable insights into the disorder's underlying molecular mechanisms and the potential of microRNA-based ML models in the clinical management of MDD patients. Overall, our results showed that microRNA is highly predictive of MDD status, and moderately predictive of depression severity. However, predicting depression severity proved more challenging. Lastly, we found that predicting antidepressant response using baseline microRNA expressions was not feasible with the data we had.

In Chapter 5, we continue exploring the application of ML in precision medicine, as part of optimizing the treatment of complex diseases with a broader scope. We extend our ML approach to predicting medication usage based on genetic data. This could potentially lead to improved treatment outcomes and reduced side effects for patients with complex diseases.

We focus on the analysis of genomic data, and specifically, targeted pharmacogenomic variants, which are more directly related to medication response. We also leverage the UK Biobank to maximize the sample size for our analyses. Moreover, we utilize a graph-based ML approach, specifically graph representation learning (GRL), to integrate interconnected biomedical entities in the form of a knowledge graph as part of our ML prediction model. This

approach allows us to leverage the wealth of existing biomedical knowledge to inform our predictions and improve the performance of ML models vs. using only tabular representations of data. Furthermore, we developed a ranking approach for interpreting medication usage odds, which provides a more interpretable and actionable output for clinicians and patients. This approach enables the prioritization of medications based on an individual's genetic data, potentially leading to more personalized and effective treatment decisions.

Chapter 4 note:

- Supplemental table 1 containing the description of samples exists as an Excel spreadsheet and can be obtained from *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7689198/*

Chapter 4 erratum:

- On pages 152-153, the procedure for determining the significance of the top miRNA maker for antidepressant response may be argued to be inaccurately applied since the p-values obtained after adjustment are corrected for only the top miRNAs prioritized from the ML analysis (out of the initial 285 miRNAs), which would increase the risk of false discoveries. From a multiple hypothesis testing perspective, it could be argued that a more stringent p-value adjustment would be to apply a correction with all 285 miRNAs considered as hypotheses tested out of which the top candidates were selected. In this case, none of the miRNA markers would pass the significance threshold after correction. Having said this, it's important to note that the supervised ML approach we used to analyze the 285 miRNA features is not exactly the same as multiple hypothesis testing. Correcting for multiple testing with all 285 miRNAs considered as hypotheses could be

overly conservative. This is because the ML model selection procedure, which includes cross-validation and testing evaluation, is already designed to reduce potential overfitting. This in turn decreases the risk of false discoveries due to chance, even before p-value correction.

Chapter 5. Graph representation learning for the prediction of medication usage in the UK Biobank based on pharmacogenetic variants

AUTHORS

Bill Qi[1], Yannis J. Trakadis[1,2]

AUTHORS INSTITUTIONAL AFFILIATIONS

[1] Department of Human Genetics, McGill University, Montreal, QC, Canada

[2] Department of Medical Genetics, McGill University Health Center, Montreal, QC, Canada

Correspondence to

Yannis J. Trakadis, MD MSc FRCPC FCCMG

Medical Geneticist & Metabolics Specialist

Assistant Professor, Human Genetics

McGill University Health Centre

Room A04.3140, Montreal Children's Hospital

1001 Boul. Décarie, Montreal, Quebec, Canada, H4A 3J1

Tel: (514) 412-4427, Fax: (514) 412-4296

Email**:** yannis.trakadis@mcgill.ca

Abstract

Ineffective treatment and side effects are associated with high burdens for the patient and society. We investigated the application of graph representation learning (GRL) for predicting medication usage based on individual genetic data in the United Kingdom Biobank (UKBB). Graph convolutional network (GCN) was used to integrate interconnected biomedical entities in the form of a knowledge graph as part of a machine learning (ML) prediction model. Data from The Pharmacogenomics Knowledgebase (PharmGKB) was used to construct a biomedical knowledge graph. Individual genetic data (n=485754) from the UKBB was obtained and preprocessed to match with pharmacogenetic variants in the PharmGKB. Self-reported medication usage labels were obtained from UKBB data field 20003. We hypothesize that joint analysis of all pharmacogenetic variants can predict the treatment response of individuals for different medications. Moreover, we assume that an individual using a medication on a regular basis experiences a net benefit from the medication (when factoring in treatment response and potential side effects), which is consistent with the continued/chronic use of the said medication. ML models were trained to predict medication usage for 264 medications. The GCN model significantly outperformed both a baseline logistic regression model (p-value: 1.53e-9) and a deep neural network model (p-value: 8.68e-8). The GCN model also significantly outperformed a GCN model trained using a random graph (GCN-random) (p-value: 5.44e-9). A consistent trend of medications with higher sample sizes having better performance was observed, and for several medications, a high relative rank of the medication (among multiple medications) was associated with greater than 2-fold higher odds of usage of the medication. In conclusion, a graph-based ML approach could be useful in advancing precision medicine by prioritizing medications that a patient may need based on their genetic data. However, further research is

needed to improve the quality and quantity of genetic data and to validate our approach using more reliable medication labels.

Keywords

Pharmacogenetics; Machine learning; Graph representation learning; Graph convolutional network

Introduction

To identify the most effective treatment for a patient affected by a given disease, typically, several medications need to be tried before an effective medication is identified. This causes an excessive burden on the patients by prolonging the suffering and decreasing the quality of life for patients. There are also major economic implications. Ineffective treatment and side effects of non-optimal treatment are associated with annual costs of around $495-672 billion in the US (i.e., 16% of total US healthcare expenditures in 2016) [1].

Pharmacogenomics (PGx) combines pharmacology with genomics with the aim of understanding how genes affect responses to drugs. Genetic associations of medication usage have been identified via genome-wide association studies. Wu et al. identified 505 linkage disequilibrium-independent genetic loci significantly associated with self-reported medication use from 23 medication categories utilizing self-reported medication-use data from the UK Biobank (UKBB) [2]. A more recent meta-analysis of the UKBB, Estonian Biobank, and FinnGen discovered 333 independent genetic loci associated with medication use patterns in hyperlipidemia, hypertension, and type 2 diabetes [3]. These studies, along with the numerous PGx variants and genes identified to be associated with medication response in the literature [4], provide evidence for the role of genetic variations in medication response and usage patterns and a strong case for their use in enabling precision (more targeted) medicine.

Despite the number of significant genetic associations of medication response and usage at the population level, studies leveraging the use of machine learning (ML) methods to predict medication traits at the individual level have been limited. Furthermore, most of the existing studies have been focused on the field of oncology and transcriptomic data [5, 6]. However, in a recent publication from Taliaz et al., successful use of ML classifiers has been reported for

predicting responses to specific antidepressants (citalopram, sertraline, and venlafaxine) in major depressive disorder using genetic, clinical, and demographic features [7].

Considering the potential of ML analysis of genetic data to enhance the treatment of patients through medication-related predictions, we explored the application of ML to a broader range of medications. The goal of our study is to develop a machine learning (ML)-based model for medication usage prediction based on genotypes of patients in the UKBB. Our study will address the challenge of reducing the amount of trial-and-error burden on patients by predicting and prioritizing medications a patient is likely to use and benefit from. A key assumption underlying our study involves the use of self-report data on medication usage. Since self-reported medication usage from the UKBB includes only regular medications and health supplements (taken weekly, monthly, etc.), we assume that an individual using a medication on a regular basis experiences a net benefit from the medication (when factoring in treatment response and potential side-effects), which is consistent with the continued/chronic use of the said medication.

Furthermore, we explored the use of a graph representation learning (GRL) approach to integrate interconnected biomedical entities in the form of a knowledge graph as part of the ML prediction model. There are several successful examples of applications of GRL in the biomedical domain at the molecular, genomic, therapeutic, and healthcare levels [8]. However, to our knowledge, no study has applied a GRL approach for medication usage prediction based on individual genotype data. A biomedical graph capturing the relationships between genetic variants, genes, diseases, and medications could potentially increase the intelligence of ML models and prediction performance by introducing biomedical domain knowledge and dependencies between entities.

Methods

Graph convolutional network

The class of GRL we utilized is the graph convolutional network (GCN) introduced by Kipf & Welling [9]. GCNs extend the ideas of convolution neural networks defined for Euclidean data to irregular graph data. Therefore, we can apply existing neural network operations on graph objects. To illustrate how the GCN works, we first define a graph as consisting of a set of nodes and a set of edges between nodes.

The goal of the GCN is to learn a function (f) of the node features and graph structure. The node features are provided by a feature matrix X with dimensions $N \times D$, where N is the number of nodes in the graph, and D is the number of features per node. The graph structure is represented using an $N \times N$ dimensional adjacency matrix A, and an element A_{ij} in the matrix denotes whether an edge exists between nodes i and j.

The function f is a neural network layer consisting of a learnable weight matrix W and a non-linear activation function σ, and takes as input the node feature matrix X and adjacency matrix A:

$$Z = f(X, A) = \sigma(\widehat{D}^{-\frac{1}{2}}\hat{A}\widehat{D}^{-\frac{1}{2}}XW)$$

Here, $\hat{A} = A + I$, where I is the identity matrix. $\widehat{D}$ is the diagonal node degree matrix of $\hat{A}$. The output of f is a matrix Z with dimensions $N \times F$, where F is the dimensionality of the transformed output. The product of X and W is a transformed node feature matrix, and subsequent multiplication with $\widehat{D}^{-\frac{1}{2}}\hat{A}\widehat{D}^{-\frac{1}{2}}$ results in a normalized aggregation of all the transformed neighboring nodes for each node and itself.

The GCN consists of multiple stacked layers of transformations and aggregations for increased expressivity. With $X = H^{(l)}$, $W = W^{(l)}$, and $Z = H^{(l+1)}$, where l is an index denoting the layer of the GCN model, we obtain the layer-wise propagation rule introduced in [9]:

$$H^{(l+1)} = f\left(H^{(l)}, A\right) = \sigma(\widehat{D}^{-\frac{1}{2}}\hat{A}\widehat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)})$$

The final layer output of the GCN is taken as the embeddings of the nodes in the graph capturing the local neighborhood structure around each node. The embeddings can then be used in downstream tasks (e.g., node classification or link prediction between nodes).

Biomedical knowledge graph

The graph data source we utilized for GRL is the Pharmacogenomics Knowledge Base (PharmGKB) [4]. PharmGKB consists of curated relationships between variants, genes, medications, and diseases extracted from PubMed articles using manual curation with the support of natural language processing. The "relationships" table in PharmGKB summarized the curated edges between pairwise entities. From the "relationships" table, we filtered for all edges which have the attribute "associated," and discarded those which are "ambiguous" or "not associated, to create a new "filtered relationships" table. Based on the table, a visualization of the graph was generated using the ForceAtlas2 algorithm [10], and partitioned using a community detection algorithm [11], in Gephi (v. 0.10.1) [12]. The visualization of the graph is shown in Supplemental Figure 1.

Individual genotype data

For individual genetic data, we used the imputed genotypes from the UKBB (~96 million variants, aligned to + strand of GRCh37 reference genome). We used PLINK2 [13, 14] to

remove variants with a Hardy-Weinberg equilibrium test p-value lower than 1e-15, as well as variants with missing call rates exceeding 0.1. We limited our analysis to the autosomes (chromosomes 1 to 22). All variant coordinates were subsequently uplifted to GRCh38, to be consistent with PharmGKB coordinates. Further sample-based filtering was applied based on UKBB recommendations, including the following data fields: 22010 (recommended genomic analysis exclusions), 22019 (sex chromosome aneuploidy), 22027 (outliers for heterozygosity or missing rate), and 22051 (UKBiLEVE genotype quality control for samples). A list of individuals withdrawn from the UKBB as of February 22nd, 2022 was removed. After filtering, a total of 485754 individuals remained.

Variants present in the PharmGKB were matched to those present in the UKBB imputed genotypes dataset. To maximize the number of relevant variants included in the analysis, we matched with variants annotated in the "variants" table in PharmGKB. For variants not part of the "filtered relationships" table, we created additional ad-hoc variant-to-gene and gene-to-variant edges, if the gene corresponding to the variant exists in the "filtered relationships" table. Overall, a total of 3962 variants overlapped between the PharmGKB and UKBB datasets.

Furthermore, we used PGxPOP (https://github.com/PharmGKB/PGxPOP) to obtain inferred haplotype calls based on allele definitions for PGx genes [15]. These haplotypes were matched with haplotypes present in the "filtered relationships" table. For any haplotypes that were not present in the "filtered relationships" table, we created additional ad-hoc edges linking the haplotype to its corresponding gene. Overall, a total of 175 haplotypes overlapped between the PharmGKB and UKBB datasets.

As a final filtering step to reduce the number of uninformative features for ML analysis, we filtered out any features from the UKBB genotype data with less than 10 non-zero values, resulting in 3890 variant and 156 haplotype features used for subsequent ML analysis.

Supplemental Figure 2 provides a summary of the types of nodes present in the final PharmGKB graph used in our analysis. Supplemental Figure 3 provides a summary of the edge types that exist between node types. The graph is undirected, thus for every edge from node i to j an opposing edge exists from node j to i.

Medication usage labels

We obtained medication usage data from UKBB data field 20003 (treatment/medication code). The medication codes are derived from self-reported regular treatments (taken weekly, monthly, etc.). Thus, we assume that an individual using a medication on a regular basis experiences a net benefit from the medication (when factoring in treatment response and potential side effects), which is consistent with the continued/chronic use of the said medication. To create medication usage labels, we obtained a mapping of UKBB medication codes to ATC codes and active ingredient names from the supplementary data in the study by Wu et. al [2]. Next, for each patient, we converted their UKBB medication codes to ATC codes and active ingredient names. Finally, we filtered for only medications present in the PharmGKB based on a match with either the ATC code or active ingredient name of the medication. To ensure an adequate number of samples for ML analysis, we kept medications taken by at least 100 patients. In total, 264 medications met the threshold. For each patient, a 264-dimensional vector (y_i) is created with a value of 1 indicating that the patient is taking the medication, and 0 otherwise.

GCN model architecture

A supervised ML approach based on GCN was developed for the prediction of medication usage based on individual genotype data. Our approach consists of three steps: 1) learning node embedding using graph convolutions, 2) aggregation of feature node embeddings weighted by feature values, 3) prediction of class probabilities. Figure 1 shows an illustration of the approach. In step 1, embeddings for nodes in the graph are learned through graph convolutions, i.e., normalized aggregation of all the transformed neighboring nodes for each node and itself (example shown for node V_1). The initial features for each node (X) are initialized as an identity matrix. The reasoning for using an uninformative identify matrix is so that the model will leverage only structural information of the graph as part of the predictions, thus emphasizing the benefits of the specific graph structure, and providing a fair comparison against non-graph-based ML approaches. In step 2, the embeddings of nodes corresponding to features (i.e., genetic variant nodes V_1, V_2, V_3, and V_4), are multiplied with the feature values (i.e., genotypes of a patient i (G_i)) to yield a feature value-weighted aggregation of the embeddings, i.e., patient-specific aggregated embedding (Z_i). In step 3, Z_i is fed into a neural network model to output class probabilities (i.e., predict the medication(s) that patient i is taking). The neural network model can be viewed as a function of the genotype inputs and the graph convolution embeddings (which itself is a function of the PharmGKB graph). The architecture consists of fully differentiable operators and the parameters of the neural network model, and the graph convolution layers are jointly optimized through an iterative process backpropagation and gradient descent.

ML analysis

We developed supervised ML models using the 3890 variants and 156 haplotypes extracted as described above as features. We performed cross-validation with the data split as 70% training, 10% validation, and 20% testing using a stratified split of individuals to maintain a balanced distribution of each medication across the splits. For cross-validation, we allocated 340027 individuals for training and 48576 for validation. A final set of 97151 individuals were not used in the cross-validation process and were reserved for the final evaluation of the trained models.

To assess the GCN model performance, we compared it with a logistic regression (baseline), a regular deep neural network (DNN), and a GCN model with a randomized graph (GCN-random). We defined the GCN model with two graph convolution layers consisting of a 256-dimensional output (i.e., $H^{(l)}$) with rectified linear unit (ReLU) as the non-linear activation function. The neural network model composed over the graph convolution outputs consists of 2 fully-connected layers (i.e., each layer has a linear transformation of their input followed by non-linear activation of the transformed input) with 512-dimensional outputs, followed by a layer with 264 nodes and sigmoid activation to obtain the final prediction output as probabilities. The DNN model has the same architecture as the neural network model except genotype values are used as the input rather than the output from graph convolution integrated with genotypes. The GCN-random model is the same as the GCN model except the original PharmGKB graph edges are randomly permuted such that the associations information between nodes are randomized.

All models were trained and evaluated on the same dataset splits. We trained the model using the following multilabel binary cross-entropy loss function with minibatch stochastic gradient descent (SGD) using the Adam optimizer [16]:

$$Loss_{BCE} = -\frac{1}{n}\sum_{i=0}^{n}\sum_{j=0}^{k}(y_{ij} * \log(\hat{y}_{ij}) + (1 - y_{ij}) * \log(1 - \hat{y}_{ij}))$$

where n is the number of patients in each minibatch for SGD (n=4096), k is the number of medications (k=264), y_{ij} is the true label value for patient i and medication j, and $\hat{y}_{ij}$ is the model prediction for patient i and medication j.

To select the final model used for evaluation, we ran minibatch SGD until convergence, and selected the model with the lowest $Loss_{BCE}$ value on the validation set. The evaluation metric used to assess and compare model performance is the area under the receiver operating characteristic curve (AUC) on the testing set. An AUC value is calculated for each medication independently. Furthermore, we performed comparisons between approaches using paired (i.e., for each medication) t-tests to assess whether the model performance differences are statistically significant. Lastly, we analyzed the mean AUCs across five medication sample size percentile ranges to assess whether there is a relationship between the number of users of a medication and model performance.

Ranking interpretation of model predictions

We performed a downstream analysis assessing the significance of medication usage predictions. Using a ranking approach, we examined whether an individual having a higher rank for a medication (out of all medications) means higher odds of taking the medication. First, for each medication, we standardized the predicted probabilities for each patient into Z-scores with a mean of 0 and standard deviation of 1. Next, for each patient, the Z-scores are converted into ranks (i.e., the higher the Z-score, the higher the rank). For each medication, we use logistic regression to assess the association of having a rank value of the medication ranked in the top 5

out of all medications with actual medication usage. We limited the analysis to medications with a sample size of at least 5000.

Results

We selected the models with the lowest $Loss_{BCE}$ value on the validation set for each of the compared approaches (i.e., Baseline, DNN, GCN, and GCN-random). Final evaluations of the selected models were performed on the testing set. The overall mean AUC values on the testing set over all medications for each approach are shown in Table 1. We also provide an overview of the prediction performance over all medications in Figure 2 using a swarm plot of the AUC value of each medication. Overall, the GCN model outperformed the Baseline and DNN models. Using paired t-tests, we find statistically significant improvements of the GCN model over baseline (p-value: 1.53e-9) and DNN (p-value: 8.68e-8). Furthermore, the GCN model significantly outperforms the GCN-random model (p-value: 5.44e-9), suggesting that a GCN model utilizing a specific graph structure is significantly better than one with a randomized graph. In summary, the prediction AUCs are statistically better than what would be expected from null predictions, however, in absolute terms, the prediction performance of all approaches was low, ranging from an AUC of 0.510 to 0.527.

To explore how model prediction performance is related to the medication sample size (i.e., do medications with a higher number of users have higher AUCs), we plotted the mean AUC value of each bin after binning medications into percentile ranges (Figure 3). Overall, we observe an increase in mean AUC as medication sample size increases. We also observe that the performance of the GCN model is the highest in all five percentile ranges.

Despite the overall prediction performance in terms of AUC being low, our analysis using the medication ranking approach (described in *Ranking interpretation of model predictions*) revealed a significant association of medication usage with having a high rank (having a rank in the top 5 relative to all other medications) for many medications. Figure 4 shows the odds ratio of the associations accompanied by 95% confidence intervals for each medication with a sample size of at least 5000. Furthermore, having a high rank for several medications is associated with greater than 2-fold higher odds of usage of the medication (e.g., iron, insulins and analogues, metformin).

Discussion

We introduced a novel application of a graph-based ML approach for medication usage prediction using curated biomedical domain knowledge from the PharmGKB and targeted pharmacogenetic data from 485754 individuals from the UKBB. Our findings revealed the predictability of 264 commonly used medications based on PGx features (3890 variants and 156 haplotypes). Although the overall performance for medication usage prediction was low, we found that the GCN approach outperformed all other models in our comparison, including the same GCN approach with a random graph, suggesting that the graph structure contained specifically in the PharmGKB graph could be useful to improve the performance of prediction models.

The relatively poor performance of the models could be due to a combination of different factors. The first is that of low medication sample sizes for the majority of medications. In support of this factor, we saw a consistent trend of medications with higher sample sizes having better performance. Second, our input genetic variants are limited in terms of quality of

imputation and quantity (in the order of thousands). It may be important to further enrich the current graph to include additional sources of genetic variants and genes with impacts on the metabolism of medications. One potential approach would be to leverage the use of pharmacogenomic variant effect classifiers at the variant level [17], to prioritize relevant variants as features for medication trait prediction at the individual patient level. Third, the use of self-reported medication use labels may not be an accurate reflection of response to the medication. To avoid making assumptions about the training labels, data from randomized controlled trials (RCT) or N-of-1 clinical trials [18], capturing medication responses and side effects at the individual level would be required. However, RCTs are expensive to conduct and often have low sample sizes. As a future direction, it may be feasible to explore transfer learning [19], to leverage the knowledge stored in the current model when training a new model from high-quality RCT or N-of-1 data.

Despite the relatively poor prediction of individual medications, we found that for several medications, a patient with a higher relative rank of the medication (among all medications) is significantly more likely to be using the medication. The ranking interpretation of model predictions could also be more clinically meaningful in the context of precision medicine by providing relative ranks of medications (e.g., for comparing between multiple potential medications).

Using a ranking interpretation of model predictions, we highlighted several medications for which our method had the highest performance, including iron, insulin, and metformin, all of which are widely used medications. Iron supplementation is used in the treatment of iron deficiency of varying causes including nutritional deficiency, malabsorption, chronic inflammatory state, blood loss, and others [20]. Insulin is used in the treatment of type 1 diabetes

and less commonly in type-2 diabetes by binding to insulin receptors on cells to reduce blood glucose levels [21]. Similarly, metformin is used primarily for the treatment of type-2 diabetes through several mechanisms, including decreasing glucose production in the liver and intestinal absorption, as well as increasing insulin sensitivity [22]. All three medications are associated with greater than 2-fold higher odds of usage of the medication given a rank of the medication within the top five, suggesting that individuals with specific genetic profiles have a higher likelihood of needing these medications. While our method is able to identify individuals requiring iron supplementation, our current method cannot identify individuals who would experience adverse effects of medications due to a lack of ground truth labels for these events. To capture adverse effects in future studies, integration of data on medications and adverse events from resources such as electronic health records with natural language processing methods would be required [23]. Lastly, adjustments to the current graph approach, such as enriching the current biomedical knowledge graph with adverse effect nodes and edges to medications, could enhance the prediction of adverse effects of medications.

In conclusion, a graph-based ML analysis of targeted pharmacogenomic variants could be useful in advancing precision medicine by prioritizing medications that a patient may need based on their genetic data. This could help reduce the need for trial-and-error in finding an effective medication. However, further research is needed to validate the findings, and future studies should focus on improving the quality and quantity of genetic data and medication response data, including more informative genetic variants involved in the metabolism of medications and more reliable medication labels.

Acknowledgments

We thank the United Kingdom Biobank (UKBB), the Healthy Brains, Healthy Lives (HBHL) program, Neurohub, and The Pharmacogenomics Knowledgebase (PharmGKB) for providing access to the datasets used in our study. We would also like to thank Dr. Celia Greenwood and Dr. Jeff Xia for their suggestions and feedback on the methodology and results.

Contributors

Bill Qi performed the machine learning analyses and drafted the manuscript under the supervision of Yannis Trakadis who conceived and coordinated the project. Bill Qi and Yannis Trakadis designed the original methodology. All authors reviewed and provided feedback on the manuscript.

Availability of data and materials

The genotype and phenotype datasets used in the preparation of this manuscript were obtained from the United Kingdom Biobank (UKBB) through a partnership with the Healthy Brains, Healthy Lives (HBHL) program and Neurohub. The data used to create the graph are available at The Pharmacogenomics Knowledgebase (PharmGKB).

Role of funding source

Yannis Trakadis is supported by the McGill University Health Centre Research Institute and the Canada First Research Excellence Fund (McGill University Healthy Brains for Healthy Lives Initiative). *The funding source had no further role in study design; in the collection, analysis and*

interpretation of data; in the writing of the report; and in the decision to submit the paper for publication.

Conflict of interest

None

Tables and figures

Figure 1. GCN model architecture.

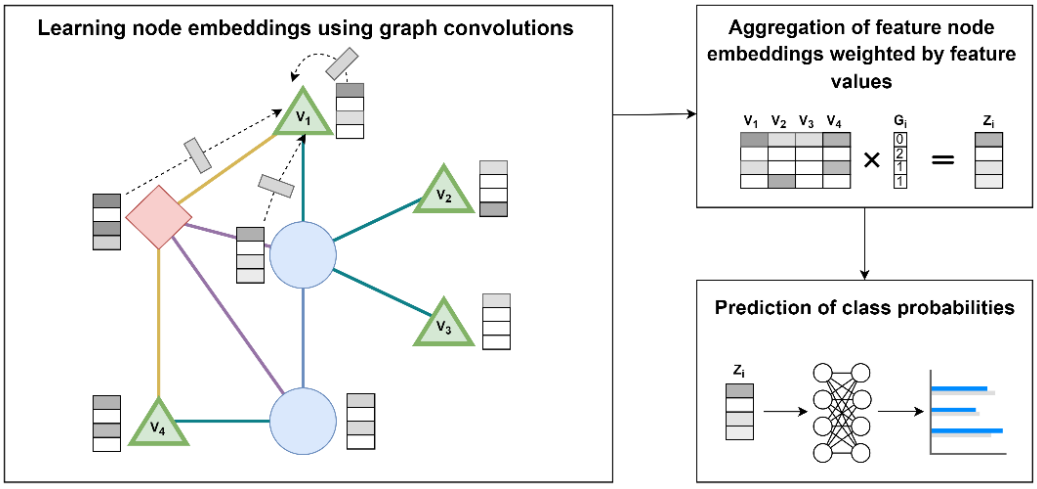


Figure 1. A simplified illustration of the three steps of the GCN model architecture as described in the section *GCN model architecture* is shown. The green nodes represent genetic variants in the graph, while the blue and red nodes are other node types in the graph such as genes or medications.

Figure 2. Distribution of testing set AUC values for each approach.

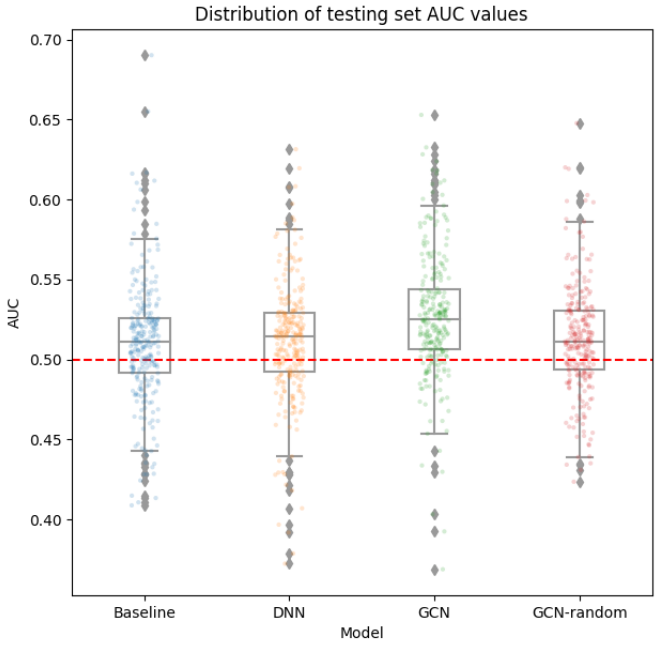


Figure 2. The performance of each of the final models for each approach is shown. The y-axis shows the AUC values while the x-axis are labels indicating each approach. The box plots overlayed show a summary of the AUC ranges, while each point in the strip plot corresponds to the AUC value for a specific medication. The GCN approach significantly outperforms the Baseline (p-value: 1.53e-9), DNN (p-value: 8.68e-8), and GCN-random (p-value: 5.44e-9) approaches in terms of AUC values, albeit the actual AUC values were low.

Figure 3. Mean AUC at each medication sample size percentile range.

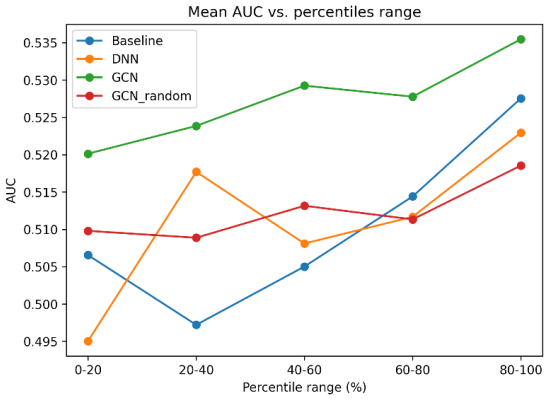


Figure 3. The relationship between medication sample size and model prediction performance is shown. The x-axis indicates a percentile range of bins of medication sample sizes. Each percentile range is a grouping of medications with sample sizes falling in the range. For example, a percentile range of 0-20 includes all medications which have a sample size that falls within the lowest 20% percentile out of all medications. The y-axis indicates the mean AUC of all medications within the percentile range. We noted that as the medication sample size increases, the mean AUC value also increases, suggesting that prediction performance is higher for medications with a higher number of users. Moreover, the GCN model consistently demonstrates the highest performance across all five percentile ranges.

Figure 4. Odds ratio between usage of a medication and having a rank value within the top five.

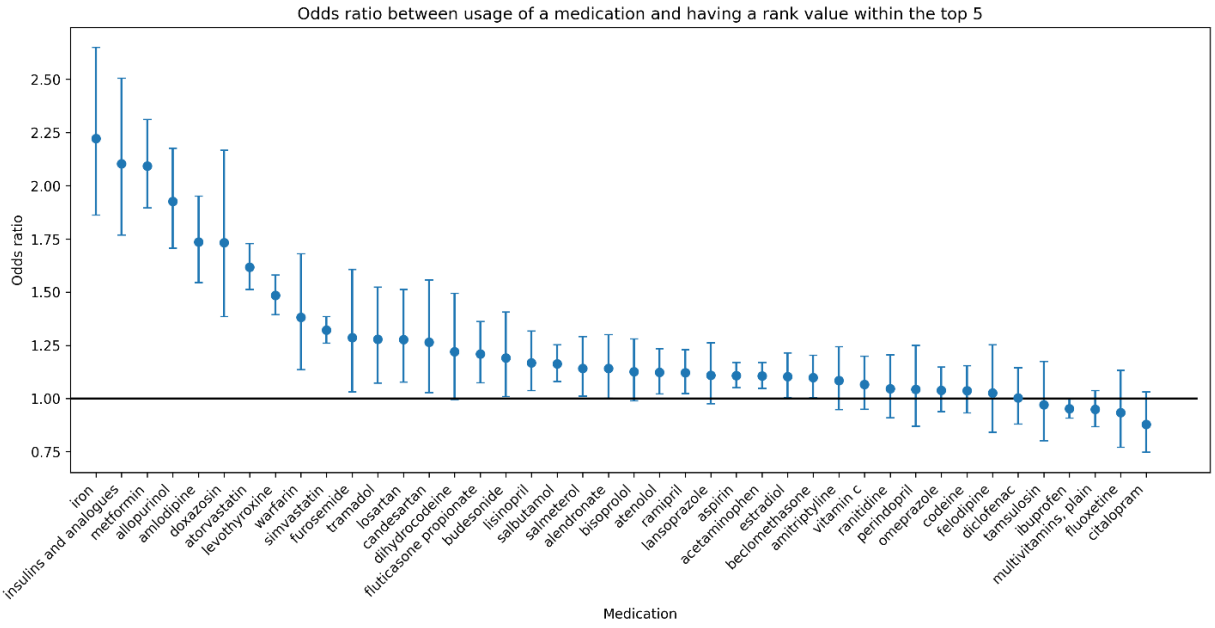


Figure 4. The odds ratio of the associations between the usage of a medication and having a rank value within the top five are shown for each medication with a sample size of at least 5000. The x-axis indicates each medication with a sample size of at least 5000. The y-axis indicates the corresponding odds ratio for having a rank within the top five for the corresponding medication. The error bars around each point indicate lower and upper bounds of the 95% confidence interval of the odds ratio estimate. Medications are sorted from highest to lowest by odds ratios. We noted that having a high rank for several widely used medications, including iron, insulins and analogues, and metformin, is associated with greater than 2-fold higher odds of usage of the medication.

Table 1. Summary of model performance.

Approach	Mean AUC
Baseline	0.510
DNN	0.511
GCN	0.527
GCN-random	0.512

References

1. Watanabe, J.H., T. McInnis, and J.D. Hirsch, *Cost of Prescription Drug-Related Morbidity and Mortality.* Ann Pharmacother, 2018. **52**(9): p. 829-837.
2. Wu, Y., et al., *Genome-wide association study of medication-use and associated disease in the UK Biobank.* Nat Commun, 2019. **10**(1): p. 1891.
3. Kiiskinen, T., et al., *Genetic predictors of lifelong medication-use patterns in cardiometabolic diseases.* Nature Medicine, 2023. **29**(1): p. 209-218.
4. Thorn, C.F., T.E. Klein, and R.B. Altman, *PharmGKB: the Pharmacogenomics Knowledge Base.* Methods Mol Biol, 2013. **1015**: p. 311-20.
5. Adam, G., et al., *Machine learning approaches to drug response prediction: challenges and recent progress.* npj Precision Oncology, 2020. **4**(1): p. 19.
6. Golriz Khatami, S., et al., *Using predictive machine learning models for drug response simulation by calibrating patient-specific pathway signatures.* npj Systems Biology and Applications, 2021. **7**(1): p. 40.

7. Taliaz, D., et al., *Optimizing prediction of response to antidepressant medications using machine learning and integrated genetic, clinical, and demographic data.* Translational Psychiatry, 2021. **11**(1): p. 381.

8. Li, M.M., K. Huang, and M. Zitnik, *Graph Representation Learning in Biomedicine.* arXiv preprint arXiv:2104.04883, 2021.

9. Kipf, T.N. and M. Welling, *Semi-supervised classification with graph convolutional networks.* arXiv preprint arXiv:1609.02907, 2016.

10. Jacomy, M., et al., *ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software.* PloS one, 2014. **9**(6): p. e98679.

11. Blondel, V.D., et al., *Fast unfolding of communities in large networks.* Journal of statistical mechanics: theory and experiment, 2008. **2008**(10): p. P10008.

12. Bastian, M., S. Heymann, and M. Jacomy. *Gephi: an open source software for exploring and manipulating networks*.

13. Chang, C.C., et al., *Second-generation PLINK: rising to the challenge of larger and richer datasets.* GigaScience, 2015. **4**(1): p. s13742-015-0047-8.

14. Shaun Purcell, C.C., *PLINK v2.00a3.3*.

15. McInnes, G., et al., *Pharmacogenetics at Scale: An Analysis of the UK Biobank.* Clinical Pharmacology & Therapeutics, 2021. **109**(6): p. 1528-1537.

16. Kingma, D.P. and J. Ba, *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980, 2014.

17. Pandi, M.T., et al., *A novel machine learning-based approach for the computational functional assessment of pharmacogenomic variants.* Hum Genomics, 2021. **15**(1): p. 51.

18. Lillie, E.O., et al., *The n-of-1 clinical trial: the ultimate strategy for individualizing medicine?* Per Med, 2011. **8**(2): p. 161-173.

19. Tan, C., et al. *A survey on deep transfer learning*. in *International conference on artificial neural networks*. 2018. Springer.

20. Nguyen, M. and P. Tadi, *Iron supplementation.* 2020.

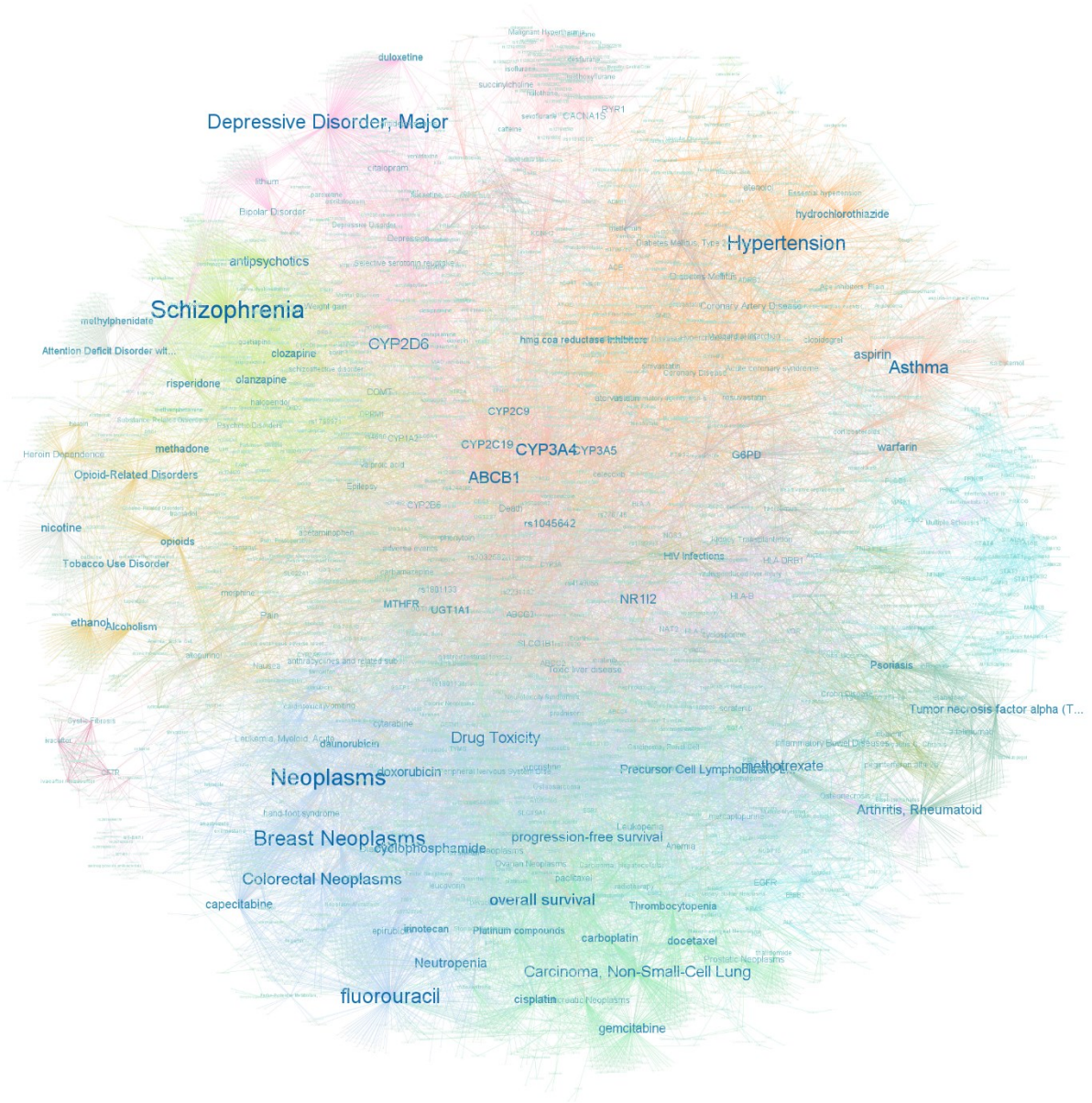
21. Thota, S. and A. Akbar, *Insulin*, in *StatPearls [Internet]*. 2022, StatPearls Publishing.

22. Corcoran, C. and T.F. Jacobs, *Metformin.* 2018.

23. Feng, C., D. Le, and A.B. McCoy, *Using Electronic Health Records to Identify Adverse Drug Events in Ambulatory Care: A Systematic Review.* Appl Clin Inform, 2019. **10**(1): p. 123-128.

Supplemental Figures

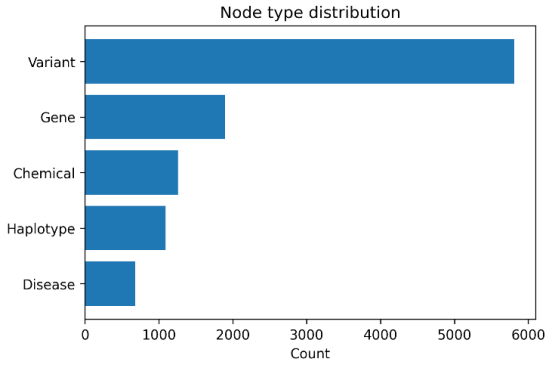
Supplemental Figure 1. Visual representation of the PharmGKB graph.



Supplemental Figure 1. A visual representation of the PharmGKB graph is shown. The size of nodes in the graph is defined by their degree (i.e., the number of edges a node has). Nodes are colored by the community they belong to as derived from community detection. The figure is

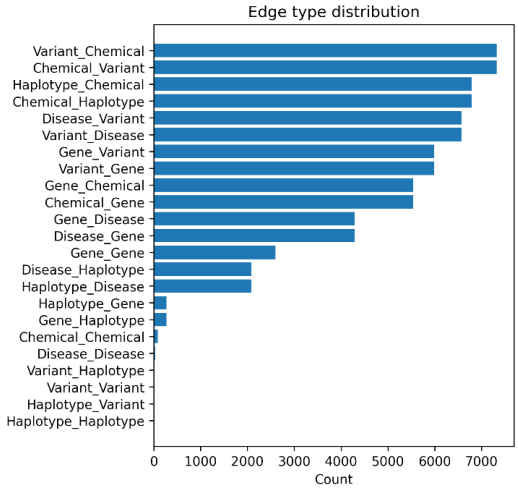
meant to highlight the overall patterns of the PharmGKB graph through a projection into two dimensions. One immediately notable pattern is that similar diseases tend to be clustered more closely.

Supplemental Figure 2. Node types present in the final PharmGKB graph.



Supplemental Figure 2. The y-axis shows each of the node types present in the final PharmGKB graph, while the x-axis shows the corresponding count. Nodes are sorted by highest to lowest count.

Supplemental Figure 3. Edge types present in the final PharmGKB graph



Supplemental Figure 3. The y-axis shows each of the edge types present in the final PharmGKB graph, while the x-axis shows the corresponding count. Edge type labels denote the pair of node types involved in the edge. Since the PharmGKB graph is undirected, an equal number of edges exist for every edge type in the reverse direction. Edge types are sorted by highest to lowest count. Medications are labeled as “Chemical” in the PharmGKB graph.

Chapter 6. General Discussion

Our overarching hypothesis of this thesis was that ML analysis of omics data can provide valuable insights into the pathophysiology of complex diseases and aid in their treatment. The objectives of this thesis were to explore the potential of ML approaches for identifying and understanding the molecular basis of complex diseases (using psychiatric disorders as our model), and to investigate the application of ML for optimization of the treatment of complex diseases. We have investigated the role of ML in achieving these objectives using various omics data sources such as gene expression, microRNA, and genetic data. We have included four studies as part of this thesis, each showcasing the effectiveness of ML analysis of omics data for producing novel insights.

General overview of studies

In our first study (Chapter 2), we tested the hypothesis that our ML methodology could effectively distinguish disease cases from controls based on gene expressions from the dorsolateral prefrontal cortex (DLPFC) better than random chance and generalize to unseen data. We utilized gene expression microarray data from the DLPFC of post-mortem SCZ cases and controls and developed a supervised ML pipeline using the XGBoost algorithm. Our results demonstrated above-chance performance in classification and generalization, with an average AUC of 0.76 on cross-validation and testing data. We also identified genes that were significantly relevant to SCZ through the integration of ML and biological gene set analysis. These initial findings supported our hypothesis that ML analysis of gene expressions can enhance our understanding of complex diseases and identify potential treatment targets. Of note,

we used transcriptomic data from the DLPFC, and although the DLPFC is a significant brain region associated with SCZ, it is important to recognize that other brain areas are also involved [111], as well as other systems outside of the brain [112], which may account for the non-perfect separation of diseases cases and controls.

Building on the findings from the first study, in Chapter 3, we further validated our hypothesis that patterns identified through ML are relevant to disease and generalizable, and strengthened our overall approach, with a case-control analysis in MDD. More specifically, we applied supervised ML to gene expression data from the DLPFC of post-mortem MDD cases and controls, achieving an average AUC of 0.72 on 10-fold cross-validation and 0.76 on testing data. In this case, we included an external validation of our initial findings using an independent cohort, observing an AUC of 0.62. Furthermore, using covariate information, we were able to identify an association between smoking and recall performance in MDD case prediction, showing that the inclusion of covariate information for assessing model performance can lead to further insights in a ML analysis. Overall, Chapter 3 shows a further refinement of the ML methodology introduced in Chapter 2, and the findings further demonstrate the effectiveness of ML applied to complex diseases. However, our analysis of blood gene expression data did not yield promising results in MDD, underlining an obstacle towards clinical translation of the approach, and suggesting that different tissues may yield different results, and this choice should be carefully considered in future studies.

In our third study (Chapter 4), we hypothesized that ML analysis of omics data could yield biomarkers for diagnosis, disease severity, and treatment response. We began our exploration into the potential application of ML for the optimization of the treatment of complex diseases. We applied supervised and unsupervised ML approaches to analyze blood microRNA

expression profiles from an MDD case-control dataset. We successfully distinguished MDD cases from healthy controls with an AUC of 0.97 on testing data. Furthermore, in addition to analyzing disease status, we also performed an analysis of disease severity levels. We were able to distinguish high vs. low severity individuals with an AUC of 0.63, partially supporting our hypothesis. We also found that unsupervised clustering of patients may improve the performance of the classifiers in predicting MDD severity. However, we could not separate antidepressant responders from non-responders with high accuracy, indicating that further improvements in our ML methodology may be needed. Of note, a potential inaccuracy in our application of false-discovery adjustment in the study which determined the statistical significance of the top miRNA marker for antidepressant response (see Chapter 4 erratum) was highlighted after the publication of the manuscript and added as note at the bottom of page 168. Due to this, it is possible that our reported significance is under-corrected and an over-estimation of the actual significance.

In our final study (Chapter 5), we hypothesized that the incorporation of pharmacogenomic domain knowledge as part of ML modelling could enhance performance in the prediction of medication usage and advance precision medicine. We continued our exploration into the application of ML for optimizing treatment by developing a GRL for predicting medication usage based using individual genetic data in the UKBB. We used a GCN model to integrate interconnected biomedical entities in the form of a knowledge graph as part of an ML prediction model. Our GCN model significantly outperformed both a baseline logistic regression model and a deep neural network model, demonstrating the benefits of integrating biomedical domain knowledge and the potential of a graph-based ML approach in advancing precision medicine by prioritizing medications that a patient may need based on their genetic data. These results support our initial hypothesis. Furthermore, although we limited the

application of the graph-based ML approach to genetic data, our findings provide a basis for future multi-omics integration in ML analysis. To achieve this, we would just need to introduce connections between different omics types in a graph, in addition to genetic data.

Original contributions to knowledge

Overall, the work presented in this thesis contributes to the growing body of research leveraging the use of ML approaches in understanding the molecular basis of complex diseases [113], and treatment optimization [114-116]. Specifically, we have contributed several novel methodologies, including (1) the ML model selection procedure for high-dimensional omics data introduced in Chapter 2, (2) the gene set analysis method of combining differential gene expression statistics with genes selected through supervised ML (Chapters 2 and 3), (3) the use of unsupervised clustering to reduce heterogeneity in the data prior to supervised ML analysis for disease severity prediction (Chapter 4), and (4) the graph approach for integrating biomedical domain knowledge in ML analysis of genetic data (Chapter 5).

Although our classification results in the studies presented in Chapters 2 and 3 were not high compared to previous studies, our study consisted of a greater number of cases and controls and the use of a training/validation and testing set split, as well as external testing data for final evaluation, which reflects a more precise estimation of model performance and a more realistic estimation of the generalizability of the patterns learned by the model.

We noted that model selection was a major challenge of these studies given the large number of features of our input data (~20000 genes). As we have seen, even based on the final set of genes prioritized by XGBoost, we saw that further classification improvements can be

obtained through an additional Bayesian optimization step to explore and identify an ensemble model, suggesting that further refinements to model selection are still possible. This two-stage method of model selection may be particularly useful for future applications of ML to high-dimensional omics data.

Furthermore, we emphasized the molecular basis of the respective diseases through the development of a novel gene set analysis (GSA) technique combining both a statistical testing method and the results from the ML analysis. The GSA method, when integrated with the findings from ML models, was able to enhance the biological interpretation of the results but also increased their robustness by reducing potential false-positive results through consensus of the two methods. This synergistic approach allowed us to identify significant molecular functions which were associated with the respective diseases in existing literature, thus providing further support for the effectiveness of the approach. Moreover, the identification of novel associations could enable us to uncover molecular targets that may be implicated in the pathophysiology of complex diseases and enable the development of new therapeutic strategies.

Chapters 4 and 5 highlighted the usefulness and limitations of ML in terms of the optimization and treatment of complex diseases. Our findings suggest that ML may be useful for the purposes of complex disease diagnosis and monitoring based on omics data as demonstrated in Chapter 4. However, we underlined several obstacles in using ML for direct treatment response and usage prediction based on omics data, and further improvements to data quality, pharmacogenomic knowledge, and analysis methodology would be necessary to address this challenge.

The study in Chapter 5 departs notably from the previous studies in both the scope, objective, and methodology, with a focus on the use of genetic data compared to transcriptomics.

The scope of the analysis in Chapter 5 covers a set of 264 medications, whereas in the previous studies we focused on a single phenotype at a time. Although it may be more complex to model multiple (vs. a single) targets, the results enable a more robust comparison between multiple algorithms/models by enabling the use of statistical testing procedures such as paired t-tests.

This thesis also contributes to advancing precision psychiatry [117], an emerging field aiming to improve the diagnosis, prognosis, and treatment of psychiatric disorders by incorporating individual-level data. Specifically, we highlighted several novel genes and molecular functions which could help provide a more comprehensive understanding of the pathophysiology of psychiatric disorders and drive the development of more effective therapeutics (Chapters 2 and 3). Furthermore, we demonstrated the potential of using blood microRNA profiles as biomarkers of MDD severity and treatment response, which could enable a biologically informed approach to disease treatment (Chapter 4). Our findings also suggest that ML approaches can facilitate the prioritization of medications based on an individual's genetic data (Chapter 5), paving the way toward more personalized treatment strategies in psychiatry. In addition to our contributions to the field of psychiatry, our work may have wider implications for other complex diseases beyond psychiatric disorders.

Machine learning challenges

Overall, the four studies comprising this thesis showed that ML approaches offer considerable potential for understanding complex diseases and optimizing their treatment. However, during the course of our research, it was evident that the challenge was not just about applying the correct ML algorithm, but the entire process from ensuring data quality, data

cleaning and preprocessing, to model selection, evaluation, and interpretation, and the correctness of algorithm implementation for each of these multiple stages. Despite careful preprocessing of omics data, there were instances where data quality and preprocessing could have been further improved. For example, a step for detecting poor quality or outlier features could have been implemented, which might have improved the performance of resulting ML models and the robustness of our findings. Furthermore, methods to adjust for class imbalances, which are often seen as a challenge for ML algorithms designed based on the assumption of balanced classes [118], could have been implemented in our analysis methodology (e.g., for the blood mRNA dataset in the MDD cases-control analysis).

Moreover, we also did not offer a more comprehensive interpretation of model performance such as precision-recall, prediction uncertainties, and calibration. We realize that such performance metrics would be crucial if ML classifiers were to be deployed in real-world clinical scenarios for classification and prediction purposes. The effectiveness of a model would need to be evaluated based on specific clinical situations such as patient outcomes and cost savings [119], and studies involving a closer collaboration between multiple expertise groups including clinicians, ethicists, and statisticians would be necessary to enable a proper analysis of the clinical utility of ML tools. However, more research focusing on improving the model performance through improving data quality, model architecture design, and incorporation of prior domain knowledge is still needed in parallel with these efforts.

In addition to performance metrics, other factors such as the interpretability and fairness of the models are also important. Interpretability is crucial for the clinical translation of our work, as it allows clinicians to understand the reasoning behind a model's predictions. Models based on a foundation of prior domain knowledge (as shown in Chapter 5 with the GCN model),

combined with feature attribution methods of neural networks such as integrated gradients [120], could offer a path towards better interpretability of ML models. Moreover, fairness ensures that the model does not discriminate against specific patient groups, which is essential for ethical considerations in healthcare. Although we did not specifically focus on addressing potential biases in our ML models, it is important to note there are methods and tools designed specifically for such purposes [121], and that such analyses would be crucial in the context of clinical prediction models in general.

In terms of ML applications, this thesis focused mainly on methods of binary or multi-class classification, which simplifies model evaluation and interpretability. However, it may be the case that regression modelling offers more insights in terms of tasks such as disease severity and treatment response prediction. By employing regression models, we can potentially uncover more nuanced relationships between variables and better understand the factors influencing disease severity and treatment response.

Furthermore, it may be difficult to directly learn a model for a complex task such as treatment response prediction eight weeks in advance due to a very low signal-to-noise ratio. This issue could arise when the relevant signal related to treatment response is obscured by a large amount of irrelevant biological and environmental influences, making it difficult for the model to discern meaningful patterns. A strategy for lowering the complexity of the task through incorporating multiple time points may be useful to improve the learning of temporal dynamics for treatment response prediction.

Data quality challenges

We encountered several further challenges related to data quality during the course of research. First, the sample sizes used in the studies, particularly for gene expression data, were small relative to the number of features in the dataset. Although we have implemented several strategies to avoid overfitting, it should be noted that another option would be to focus on strategies for creating larger gene expression datasets. Large data repositories of gene expression containing millions of samples from an aggregate of different studies are available [122], and it recently has been shown that it may be effective to combine the data with proper data normalization and batch correction for ML analysis [123].

Furthermore, when assessing treatment response using genetic data, a prospective study with better characterization of treatment response would be necessary (e.g., taking advantage of the N-of-1 study design to ensure objective treatment response classification, where possible).

Finally, the generalizability of the ML models developed in this thesis may be limited by the specific populations and data sources used. For instance, the datasets used consisted of mainly individuals with European ancestry. As a result, the performance of these models may not hold up when applied to diverse populations with different genetic backgrounds, environmental exposures, and clinical characteristics. To address this limitation, future research should focus on validating the models using independent datasets collected from various populations and demographic groups, and incorporating methods for model fairness analysis mentioned previously. This approach would help improve the robustness and generalizability of the patterns learned through ML.

Confounding and causality

Another challenge in ML analysis we encountered is due to the indirect effects of covariates, which in our context refer to explanatory variables which are not the primary interest in an investigation [124]. These covariates can confound the relationship between the variables of interest and the outcome, making it challenging to determine the true underlying causative relationships relevant to disease. For example, in the case that smoking is associated with MDD status, the effects of smoking may be reflected in the gene expression patterns of patients. In this scenario, it would be difficult to disentangle whether the gene expression differences are due to smoking, or due to disease-relevant mechanisms. There are various strategies to account for the potential confounding effects of covariates. As we have shown in Chapter 3, given that we have the smoking status of each individual in the dataset, we can perform a stratification of smoking status to examine how the model prediction performance differs in the smoker vs. non-smoker subgroups to understand the effects of smoking on model predictions of disease status. This stratification allows us to isolate the impact of smoking and better understand its role in the relationship between gene expression patterns and MDD status. Future research leveraging ML could leverage similar strategies to better evaluate the effects of covariates on model performance. Another option to address this challenge could be to leverage causal supervised learning which includes techniques specifically designed for learning causal relationships using ML [125].

Chapter 7. Conclusion and Future Directions

In conclusion, this thesis has demonstrated the potential of ML approaches to advance our understanding of the molecular basis of complex diseases and optimize their treatment. Our findings show the effectiveness of ML models in classifying psychiatric disorders using gene expression data and highlight the potential of blood microRNA profiles and genetic data as biomarkers for monitoring clinical evolution and treatment response in patients. In particular, we have illustrated the effectiveness of a graph-based ML approach for prioritizing medications based on individual genetic data. These findings together support the need to further develop ML approaches for omics data analysis that take into consideration the combined effects of multiple features for understanding of complex diseases.

Moving forward, it is important to conduct further validation of the genes we have identified in this study. Additional research could focus on testing these genes in independent cohorts and different demographic groups, and exploring their potential functional roles in disease pathophysiology. Furthermore, certain limitations and challenges also need to be addressed in future research. These include increasing sample sizes, improving the generalizability of ML models in individuals of non-European ancestry, and addressing the indirect effects of covariates on ML analysis.

Another promising direction to explore is integrating comprehensive phenotyping data into ML analysis. The use of a single target label (e.g., disease status) simplifies a ML task, however, this approach does not fully capture the details of an individual's unique disease state. The application of ML and more specifically deep learning and GRL, which have the flexibility of modelling multiple target variables (i.e., multi-label, multi-output models) [126, 127], may be able to more accurately characterize an individual's disease state and allow for more insights

with clinical relevance and impact to be learned about complex diseases. A promising way to achieve this would be to leverage the use of population-scale biobanks which provide comprehensive genotyping and phenotyping data of participants [128].

While we analyzed several different omics data types including gene expressions, microRNA expressions, and genetic variants, we did not explore the integration of multiple feature types (i.e., multi-omics integration [129]). The graph approach we introduced in Chapter 5 could be a promising approach for multi-omics integration. The current graph-based approach using genetic variants as inputs could be naturally extended to integrate multiple feature types while being able to leverage the biological connections between the features through biological processes and pathways.

In addition to integrating multiple biological omics data types, incorporating data from other modalities such as biomedical imaging in the ML analysis of complex disorders may compensate for limitations inherent to omics modalities and contribute to a more comprehensive view of complex disease etiology and treatment. For example, a recent study analyzed whole-brain structural magnetic resonance imaging integrated with cerebrospinal fluid metabolomics to associate brain regions with metabolic disruptions [130]. Furthermore, given that a large amount of descriptions of a patient's disease are in the form of clinical documentation, which consists of unstructured text information, an appropriate means of integrating important information should be integrated into complex disease research [131].

Another important future direction is to explore the use of temporal data for ML modelling for understanding of disease development, progression, and treatment responses. In our studies, we have not focused on the analysis of temporal data, which could be a potential limitation. For example, in our analysis of microRNA for antidepressant response (Chapter 4),

incorporating temporal data in the early stages of treatment may be useful in capturing early changes in microRNA levels and offer more robust signals of antidepressant response.

Lastly, identifying potential clinical applications based on research findings is crucial for advancing the treatment of complex diseases. However, the implementation of these applications will require careful consideration of ethical and economic challenges, such as data collection and privacy, minimizing potential harmful biases, and evaluating the cost-effectiveness of clinical applications [132-134]. Ultimately, the successful translation of these research findings into clinical applications will pave the way for a new era in precision medicine, revolutionizing the diagnosis, prognosis, and treatment of complex diseases and improving patient outcomes.

Chapter 8. General References

1. Mitchell, K.J., *What is complex about complex disorders?* Genome Biol, 2012. **13**(1): p. 237.
2. Manolio, T.A., *Genomewide association studies and assessment of the risk of disease.* N Engl J Med, 2010. **363**(2): p. 166-76.
3. Amberger, J., et al., *McKusick's online Mendelian inheritance in man (OMIM®).* Nucleic acids research, 2009. **37**(suppl_1): p. D793-D796.
4. Sollis, E., et al., *The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource.* Nucleic Acids Research, 2023. **51**(D1): p. D977-D985.
5. Wang, G., et al., *Additive, Epistatic, and Environmental Effects Through the Lens of Expression Variability QTL in a Twin Cohort.* Genetics, 2014. **196**(2): p. 413-425.
6. Barabási, A.L., N. Gulbahce, and J. Loscalzo, *Network medicine: a network-based approach to human disease.* Nat Rev Genet, 2011. **12**(1): p. 56-68.
7. Furman, D., et al., *Chronic inflammation in the etiology of disease across the life span.* Nature Medicine, 2019. **25**(12): p. 1822-1832.
8. Hasin, Y., M. Seldin, and A. Lusis, *Multi-omics approaches to disease.* Genome Biology, 2017. **18**(1): p. 83.
9. Lewis, C.M. and E. Vassos, *Polygenic risk scores: from research tools to clinical instruments.* Genome medicine, 2020. **12**(1): p. 1-11.
10. Cookson, W., et al., *Mapping complex disease traits with global gene expression.* Nature Reviews Genetics, 2009. **10**(3): p. 184-194.

11. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Research, 2015. **43**(7): p. e47-e47.

12. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**(1): p. 559.

13. Dai, X. and L. Shen, *Advances and Trends in Omics Technology Development.* Frontiers in Medicine, 2022. **9**.

14. Collins, F.S. and H. Varmus, *A new initiative on precision medicine.* New England journal of medicine, 2015. **372**(9): p. 793-795.

15. Relling, M.V. and W.E. Evans, *Pharmacogenomics in the clinic.* Nature, 2015. **526**(7573): p. 343-350.

16. Ciardiello, F., et al., *Delivering precision medicine in oncology today and in future-the promise and challenges of personalised cancer medicine: a position paper by the European Society for Medical Oncology (ESMO).* Ann Oncol, 2014. **25**(9): p. 1673-1678.

17. Torres, C. and P.J. Grippo, *Pancreatic cancer subtypes: a roadmap for precision medicine.* Annals of Medicine, 2018. **50**(4): p. 277-287.

18. Tsakiroglou, M., A. Evans, and M. Pirmohamed, *Leveraging transcriptomics for precision diagnosis: Lessons learned from cancer and sepsis.* Front Genet, 2023. **14**: p. 1100352.

19. Phulka, J.S., et al., *Current State and Future of Polygenic Risk Scores in Cardiometabolic Disease: A Scoping Review.* Circulation: Genomic and Precision Medicine, 2023. **16**(3): p. 286-313.

20. Johansson, Å., et al., *Precision medicine in complex diseases—Molecular subgrouping for improved prediction and treatment stratification.* Journal of Internal Medicine, 2023. **n/a**(n/a).

21. Olivier, M., et al., *The Need for Multi-Omics Biomarker Signatures in Precision Medicine.* Int J Mol Sci, 2019. **20**(19).

22. Wiedmeier, J.E., et al., *Single-Cell Sequencing in Precision Medicine.* Cancer Treat Res, 2019. **178**: p. 237-252.

23. Ota, M. and K. Fujio, *Multi-omics approach to precision medicine for immune-mediated diseases.* Inflammation and Regeneration, 2021. **41**(1): p. 23.

24. Saura, C.A., et al., *Revealing cell vulnerability in Alzheimer's disease by single-cell transcriptomics.* Semin Cell Dev Biol, 2023. **139**: p. 73-83.

25. MacEachern, S.J. and N.D. Forkert, *Machine learning for precision medicine.* Genome, 2021. **64**(4): p. 416-425.

26. Moreno-Küstner, B., C. Martín, and L. Pastor, *Prevalence of psychotic disorders and its association with methodological issues. A systematic review and meta-analyses.* PLoS One, 2018. **13**(4): p. e0195687.

27. Hasin, D.S., et al., *Epidemiology of Adult DSM-5 Major Depressive Disorder and Its Specifiers in the United States.* JAMA Psychiatry, 2018. **75**(4): p. 336-346.

28. Sullivan, P.F., K.S. Kendler, and M.C. Neale, *Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies.* Arch Gen Psychiatry, 2003. **60**(12): p. 1187-92.

29. Sullivan, P.F., M.C. Neale, and K.S. Kendler, *Genetic epidemiology of major depression: review and meta-analysis.* Am J Psychiatry, 2000. **157**(10): p. 1552-62.

30. Morgan, C. and H. Fisher, *Environment and schizophrenia: environmental factors in schizophrenia: childhood trauma--a critical review.* Schizophr Bull, 2007. **33**(1): p. 3-10.

31. Mandelli, L., C. Petrelli, and A. Serretti, *The role of specific early trauma in adult depression: A meta-analysis of published literature. Childhood trauma and adult depression.* European psychiatry, 2015. **30**(6): p. 665-680.

32. Wahbeh, M.H. and D. Avramopoulos, *Gene-environment interactions in schizophrenia: a literature review.* Genes, 2021. **12**(12): p. 1850.

33. Klengel, T. and E.B. Binder, *Gene—environment interactions in major depressive disorder.* The Canadian Journal of Psychiatry, 2013. **58**(2): p. 76-83.

34. Patel, K.R., et al., *Schizophrenia: overview and treatment options.* P t, 2014. **39**(9): p. 638-45.

35. Zhdanava, M., et al., *The Prevalence and National Burden of Treatment-Resistant Depression and Major Depressive Disorder in the United States.* J Clin Psychiatry, 2021. **82**(2).

36. Fernandes, B.S., et al., *The new field of 'precision psychiatry'.* BMC Medicine, 2017. **15**(1): p. 80.

37. American Psychiatric, A., *Diagnostic and statistical manual of mental disorders (DSM-5®)*. 2013: American Psychiatric Pub.

38. Pardiñas, A.F., et al., *Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection.* Nature genetics, 2018. **50**(3): p. 381-389.

39. Ormel, J., C.A. Hartman, and H. Snieder, *The genetics of depression: successful genome-wide association studies introduce new challenges.* Translational Psychiatry, 2019. **9**(1): p. 114.

40. Murray, G.K., et al., *Could Polygenic Risk Scores Be Useful in Psychiatry?: A Review.* JAMA Psychiatry, 2021. **78**(2): p. 210-219.

41. Bracher-Smith, M., K. Crawford, and V. Escott-Price, *Machine learning for genetic prediction of psychiatric disorders: a systematic review.* Mol Psychiatry, 2021. **26**(1): p. 70-79.

42. Gusev, A., et al., *Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights.* Nature Genetics, 2018. **50**(4): p. 538-548.

43. Sanders, A.R., et al., *Transcriptome study of differential expression in schizophrenia.* Hum Mol Genet, 2013. **22**(24): p. 5001-14.

44. Etemadikhah, M., et al., *Transcriptome analysis of fibroblasts from schizophrenia patients reveals differential expression of schizophrenia-related genes.* Scientific Reports, 2020. **10**(1): p. 630.

45. Comes, A.L., et al., *Proteomics for blood biomarker exploration of severe mental illness: pitfalls of the past and potential for the future.* Transl Psychiatry, 2018. **8**(1): p. 160.

46. Shih, P.B., *Metabolomics Biomarkers for Precision Psychiatry.* Adv Exp Med Biol, 2019. **1161**: p. 101-113.

47. Thorn, C.F., T.E. Klein, and R.B. Altman, *PharmGKB: the Pharmacogenomics Knowledge Base.* Methods Mol Biol, 2013. **1015**: p. 311-20.

48. Iyengar, R., *Complex diseases require complex therapies.* EMBO Rep, 2013. **14**(12): p. 1039-42.

49. Roden, D.M., et al., *Pharmacogenomics: the genetics of variable drug responses.* Circulation, 2011. **123**(15): p. 1661-70.

50. Caldwell, J., I. Gardner, and N. Swales, *An introduction to drug disposition: the basic principles of absorption, distribution, metabolism, and excretion.* Toxicol Pathol, 1995. **23**(2): p. 102-14.

51. Marino, M., Z. Jamal, and P.M. Zito, *Pharmacodynamics.* 2018.

52. Bzdok, D. and A. Meyer-Lindenberg, *Machine Learning for Precision Psychiatry: Opportunities and Challenges.* Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 2018. **3**(3): p. 223-230.

53. Quinlan, J.R., *Induction of decision trees.* Machine learning, 1986. **1**: p. 81-106.

54. Noble, W.S., *What is a support vector machine?* Nature biotechnology, 2006. **24**(12): p. 1565-1567.

55. Friedman, J.H., *Greedy function approximation: a gradient boosting machine.* Annals of statistics, 2001: p. 1189-1232.

56. Hartigan, J.A. and M.A. Wong, *Algorithm AS 136: A k-means clustering algorithm.* Journal of the royal statistical society. series c (applied statistics), 1979. **28**(1): p. 100-108.

57. McInnes, L., J. Healy, and J. Melville, *Umap: Uniform manifold approximation and projection for dimension reduction.* arXiv preprint arXiv:1802.03426, 2018.

58. Reel, P.S., et al., *Using machine learning approaches for multi-omics data analysis: A review.* Biotechnology Advances, 2021. **49**: p. 107739.

59. Hamilton, W.L., R. Ying, and J. Leskovec, *Representation learning on graphs: Methods and applications.* arXiv preprint arXiv:1709.05584, 2017.

60. Szklarczyk, D., et al., *STRING v10: protein–protein interaction networks, integrated over the tree of life.* Nucleic acids research, 2015. **43**(D1): p. D447-D452.

61. Kipf, T.N. and M. Welling, *Semi-supervised classification with graph convolutional networks.* arXiv preprint arXiv:1609.02907, 2016.

62. Yang, Q., et al., *Classification for psychiatric disorders including schizophrenia, bipolar disorder, and major depressive disorder using machine learning.* Comput Struct Biotechnol J, 2022. **20**: p. 5054-5064.

63. Yang, Q., et al., *A novel multi-class classification model for schizophrenia, bipolar disorder and healthy controls using comprehensive transcriptomic data.* Comput Biol Med, 2022. **148**: p. 105956.

64. Feng, Y. and J. Shen, *Machine learning-based predictive models and drug prediction for schizophrenia in multiple programmed cell death patterns.* Front Mol Neurosci, 2023. **16**: p. 1123708.

65. Liu, S., et al., *A machine learning model for predicting patients with major depressive disorder: A study based on transcriptomic data.* Front Neurosci, 2022. **16**: p. 949609.

66. Zhao, S., et al., *Identification of Diagnostic Markers for Major Depressive Disorder Using Machine Learning Methods.* Front Neurosci, 2021. **15**: p. 645998.

67. Yi, Z., et al., *Blood-based gene expression profiles models for classification of subsyndromal symptomatic depression and major depressive disorder.* PLoS One, 2012. **7**(2): p. e31283.

68. Liu, Y., et al., *Machine Learning Reduced Gene/Non-Coding RNA Features That Classify Schizophrenia Patients Accurately and Highlight Insightful Gene Clusters.* Int J Mol Sci, 2021. **22**(7).

69. Kittel-Schneider, S., et al., *Proteomic Profiling as a Diagnostic Biomarker for Discriminating Between Bipolar and Unipolar Depression.* Front Psychiatry, 2020. **11**: p. 189.

70. Habets, P.C., et al., *Multimodal Data Integration Advances Longitudinal Prediction of the Naturalistic Course of Depression and Reveals a Multimodal Signature of Remission During Two Year Follow-Up.* Biol Psychiatry, 2023.

71. Mongan, D., et al., *Development of Proteomic Prediction Models for Transition to Psychotic Disorder in the Clinical High-Risk State and Psychotic Experiences in Adolescence.* JAMA Psychiatry, 2021. **78**(1): p. 77-90.

72. Zheng, H., et al., *Predictive diagnosis of major depression using NMR-based metabolomics and least-squares support vector machine.* Clin Chim Acta, 2017. **464**: p. 223-227.

73. Liu, Y., et al., *Metabolomic biosignature differentiates melancholic depressive patients from healthy controls.* BMC Genomics, 2016. **17**(1): p. 669.

74. Takahashi, Y., et al., *Improved metabolomic data-based prediction of depressive symptoms using nonlinear machine learning with feature selection.* Transl Psychiatry, 2020. **10**(1): p. 157.

75. Tasic, L., et al., *Peripheral biomarkers allow differential diagnosis between schizophrenia and bipolar disorder.* Journal of Psychiatric Research, 2019. **119**: p. 67-75.

76. Bhak, Y., et al., *Depression and suicide risk prediction models using blood-derived multi-omics data.* Transl Psychiatry, 2019. **9**(1): p. 262.

77. Joyce, J.B., et al., *Multi-omics driven predictions of response to acute phase combination antidepressant therapy: a machine learning approach with cross-trial replication.* Transl Psychiatry, 2021. **11**(1): p. 513.

78. Guo, L.K., et al., *Prediction of treatment response to antipsychotic drugs for precision medicine approach to schizophrenia: randomized trials and multiomics analysis.* Mil Med Res, 2023. **10**(1): p. 24.

79. Hasanzad, M., et al., *Precision medicine journey through omics approach.* J Diabetes Metab Disord, 2022. **21**(1): p. 881-888.

80. O'Brien, J., et al., *Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation.* Front Endocrinol (Lausanne), 2018. **9**: p. 402.

81. Al-Amrani, S., et al., *Proteomics: Concepts and applications in human medicine.* World J Biol Chem, 2021. **12**(5): p. 57-69.

82. Doerr, A., *Global metabolomics.* Nature Methods, 2017. **14**(1): p. 32-32.

83. Stears, R.L., T. Martinsky, and M. Schena, *Trends in microarray analysis.* Nature Medicine, 2003. **9**(1): p. 140-145.

84. Goodwin, S., J.D. McPherson, and W.R. McCombie, *Coming of age: ten years of next-generation sequencing technologies.* Nature Reviews Genetics, 2016. **17**(6): p. 333-351.

85. Evans, C., J. Hardin, and D.M. Stoebel, *Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions.* Brief Bioinform, 2018. **19**(5): p. 776-792.

86. Tryka, K.A., et al., *NCBI's Database of Genotypes and Phenotypes: dbGaP.* Nucleic Acids Research, 2014. **42**(D1): p. D975-D979.

87. Clough, E. and T. Barrett, *The Gene Expression Omnibus Database.* Methods Mol Biol, 2016. **1418**: p. 93-110.

88. Sudlow, C., et al., *UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age.* PLoS Med, 2015. **12**(3): p. e1001779.

89. Liberzon, A., et al., *The Molecular Signatures Database (MSigDB) hallmark gene set collection.* Cell Syst, 2015. **1**(6): p. 417-425.

90. Barbarino, J.M., et al., *PharmGKB: A worldwide resource for pharmacogenomic information.* Wiley Interdiscip Rev Syst Biol Med, 2018. **10**(4): p. e1417.

91. Armstrong, R.A., *When to use the Bonferroni correction.* Ophthalmic Physiol Opt, 2014. **34**(5): p. 502-8.

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

93. Hinton, G.E., *Distributed representations.* 1984.

94. Adhao, R. and V. Pachghare, *Feature selection using principal component analysis and genetic algorithm.* Journal of Discrete Mathematical Sciences and Cryptography, 2020. **23**(2): p. 595-602.

95. Anuradha, T., et al. *Feature Extraction and Representation Learning via Deep Neural Network*. in *Computer Networks, Big Data and IoT*. 2022. Singapore: Springer Nature Singapore.

96. Russell, S.J., *Artificial intelligence a modern approach*. 2010: Pearson Education, Inc.

97. Rokach, L. and O. Maimon, *Decision trees.* Data mining and knowledge discovery handbook, 2005: p. 165-192.

98. Akiba, T., et al. *Optuna: A next-generation hyperparameter optimization framework*.

99. Karmaker, S.K., et al., *Automl to date and beyond: Challenges and opportunities.* ACM Computing Surveys (CSUR), 2021. **54**(8): p. 1-36.

100. Chen, T. and C. Guestrin. *Xgboost: A scalable tree boosting system*.

101. O'Shea, K. and R. Nash, *An introduction to convolutional neural networks.* arXiv preprint arXiv:1511.08458, 2015.

102. Yu, Y., et al., *A review of recurrent neural networks: LSTM cells and network architectures.* Neural computation, 2019. **31**(7): p. 1235-1270.

103. Zhang, S., et al., *Graph convolutional networks: a comprehensive review.* Computational Social Networks, 2019. **6**(1): p. 11.

104. Fensel, D., et al., *Introduction: what is a knowledge graph?* Knowledge graphs: Methodology, tools and selected use cases, 2020: p. 1-10.

105. Ji, S., et al., *A survey on knowledge graphs: Representation, acquisition, and applications.* IEEE transactions on neural networks and learning systems, 2021. **33**(2): p. 494-514.

106. Nicholson, D.N. and C.S. Greene, *Constructing knowledge graphs and their biomedical applications.* Computational and Structural Biotechnology Journal, 2020. **18**: p. 1414-1428.

107. Hänsel, K., et al., *From Data to Wisdom: Biomedical Knowledge Graphs for Real-World Data Insights.* J Med Syst, 2023. **47**(1): p. 65.

108. Bartel, D.P., *MicroRNAs: genomics, biogenesis, mechanism, and function.* Cell, 2004. **116**(2): p. 281-97.

109. Ardekani, A.M. and M.M. Naeini, *The Role of MicroRNAs in Human Diseases.* Avicenna J Med Biotechnol, 2010. **2**(4): p. 161-79.

110. Dwivedi, Y., *Emerging role of microRNAs in major depressive disorder: diagnosis and therapeutic implications.* Dialogues Clin Neurosci, 2014. **16**(1): p. 43-61.

111. Karlsgodt, K.H., D. Sun, and T.D. Cannon, *Structural and Functional Brain Abnormalities in Schizophrenia.* Curr Dir Psychol Sci, 2010. **19**(4): p. 226-231.

112. Severance, E.G., et al., *Gastroenterology issues in schizophrenia: why the gut matters.* Curr Psychiatry Rep, 2015. **17**(5): p. 27.

113. Li, R., et al., *Machine learning meets omics: applications and perspectives.* Briefings in Bioinformatics, 2022. **23**(1): p. bbab460.

114. Taliaz, D., et al., *Optimizing prediction of response to antidepressant medications using machine learning and integrated genetic, clinical, and demographic data.* Translational Psychiatry, 2021. **11**(1): p. 381.

115. Adam, G., et al., *Machine learning approaches to drug response prediction: challenges and recent progress.* npj Precision Oncology, 2020. **4**(1): p. 19.

116. Hosseini, M.M., et al., *Leveraging machine learning and big data for optimizing medication prescriptions in complex diseases: a case study in diabetes management.* Journal of Big Data, 2020. **7**(1): p. 26.

117. Fernandes, B.S., et al., *The new field of 'precision psychiatry'.* BMC Med, 2017. **15**(1): p. 80.

118. Japkowicz, N. and S. Stephen, *The class imbalance problem: A systematic study.* Intelligent data analysis, 2002. **6**(5): p. 429-449.

119. Mišić, V.V., K. Rajaram, and E. Gabel, *A simulation-based evaluation of machine learning models for clinical decision support: application and analysis using hospital readmission.* npj Digital Medicine, 2021. **4**(1): p. 98.

120. Sundararajan, M., A. Taly, and Q. Yan. *Axiomatic attribution for deep networks*. PMLR.

121. Adebayo, J.A., *FairML: ToolBox for diagnosing bias in predictive modeling.* 2016.

122. Edgar, R., M. Domrachev, and A.E. Lash, *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.* Nucleic acids research, 2002. **30**(1): p. 207-210.

123. Foltz, S.M., C.S. Greene, and J.N. Taroni, *Cross-platform normalization enables machine learning model training on microarray and RNA-seq data simultaneously.* Communications Biology, 2023. **6**(1): p. 222.

124. Fan, S., *Encyclopedia of Research Design*. 2010, SAGE Publications, Inc.: Thousand Oaks

Thousand Oaks, California.

125. Kaddour, J., et al., *Causal machine learning: A survey and open problems.* arXiv preprint arXiv:2206.15475, 2022.

126. Liu, J., et al. *Deep learning for extreme multi-label text classification*.

127. Read, J. and F. Perez-Cruz, *Deep learning for multi-label classification.* arXiv preprint arXiv:1502.05988, 2014.

128. Malsagova, K., et al., *Biobanks—a platform for scientific and biomedical research.* Diagnostics, 2020. **10**(7): p. 485.

129. Subramanian, I., et al., *Multi-omics Data Integration, Interpretation, and Its Application.* Bioinform Biol Insights, 2020. **14**: p. 1177932219899051.

130. Eldridge, R.C., et al., *Multiomics analysis of structural magnetic resonance imaging of the brain and cerebrospinal fluid metabolomics in cognitively normal and impaired adults.* Frontiers in Aging Neuroscience, 2022. **13**: p. 997.

131. Sheikhalishahi, S., et al., *Natural language processing of clinical notes on chronic diseases: systematic review.* JMIR medical informatics, 2019. **7**(2): p. e12239.

132. Prosperi, M., et al., *Big data hurdles in precision medicine and precision public health.* BMC Med Inform Decis Mak, 2018. **18**(1): p. 139.

133. Fröhlich, H., et al., *From hype to reality: data science enabling personalized medicine.* BMC Med, 2018. **16**(1): p. 150.

134. Gavan, S.P., A.J. Thompson, and K. Payne, *The economic case for precision medicine.* Expert Rev Precis Med Drug Dev, 2018. **3**(1): p. 1-9.

Appendix

Chapter 2 top brain mRNA model hyperparameter search results:

- Mean validation score: 0.762 (std: 0.050)
- AUC per fold: [0.8016304347826086, 0.7771739130434783, 0.7989130434782609, 0.7159090909090909, 0.7897727272727273, 0.8636363636363636, 0.6875, 0.7272727272727273, 0.7215909090909092, 0.7357954545454546]
- Optimal threshold for classification: 0.443379976263868
- Parameters: {'gamma': 0.01768780961325722, 'learning_rate': 0.07025531638428649, 'max_depth': 3, 'n_estimators': 35}

Chapter 3 top brain mRNA model hyperparameters search results:

- Mean validation score: 0.715 (std: 0.100)
- AUC per fold: [0.8181818181818181, 0.818181818181818, 0.7159090909090909, 0.7000000000000001, 0.8, 0.7375, 0.5375, 0.775, 0.725, 0.525]
- Optimal threshold for classification: 0.4973402112722397
- Parameters: {'gamma': 0.5151736824068126, 'learning_rate': 0.10755413156827563, 'max_depth': 1, 'n_estimators': 122, 'reg_alpha': 1}

Chapter 3 top covariates only model hyperparameters search results:

- Mean validation score: 0.834 (std: 0.075)
- AUC per fold: [0.9431818181818181, 0.6534090909090908, 0.8636363636363636, 0.85, 0.8562500000000001, 0.9, 0.79375, 0.875, 0.8187500000000001, 0.7875]
- Optimal threshold for classification: 0.5238638237118721
- Parameters: {'gamma': 0.35188483434700646, 'learning_rate': 0.0805615304784293, 'max_depth': 5, 'n_estimators': 140, 'reg_alpha': 0}

Chapter 3 top brain mRNA + covariates model hyperparameters search results:

- Mean validation score: 0.705 (std: 0.087)
- AUC per fold: [0.7613636363636364, 0.7045454545454546, 0.75, 0.75, 0.7875000000000001, 0.7625000000000001, 0.5375000000000001, 0.75, 0.7125, 0.5375]
- Optimal threshold for classification: 0.5293224930763245
- Parameters: {'gamma': 0.7486721752509908, 'learning_rate': 0.09279249545453125, 'max_depth': 1, 'n_estimators': 87, 'reg_alpha': 1}

Chapter 3 top blood mRNA model hyperparameter search results:

- Mean validation score: 0.640 (std: 0.041)
- AUC per fold: [0.5859353787673093, 0.6073852837360847, 0.6334509910399131, 0.6663046429541135, 0.633722508824328, 0.7225088243279936, 0.624490904154222, 0.6023809523809524, 0.6962962962962964, 0.6304232804232804]
- Optimal threshold for classification: 0.9334825277328491
- Parameters: {'gamma': 0.3866201726338342, 'learning_rate': 0.07624839469416081, 'max_depth': 4, 'n_estimators': 138, 'reg_alpha': 0}

Chapter 4 top case-control classification model hyperparameter search results:

- Mean validation score: 0.929 (std: 0.063)

- AUC per fold: [1.0, 0.95, 0.9624999999999999, 0.8157894736842105, 0.912280701754386]
- Optimal threshold for classification: 0.7485573918391497
- Parameters: {'gamma': 0.05816379550547579, 'learning_rate': 0.10948663171697672, 'max_depth': 1, 'n_estimators': 163}

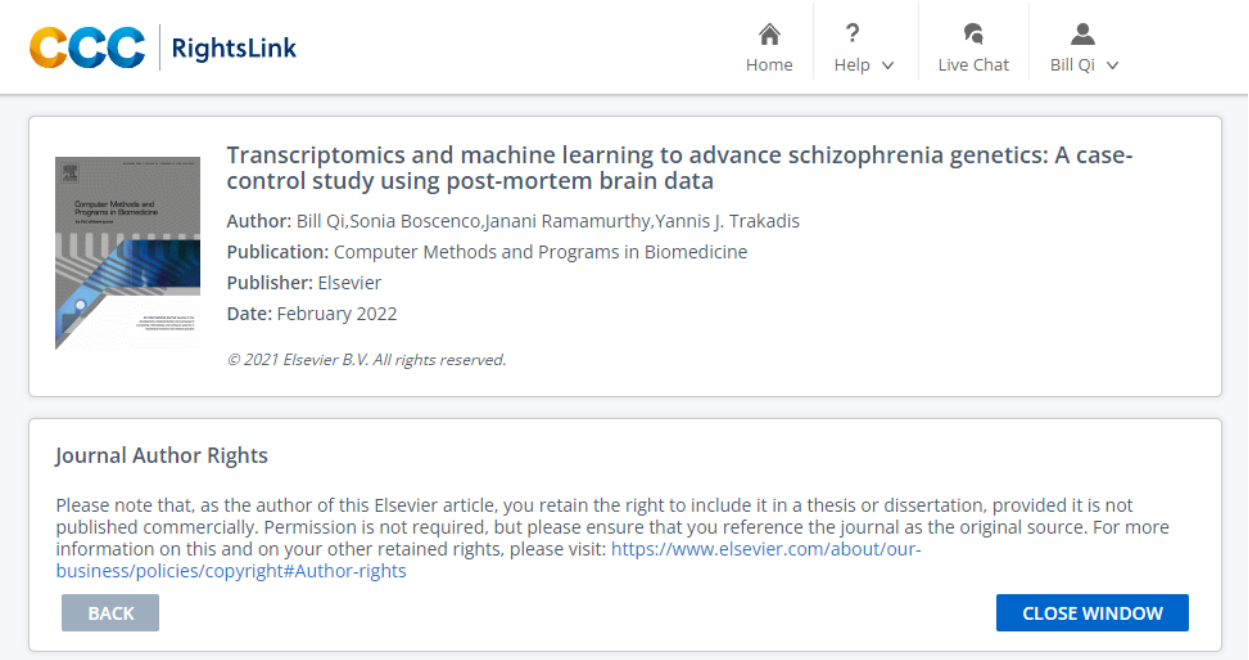
Chapter 4 top disease severity model hyperparameter search results:

- Mean validation score: 0.756 (std: 0.132)
- AUC per fold: [0.8402777777777778, 0.5, 0.7666666666666666, 0.7916666666666666, 0.875]
- Optimal threshold for classification: 0.4780568954272148
- Parameters: {'gamma': 0.14551548278735915, 'learning_rate': 0.09771275618985018, 'max_depth': 2, 'n_estimators': 145}

Chapter 4 top treatment response model hyperparameter search results:

- Mean validation score: 0.622 (std: 0.127)
- AUC per fold: [0.744047619047619, 0.42948717948717946, 0.6217948717948717, 0.5384615384615384, 0.7692307692307693]
- Optimal threshold for classification: 0.5066269127031168
- Parameters: {'gamma': 0.3772859652420726, 'learning_rate': 0.010259502438364656, 'max_depth': 4, 'n_estimators': 5}

Chapter 2 copyright permissions:



Chapter 3 copyright permissions:

JOHN WILEY AND SONS LICENSE
TERMS AND CONDITIONS
Apr 14, 2023

This Agreement between Mr. Bill Qi ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	5527840955343
License date	Apr 14, 2023
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	American Journal of Medical Genetics Part A
Licensed Content Title	Machine learning and bioinformatic analysis of brain and blood mRNA profiles in major depressive disorder: A case–control study
Licensed Content Author	Bill Qi, Janani Ramamurthy, Imane Bennani, et al
Licensed Content Date	Mar 1, 2021
Licensed Content Volume	186
Licensed Content Issue	2
Licensed Content Pages	12
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Print and electronic
Portion	Full article
Will you be	No

translating?

Title	Advancing the Understanding and Treatment of Complex Diseases through Machine Learning and Omics: Insights from Analyses of Psychiatric Disorders and Medications Data
Institution name	McGill University
Expected presentation date	Oct 2023
Order reference number	3
Requestor Location	Mr. Bill Qi 2155 Prud'homme Avenue Apartment 21 Montreal, QC H4A3H3 Canada Attn: Mr. Bill Qi
Publisher Tax ID	EU826007151
Total	0.00 CAD

Terms and Conditions

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a"Wiley Company") or handled on behalf of a society with which a Wiley Company has exclusive publishing rights in relation to a particular work (collectively "WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your RightsLink account (these are available at any time at http://myaccount.copyright.com).

Terms and Conditions

- The materials you have requested permission to reproduce or reuse (the "Wiley Materials") are protected by copyright.

- You are hereby granted a personal, non-exclusive, non-sub licensable (on a stand-alone basis), non-transferable, worldwide, limited license to reproduce the Wiley Materials for the purpose specified in the licensing process. This license, **and any CONTENT (PDF or image file) purchased as part of your order,** is for a one-time use only and limited to any maximum distribution number specified in the license. The first instance of republication or reuse granted by this license must be completed within two years of the date of the grant of this license (although copies prepared before the end date may be distributed thereafter). The Wiley Materials shall not be used in any other manner or for any other purpose, beyond what is granted in the license. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Wiley Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Wiley Material. Any third party content is expressly excluded from this permission.

- With respect to the Wiley Materials, all rights are reserved. Except as expressly granted by the terms of the license, no part of the Wiley Materials may be copied, modified, adapted (except for minor reformatting required by the new Publication), translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Wiley Materials without the prior permission of the respective copyright owner.**For STM Signatory Publishers clearing permission under the terms of the STM Permissions Guidelines only, the terms of the license are extended to include subsequent editions and for editions in other languages, provided such editions are for the work as a whole in situ and does not involve the separate exploitation of the permitted figures or extracts,** You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Wiley Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Wiley Materials on a stand-alone basis, or any of the rights granted to you hereunder to any other person.

- The Wiley Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc,

the Wiley Companies, or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Wiley Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Wiley Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto

- NEITHER WILEY NOR ITS LICENSORS MAKES ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND ITS LICENSORS AND WAIVED BY YOU.

- WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.

- You shall indemnify, defend and hold harmless WILEY, its Licensors and their respective directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.

- IN NO EVENT SHALL WILEY OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT

THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

- Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.

- The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.

- This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.

- Any fee required for this permission shall be non-refundable after thirty (30) days from receipt by the CCC.

- These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.

- In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.

- WILEY expressly reserves all rights not specifically granted in the

combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

- This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.

- This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

WILEY OPEN ACCESS TERMS AND CONDITIONS

Wiley Publishes Open Access Articles in fully Open Access Journals and in Subscription journals offering Online Open. Although most of the fully Open Access journals publish open access articles under the terms of the Creative Commons Attribution (CC BY) License only, the subscription journals and a few of the Open Access Journals offer a choice of Creative Commons Licenses. The license type is clearly identified on the article.

The Creative Commons Attribution License

The Creative Commons Attribution License (CC-BY) allows users to copy, distribute and transmit an article, adapt the article and make commercial use of the article. The CC-BY license permits commercial and non-

Creative Commons Attribution Non-Commercial License

The Creative Commons Attribution Non-Commercial (CC-BY-NC)License permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.(see below)

Creative Commons Attribution-Non-Commercial-NoDerivs License

The Creative Commons Attribution Non-Commercial-NoDerivs License (CC-BY-NC-ND) permits use, distribution and reproduction in any medium, provided the original work is properly cited, is not used for commercial purposes and no modifications or adaptations are made. (see below)

Use by commercial "for-profit" organizations

Use of Wiley Open Access articles for commercial, promotional, or marketing purposes requires further explicit permission from Wiley and will be subject to a fee.

Further details can be found on Wiley Online Library http://olabout.wiley.com/WileyCDA/Section/id-410895.html

Other Terms and Conditions:

v1.10 Last updated September 2015

Questions? customercare@copyright.com.

Chapter 4 copyright permissions:

OXFORD UNIVERSITY PRESS LICENSE
TERMS AND CONDITIONS

Apr 14, 2023

This Agreement between Mr. Bill Qi ("You") and Oxford University Press ("Oxford University Press") consists of your license details and the terms and conditions provided by Oxford University Press and Copyright Clearance Center.

License Number	5527841192818
License date	Apr 14, 2023
Licensed content publisher	Oxford University Press
Licensed content publication	International Journal of Neuropsychopharmacology
Licensed content title	Machine Learning Analysis of Blood microRNA Data in Major Depression: A Case-Control Study for Biomarker Discovery
Licensed content author	Qi, Bill; Fiori, Laura M
Licensed content date	May 4, 2020
Type of Use	Thesis/Dissertation
Institution name	
Title of your work	Advancing the Understanding and Treatment of Complex Diseases through Machine Learning and Omics: Insights from Analyses of Psychiatric Disorders and Medications Data
Publisher of your work	McGill University
Expected publication date	Oct 2023
Permissions cost	0.00 CAD
Value added tax	0.00 CAD
Total	0.00 CAD
Title	Advancing the Understanding and Treatment of Complex Diseases through Machine Learning and Omics: Insights from Analyses of Psychiatric Disorders and Medications Data
Institution name	McGill University

Expected presentation date	Oct 2023
Order reference number	4
Portions	Full Text
Requestor Location	Mr. Bill Qi 2155 Prud'homme Avenue Apartment 21 Montreal, QC H4A3H3 Canada Attn: Mr. Bill Qi
Publisher Tax ID	GB125506730
Total	0.00 CAD

Terms and Conditions

STANDARD TERMS AND CONDITIONS FOR REPRODUCTION OF MATERIAL FROM AN OXFORD UNIVERSITY PRESS JOURNAL

1. Use of the material is restricted to the type of use specified in your order details.

2. This permission covers the use of the material in the English language in the following territory: world. If you have requested additional permission to translate this material, the terms and conditions of this reuse will be set out in clause 12.

3. This permission is limited to the particular use authorized in (1) above and does not allow you to sanction its use elsewhere in any other format other than specified above, nor does it apply to quotations, images, artistic works etc that have been reproduced from other sources which may be part of the material to be used.

4. No alteration, omission or addition is made to the material without our written consent. Permission must be re-cleared with Oxford University Press if/when you decide to reprint.

5. The following credit line appears wherever the material is used: author, title, journal, year, volume, issue number, pagination, by permission of Oxford

University Press or the sponsoring society if the journal is a society journal. Where a journal is being published on behalf of a learned society, the details of that society must be included in the credit line.

6. For the reproduction of a full article from an Oxford University Press journal for whatever purpose, the corresponding author of the material concerned should be informed of the proposed use. Contact details for the corresponding authors of all Oxford University Press journal contact can be found alongside either the abstract or full text of the article concerned, accessible from www.oxfordjournals.org Should there be a problem clearing these rights, please contact journals.permissions@oup.com

7. If the credit line or acknowledgement in our publication indicates that any of the figures, images or photos was reproduced, drawn or modified from an earlier source it will be necessary for you to clear this permission with the original publisher as well. If this permission has not been obtained, please note that this material cannot be included in your publication/photocopies.

8. While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by Oxford University Press or by Copyright Clearance Center (CCC)) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and Oxford University Press reserves the right to take any and all action to protect its copyright in the materials.

9. This license is personal to you and may not be sublicensed, assigned or transferred by you to any other person without Oxford University Press's written permission.

10. Oxford University Press reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's

Billing and Payment terms and conditions.

11. You hereby indemnify and agree to hold harmless Oxford University Press and CCC, and their respective officers, directors, employs and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

12. Other Terms and Conditions:

v1.4

Questions? customercare@copyright.com.