Pharmacology & Therapeutics 250 (2023) 108530

Contents lists available at ScienceDirect

Pharmacology & Therapeutics

journal homepage: www.elsevier.com/locate/pharmthera

Use of big data and machine learning algorithms to extract possible treatment targets in neurodevelopmental disorders

Muhammad Ammar Malik [a], Stephen V. Faraone [b], Tom Michoel [a], Jan Haavik [c,d,*]

[a] Computational Biology Unit, Department of Informatics, University of Bergen, PO BOX 7803, 5020 Bergen, Norway
[b] Department of Psychiatry, Norton College of Medicine at SUNY Upstate Medical University, 13210, NY, USA
[c] Department of Biomedicine, University of Bergen, PO BOX 7804, 5020 Bergen, Norway
[d] Bergen Center for Brain Plasticity, Division of Psychiatry, Haukeland University Hospital, PO BOX 1400, 5021 Bergen, Norway

ARTICLE INFO

Available online 12 September 2023

Associate Editor: J. Harro

Keywords:
Neurodevelopmental disorders
Neuropsychiatric disorders
Machine learning
Therapeutic targets
Drug repurposing
Genomics

ABSTRACT

Neurodevelopmental disorders (NDDs) impact multiple aspects of an individual's functioning, including social interactions, communication, and behaviors. The underlying biological mechanisms of NDDs are not yet fully understood, and pharmacological treatments have been limited in their effectiveness, in part due to the complex nature of these disorders and the heterogeneity of symptoms across individuals.

Identifying genetic loci associated with NDDs can help in understanding biological mechanisms and potentially lead to the development of new treatments. However, the polygenic nature of these complex disorders has made identifying new treatment targets from genome-wide association studies (GWAS) challenging.

Recent advances in the fields of big data and high-throughput tools have provided radically new insights into the underlying biological mechanism of NDDs. This paper reviews various big data approaches, including classical and more recent techniques like deep learning, which can identify potential treatment targets from GWAS and other omics data, with a particular emphasis on NDDs. We also emphasize the increasing importance of explainable and causal machine learning (ML) methods that can aid in identifying genes, molecular pathways, and more complex biological processes that may be future targets of intervention in these disorders.

We conclude that these new developments in genetics and ML hold promise for advancing our understanding of NDDs and identifying novel treatment targets.

Contents

Abbreviations: ADHD, Attention deficit hyperactivity disorder; AI, Artificial intelligence; ASD, Autism spectrum disorder; GWAS, Genome-wide association study; ML, Machine learning; MR, Mendelian randomization; NDD, Neurodevelopmental disorders; NPD, Neuropsychiatric disorders; PGC, Psychiatric Genomics Consortium; SNP, Single-nucleotide polymorphisms.

* Corresponding author at: Department of Biomedicine, University of Bergen, PO BOX 7804, 5020 Bergen, Norway.
E-mail address: Jan.Haavik@uib.no (J. Haavik).

https://doi.org/10.1016/j.pharmthera.2023.108530

1. Introduction

Psychiatric disorders affect 20–25% of the population in any given year (Vos et al., 2017) and >50% of the general population in middle- and high-income countries will suffer from at least one such disorder during their lives (Steel et al., 2014). Anxiety, depression and substance use disorders (SUDs) are the most prevalent psychiatric disorders in adults, while anxiety and neurodevelopmental disorders (NDDs) such as attention deficit hyperactivity disorder (ADHD) are common in children (Kessler & Wang, 2008).

NDDs typically manifest early in development, often before the child enters grade school, and are characterized by developmental deficits that impair personal, social, academic, or occupational functioning (Association, A. P, 2022). According to current psychiatric classification, the NDDs include ADHD, as well as intellectual-, communication- and autism spectrum disorder (ASD), specific learning disorders, motor disorders, tic disorders and other neurodevelopmental disorders (Association, A. P, 2022). The NDDs have variable clinical presentations, a range of severities and often coexist (are comorbid with) other NDDs, as well as other psychiatric disorders. Different NDDs share genetic and environmental risk factors and share some genomic causes. Except for ADHD, where stimulant therapy was discovered in 1937, other NDDs have no effective pharmacological treatments for their core symptoms.

There are many explanations for this lack of progress, including the complex and heterogeneous biology and lack of valid animal models for NDDs. Furthermore, current classification systems of neuropsychiatric disorders (NPDs), as described in the Diagnostic and Statistical Manual of Mental Disorders (DSM) and International Classification of Disease (ICD) diagnostic manuals, are mainly based on tradition and practical utility, not on underlying biology. Current diagnostic entities probably include conditions with different underlying biology and potentially treatable targets.

Furthermore, recent clinical, epidemiological, and biomarker studies (including genetics and brain imaging studies) have reported strong biological relationships between conditions that previously were considered to be distinct diagnostic categories within clinical psychiatry or neurology (Radonjić et al., 2021; Smoller, 2019). For instance, genetic variants affecting human ion channels are associated with several rare and common disorders within neurology and psychiatry, as well as cardiology and endocrinology (Alam, Svalastoga, Martinez, Glennon, & Haavik, 2023). Together, these shortcomings have impeded the development of new, effective treatments. Recently there has been more optimism in this field, partially due to increased mechanistic insights and the introduction of new research tools. The limitations of traditional diagnostic systems triggered an intensive search for new, biologically informed classifications of mental disorders, including the Research Domain Criteria (Sanislow, Ferrante, Pacheco, Rudorfer, & Morris, 2019). Probably, the most significant advances introduced in this field have been in the use of big data, especially in genomics and new high-throughput tools to analyze such data.

Twin, family, and molecular genetic studies have demonstrated a strong genetic contribution to all common NDDs (Smoller, 2019). For example, ADHD and ASD have reported heritability estimates of approximately 74% and 83%, respectively (Faraone & Larsson, 2019; Sandin et al., 2017). Most NDDs are polygenic disorders, with both common and rare genetic variation contributing to their etiology. Genome-wide association (GWA) studies (GWAS) conducted by the Psychiatric Genomics Consortium (PGC) and others have revealed strongly associated genetic loci for NDDs, as well as other common psychiatric disorders, such as major depression, bipolar disorder and schizophrenia (Smoller, 2019). For instance, recent meta-analyses found 27 independent risk loci for ADHD (Demontis et al., 2023), 12 loci for ASD (Grove et al., 2019), 64 loci for bipolar disorder (Mullins et al., 2021) and 287 for schizophrenia (Trubetskoy et al., 2022).

Those loci are typically enriched for genes expressed in brain cells and genes linked to the mechanisms of drugs used to treat psychiatric disorders, as well as drugs in other drug classes, opening up possibilities for drug repurposing (Wu et al., 2019).

Especially during the past decade, many large genetic studies and other biomarker investigations have provided new insights into genetic risk factors underlying common psychiatric disorders. For instance, in July 2022 the GWAS central webpage listed >70 million associations between 3.3 million unique SNPs and 1451 unique MeSH/disease/phenotype descriptions. Altogether, these impressive results have clearly demonstrated that genetic risk factors for psychiatric disorders are not confined to discrete categories as defined in diagnostic manuals. In a large study of psychiatric genetics (232,964 cases and 494,162 controls across eight disorders), 109 distinct genetic loci were associated with at least two psychiatric disorders. These included 23 loci with pleiotropic effects on four or more disorders and 11 loci with antagonistic effects on several disorders. Several of the identified risk loci contain genes encoding proteins recently implicated in multiple NPDs, such as calcium channel subunits and proteins involved in glutamatergic neurotransmission as well as completely unexplored, potentially druggable proteins and pathways (Lee et al., 2019). More recently, it has been shown that rare protein coding variants also contribute strongly to common NPDs and that many of these protein coding variants, as well as common variants, fall into biological pathways shared across multiple disorders (Lal et al., 2020; Satterstrom et al., 2019).

Drug discovery and development are complex processes that have applied various automated procedures for screening candidate drugs and analyzing large data sets. More recently, artificial intelligence (AI), in particular machine learning (ML) algorithms have been adopted by the pharmaceutical industry and academic laboratories involved in drug discovery and development. The use of ML in target identification and validation, compound screening and lead discovery, as well as preclinical and clinical studies has been subject to recent reviews and commentaries (Dara, Dhamercherla, Jadav, Babu, & Ahsan, 2022; Vamathevan et al., 2019).

Here we review both classical and more modern (e.g., deep learning) "big data" approaches to find potential treatment targets from GWAS and related omics data, with a particular focus on NDDs. We highlight the emergence of explainable and causal ML methods to identify genes, molecular pathways, and higher-order biological processes implicated in disease. Together, these approaches have the potential to dramatically improve our ability to discover new treatments for human diseases.

2. Genome guided target identification

Many complex diseases are caused in part by alterations of DNA sequences. Drug targets that are supported by human genetic evidence are more likely to lead to approved drugs (King, Davis, & Degner, 2019; Nelson et al., 2015). This has been shown for common variants in common disorders, as well as for rare Mendelian diseases (King et al., 2019). Thus, it appears logical to use genetic data to explore the underlying biology and search for druggable targets for NPDs and other complex disorders. However, so far this approach has not been very successful. This failure has many explanations, including the polygenic nature of complex conditions and the fact that the genomic loci causing a disorder can be different from treatment response.

There are hundreds or thousands of common genetic variants that increase the risk for any complex disorder, with each variant typically contributing only a small risk. For example, Demontis et al. estimated that ≥7000 common variants contribute to the polygenic risk for ADHD (Demontis et al., 2023). As the vast majority of such risk variants of common disorders are located in noncoding regulatory regions (Shendure, Findlay, & Snyder, 2019), it is often challenging to identify the specific genes, biological pathways and cells affected by these variants.

Most earlier studies have lacked the resources to interpret the functional roles of genetic variants associated with NPDs and other complex

conditions. However, this situation is now changing dramatically. As more functional annotations of the genome in specific tissues (e.g., ENCODE (de Souza, 2012), PsychENCODE (Consortium*, P, 2018), the Allen Brain Atlas (Shen, Overly, & Jones, 2012), the Brainnetome (Jiang, 2013), GTEx (Lonsdale et al., 2013), PGC (Sullivan et al., 2018) etc.) as well as information on rare variants (i.e., exome sequencing in thousands of individuals) become available, fine mapping tools (e.g., Summary based Mendelian Randomization (SMR), HEterogeneity In Dependent Instrument (HEIDI) (Wu et al., 2019) and sc-linker (Jagadeesh et al., 2022)) to prioritize genes and cell types with common variant association signals have been developed. Such fine mapping tools use the publicly available summary statistics of GWAS (thus, no ethically sensitive individual level genetic data is needed). They can interpret the global enrichment of association signals within NPD-associated genes, pinpointing genes with specific functions and, thus, relevant for drug discovery. Fine mapping also aids in identifying disorder-specific drug targets. For example, one can leverage the specificity of rare variant signals in one disorder (e.g., schizophrenia (SCZ)) to interpret loci with common variants associated with multiple NPDs.

Genome-guided target identification for selection and prioritization of drug targets typically includes (i) discovering coincident genetic variants associated with both disease risk and other quantitative traits (e.g., brain imaging phenotypes); (ii) finding the causal genes responsible for these coincident associations and determining the direction of effects; and (iii) refining of the causal relationships and collecting further evidence for the significant role of the therapeutic target in the disease process, based on the available biological information and new focused experiments (Fig. 1).

As already mentioned, large-scale GWASs organized by PGC and others have revealed promising candidate genes or pathways for common NPDs for further study along this genome-guided target identification pipeline. Although available sample sizes have so far been limiting the success of GWASs and sequencing efforts for many childhood onset conditions, including most NDDs, these conditions can also benefit from the successful genetic studies of other psychiatric disorders that show genetic overlap with NDDs (Lee et al., 2019). Thus, information obtained for one diagnostic category may also be relevant for other (comorbid) conditions. Below follows a brief review of findings from recent GWASs on NDDs and other relevant NPDs.

A genome-wide association meta-analysis of 20,183 individuals diagnosed with ADHD and 35,191 controls and found 12 independent genome wide significant loci (Demontis et al., 2019). Later, they increased the sample size to 38,691 cases with ADHD and 186,843 controls and identified 27 loci with 76 potential risk genes. These hits were enriched among genes expressed in early brain development (Demontis et al., 2023). Although none of the known pharmacological targets of drugs used to treat ADHD were among the top hits, several of the implicated risk genes were considered to be druggable, also suggesting that existing drugs might be repurposed for ADHD treatment (Hegvik et al., 2021).

The Autism Spectrum Disorder Group of the PGC conducted a meta-analysis and replication study of 16,539 individuals with ASD and a total of 157,234 controls (Consortium, T. A. S. D. W. G. of T. P. G, 2017). They identified a genome-wide significant locus (rs1409313-T) at 10q24.32 which contains several genes including PITX3, a transcription factor, and CUEDC2, a cell cycle regulator. Similarly, Grove et al. (Grove et al., 2019) conducted a meta-GWAS of 18,381 ASD cases and 27,969 controls from a unique Danish population resource. The study identified 5 genome-wide-significant loci. Moreover, 7 additional loci were identified by utilizing the GWAS results from three phenotypes with significantly overlapping genetic architectures (schizophrenia, major depression, and educational attainment).

A large-scale GWAS using the data from 41,917 cases of bipolar disorder and 371,549 controls of European ancestry identified 64 associated genomic loci (Mullins et al., 2021). The study found significant signal enrichment in genes encoding targets of antipsychotics, calcium channel blocker, antiepileptic and anesthetics. Furthermore, the study integrated expression quantitative trait locus (eQTL) data and identified 15 genes strongly associated to bipolar disorder via gene expression, encoding druggable targets such as *HTR6*, *MCHR1*, *DCLK3* and *FURIN*.

Schizophrenia has been subject to multiple GWAS, with gradually increasing sample sizes and number of genome wide significant hits. In one of the largest genetic studies on schizophrenia a two-stage GWAS of 76,775 cases and 243,649 controls was conducted (Trubetskoy et al., 2022). They identified common variant associations at 287 distinct genomic loci. Further analysis showed that the fine-mapped candidates were enriched for genes associated with rare disruptive coding variants, including the glutamate receptor subunit *GRIN2A* and transcription factor *SP4*.

The standardized methods and data output format from molecular genetic studies has allowed for combination of multiple data sets and comparison of findings across different sites and disorders. The Cross-Disorder Group of the PGC analyzed data from 33,332 cases and 27,888 controls and identified four risk loci that have shared effects on five major psychiatric disorders, including autism spectrum disorder, attention deficit hyperactivity disorder, bipolar disorder, major depressive disorder, and schizophrenia (Consortium, C.-D. G. of the P. G, 2013). These loci include regions on chromosomes 3p21 and 10q24, and SNPs within two L-type voltage-gated calcium channel subunits, *CACNA1C* and *CACNB2*. The cross-disorder group conducted a larger study in 2019 on data from 232,964 cases and 494,162 controls across eight psychiatric disorders i.e., 3 disorders (anorexia nervosa, obsessive-compulsive disorder, Tourette syndrome) in addition to the five disorders from the previous study (Lee et al., 2019). The study identified 109 loci associated with at least two of the eight psychiatric disorders; 23 of these loci were shown to have pleiotropic effects on four or more disorders and 11 loci showed antagonistic effects on multiple disorders. The identified pleiotropic loci are located within genes that show heightened expression in the brain and play important roles in neurodevelopmental processes.

Further exploitation of these genetic loci and candidate genes for drug discovery has been challenging because identification of genes in GWAS are traditionally based on physical proximity to the genetic association signal, and therefore no direct causal effect of the gene on the disease is established. However, by integrating summary data from GWAS and studies examining the effect of genetic variants on gene or protein expression levels in specific cells types, some of these limitations can be addressed (Jagadeesh et al., 2022; Reay & Cairns, 2021; Zhu et al., 2016).

Colocalization analysis is a statistical method used to determine whether multiple genetic variants associated with different traits are located in the same region of the genome. It is used to identify potential causal genes and pathways that may be involved in disease development by comparing the patterns of association between the genetic variants and the traits of interest. If the variants show a high degree of correlation with each other and with the traits, it suggests that they are likely affecting the same biological process. On the other hand, if the patterns of association are different between the variants and the traits, it suggests that they are affecting different pathways.

A probabilistic method eCAVIAR (eQTL and GWAS Causal Variant Identification in Associated Regions) has been proposed for estimating the posterior probability that a same genetic variant is causal for both GWAS and eQTL study (Hormozdiari et al., 2016). eCAVIAR can account for more than one causal variant in a given genomic locus. Moreover, it can use summary statistics without the need of the individual genotype data.

In a recent study (Wallace, 2021), it was shown that the Sum of Single Effects (SuSiE) regression framework can be used for colocalization analysis to evaluate evidence for association at multiple genetic variants simultaneously, while separating the statistical support for each variant based on the causal signal being examined. SuSiE can identify clusters of genetic variants that are likely to share

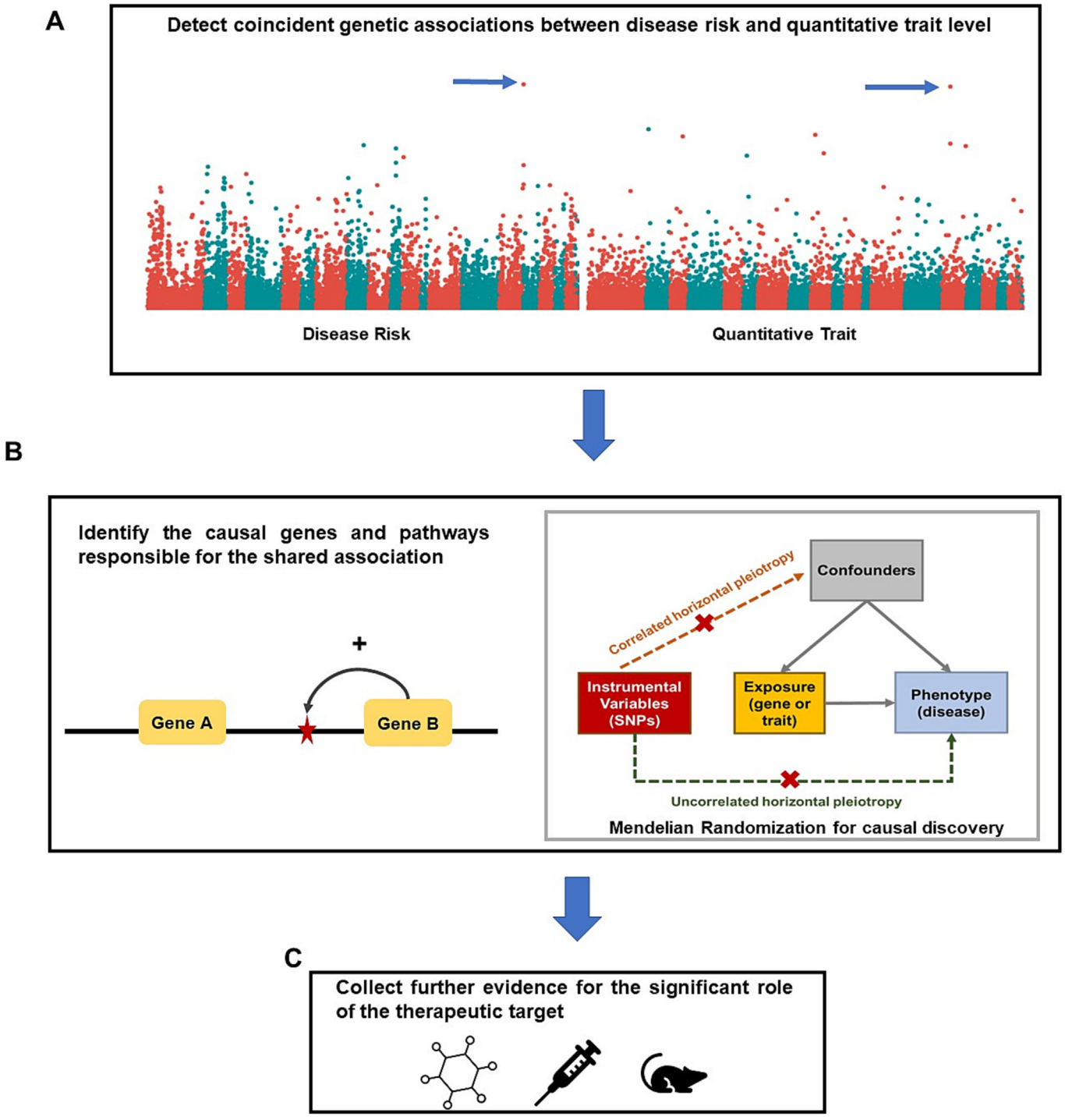

Fig. 1. Genetically driven identification of therapeutic targets. (**A**) Identification of common genetic associations between disease risk and quantitative trait such as brain imaging phenotype (**B**) Inferring the causal genes and pathways responsible for the shared association using Mendelian Randomization. (**C**) Gathering more evidence for a therapeutic target's crucial role in relation to the disorder.

a common causal signal and thus are more likely to be involved in the same biological pathway or mechanism underlying the traits of interest. This approach can improve the accuracy of identifying the causal variants involved in complex diseases and can help in the identification of novel therapeutic targets.

Similarly, ColocQuiaL pipeline has been proposed to provide a framework for performing the colocalization analyses (Chen et al., 2022) and the sc-linker pipeline for integrating single-cell RNA-sequencing, epigenomic SNP-to-gene maps and GWAS summary statistics to infer cell types and processes involved in disease (Jagadeesh et al., 2022).

Mendelian Randomization (MR) is a statistical technique that uses genetic variants that are associated with both the exposure (e.g., expression of a candidate gene in a relevant tissue) and outcome of interest (e.g., disease risk) as instrumental variables (IV) to estimate the causal effect of the exposure on the outcome (Davey Smith & Hemani, 2014; Davies, Holmes, & Smith, 2018; Ebrahim & Davey Smith, 2008; Evans & Davey Smith, 2015). MR assumes that

the genetic variants are associated with the exposure of interest, that they do not affect the outcome through any pathway other than the exposure, and that they are not associated with any confounding variables that might bias the estimates of the causal effect (Fig. 1**B**). When these assumptions are valid, the causal effect of the exposure on the outcome can be estimated by regressing the outcome on the genetically predicted values of the exposure.

Applications of MR for identifying candidate drug targets for NPDs have begun to appear. In a recent study (Liu et al., 2022), MR analysis integrated GWAS and brain-derived transcriptome and proteome data of 1263 actionable proteins; 25 potential drug targets for schizophrenia, bipolar disorder, depression and ADHD were identified. Similarly, (Wingo et al., 2021) identified 19 genes as causal factors for depression by integrating GWAS results with human brain proteomes.

Although MR can provide evidence that a candidate GWAS gene has a causal effect on disease, it is still limited by being focused on a single locus and single candidate gene at a time. To gain a more holistic perspective, including to identify downstream pathways affected by GWAS genes, more data-intensive ML approaches are required.

3. Machine learning primer

Machine learning is a field of study that aims to develop algorithms that can learn patterns and relationships from data. There are two main types of ML: supervised and unsupervised. In supervised learning, the algorithm is trained on labeled data, that is, data from both the inputs (predictors) and outputs, to predict the outputs from the input, while in unsupervised learning, the algorithm is only provided with inputs and must find patterns or relationships within the data without the use of labeled outputs.

The typical ML process begins with collecting data and dividing it into three subsets: *training*, *validation*, and *test data*. The algorithm is then trained using the training data, and the validation data is used to fine-tune the model and improve its performance. After training the ML model, its performance is evaluated using the unseen test data (Fig. 2**A**). Careful preprocessing of the data is essential to avoid "data leakage", that is, accidental transfer of information from the training to the validation and test data (e.g., when variables are standardized prior to splitting the data) (Kaufman, Rosset, Perlich, & Stitelman, 2012). Data leakage has been documented as affecting many genomic ML studies (Barnett, Zhang-James, & Faraone, 2022).

A model is a mathematical representation of a problem or relationship in the data. ML models are created by training algorithms on data and then used to make predictions or decisions. Feature extraction or feature selection is a crucial step in the ML process. A feature is an attribute or characteristic of the data that the algorithm uses to make predictions or decisions. The loss function is a measure of how well the model is performing, used to guide the training process, for instance a

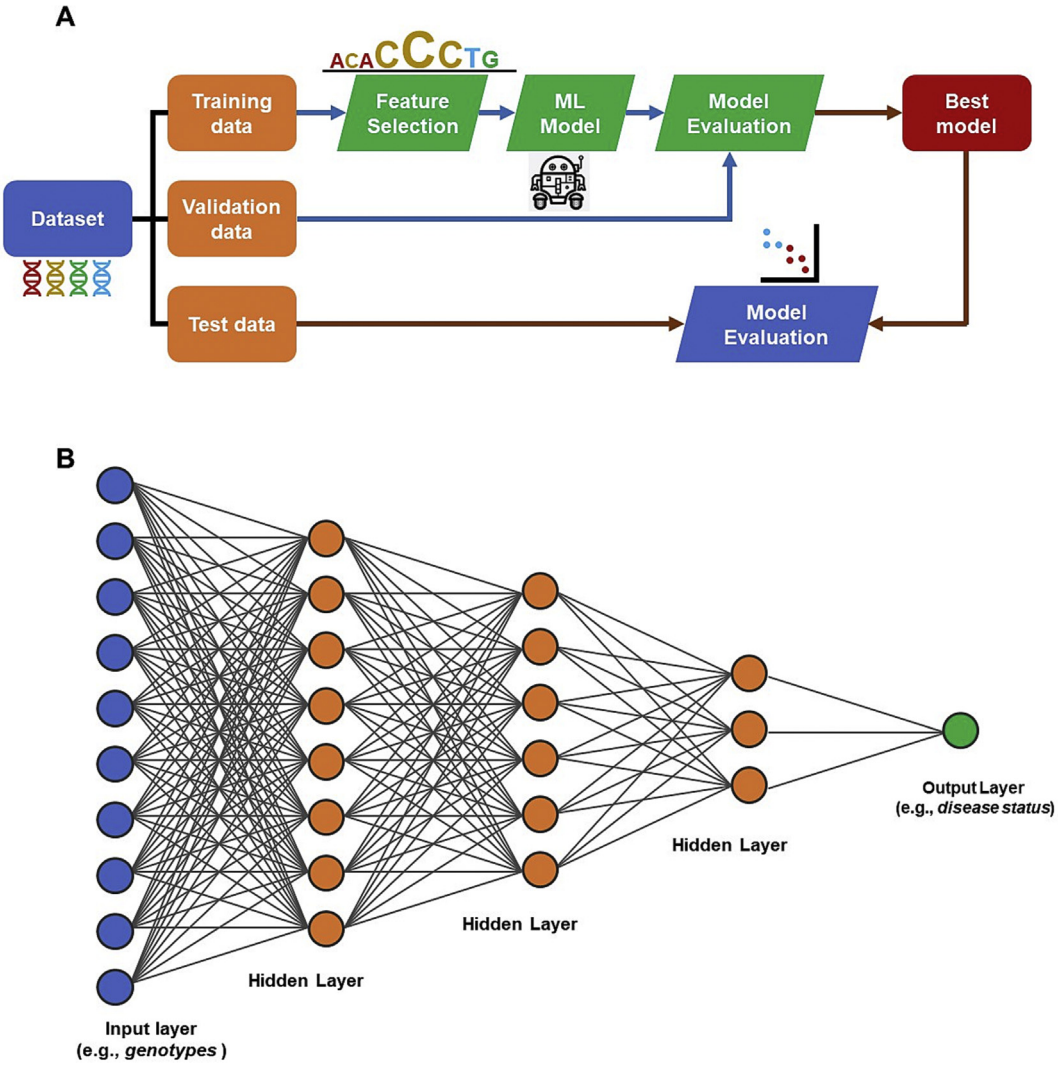

Fig. 2. (**A**) Illustration of a typical ML process using genetic data to identify underlying patterns. (**B**) Illustration of typical deep neural network (DNN) architecture using genotype data to predict a phenotype such as disease status.

Table 1
Glossary of some machine learning algorithms.

Linear Regression: A simple algorithm for performing regression, where the relationship between the independent and dependent variables is modeled as a linear function.
Logistic Regression: A type of regression algorithm used for binary classification problems, where the goal is to predict a binary outcome (e.g., yes/no, true/false).
Decision Trees: An algorithm that creates a tree-like model of decisions and their possible consequences, used for both classification and regression tasks.
Random Forest: An ensemble learning method that operates by constructing multiple decision trees and aggregating their results to make a final prediction.
Naive Bayes: A probabilistic algorithm used for classification, based on Bayes' Theorem, which assumes independence between features.
k-Nearest Neighbors (*k*−NN): A non-parametric, instance-based algorithm for classification and regression, where the prediction for a given data point is based on its k-nearest neighbors.
Support Vector Machines (SVMs): A type of algorithm for classification and regression analysis, which seeks to find the best boundary (hyperplane) to separate data into classes by maximizing the margin between the classes.
Convolutional Neural Networks (ConvNets or CNNs): A type of deep neural network architecture commonly used in computer vision tasks, that uses convolutional layers to scan and analyze local features in images.
Graph Neural Networks (GNNs): A class of deep learning models designed to process and learn from data represented in graph structures.
Recurrent Neural Networks (RNNs): A type of neural network specialized in processing sequential data, where the output at each step is influenced by previous computations.
Principal Component Analysis (PCA): A dimensionality reduction technique that seeks to find the principal components, or directions of maximum variance, in the data.
k-Means Clustering: An unsupervised learning algorithm for grouping data into clusters based on similarity.
Bayesian Networks: probabilistic graphical models for representing relationships between variables and making predictions based on available data.

likelihood function that expresses how probable it is to observe the training data given the current model parameters.

Traditional ML methods such as *linear regression*, *random forests*, *support vector machines* etc. (Table 1) rely on the use of domain knowledge to transform features from the raw data before model training. A simple example is transforming genotypes into polygenic risk scores or gene set polygenic risk scores (Barnett et al., 2022).

Neural network-based ML models can eliminate the need for handcrafting features by translating the data into compact intermediate representations. Artificial neural networks (ANNs) are inspired by the structure and function of the human brain and consist of interconnected nodes or "neurons". ANNs with multiple hidden layers, allowing for end-to-end training on large datasets, can learn and make predictions or decisions, mimicking the way the human brain works. Neural network models with more than two hidden layers are typically referred to as "deep neural networks (DNNs)" (Fig. 2**B**). Neural network models such as convolutional neural networks (CNNs), recursive neural networks (RNNs) and graph neural networks (GNN) are different variants of DNN architectures. Some of the common DNN architectures are briefly summarized in Table 1.

However, training machine learning models is challenging due to issues such as overfitting, where the model fits too closely to the training data and is not able to generalize well to new, unseen data, or underfitting, where the model is too simple and does not fit the data well enough, leading to poor performance on the training data. Regularization is a technique used to prevent overfitting by adding a model complexity penalty term to the loss function or using dropout (randomly disabling neurons in a neural network) during training.

4. Applications of machine learning in drug discovery

Many ML methods have been developed for analyzing the rapidly growing amounts of human genetic data (currently including millions of individuals) (Lal et al., 2020; Pardiñas et al., 2018). These methods can explore patterns across traditional diagnostic and genome-wide association boundaries to find shared and unique signatures of their biology and previously unknown therapeutic targets and link this information with existing therapeutic targets and ligand libraries (Hegvik et al., 2021).

In drug discovery studies for schizophrenia, AI/ML methods have been used for tasks including drug target identification (Hsu & Wang, 2017; Yang et al., 2019), developing quantitative structure–activity relationship (QSAR) models (Marunnan et al., 2017), monitoring dosing compliance (Bain et al., 2017), predicting G protein-coupled receptors (GPCRs) targeting compounds (Yang et al., 2019), and drug repositioning (Zhao & So, 2018). For example, in (Yang et al., 2019) an SVM-RFE (recursive feature elimination)-based feature selection and classification method was used to identify a biomarker signature for presynaptic dopamine overactivity, which may be responsible for schizophrenia. SVM classifiers have also been useful for predicting QSAR models of the GABA (gamma aminobutyric acid) uptake inhibitor drugs that are helpful in the treatment of schizophrenia (Marunnan et al., 2017). In addition, SVMs trained on drug-response expression profiles from the Connectivity Map (Lamb et al., 2006) showed better performance compared to other ML methods for predicting drug repositioning candidates for SCZ (Zhao & So, 2018).

AI/ML based methods have also been used in drug discovery studies for ASD. For example, an improved performance in drug response prediction of ASD patients was observed using cluster analysis (i.e., affinity propagation and k-medoids) of clinical data (i.e., signs and biomarkers) (Obara et al., 2018). In (Ekins et al., 2019), a Bayesian ML model was trained on high-throughput screening data, and it revealed a repurposing potential of nicardipine or other dihydropyridine calcium channel antagonists for the treatment of Pitt Hopkins syndrome, a rare genetic disorder exhibiting features of ASD. Moreover, ML algorithms have been successful for predicting the functional effects of variants in voltage-based sodium and calcium ion channels, that have been known to be associated with ASD, schizophrenia and developmental encephalopathy (Heyne et al., 2020). Here, ML models were trained on sequence- and structure-based features to predict the gain or loss of function effects of potential pathogenic missense variants in ion channels and exome-wide data was used for result validation.

Drug discovery research for NDDs can potentially benefit from ML-based drug discoveries for other disorders, including NPDs. Recently, (Pan et al., 2023) developed a deep learning-based prediction framework (AI-DrugNet) for identifying drug repurposing opportunities for Alzheimer's disease (AD). The authors first built a network of drug-target pairs (DTP) based on multiple features related to the drugs and targets. They then incorporated additional information about the relationships between drugs and targets both within and outside of DTPs and trained a model to predict synergistic drug combinations.

In (Luo et al., 2017), the authors propose a computational pipeline called DTINet for predicting novel drug-target interactions. DTINet uses a heterogeneous network that integrates diverse drug-related information and focuses on learning a low-dimensional vector representation of features to accurately explain the topological properties of individual nodes in the network, and then uses these representations for predicting new drug-target interactions. Similarly, deepDR (https://github.com/ChengF-Lab/deepDR) is a network-based deep learning approach for in silico drug repurposing. The approach integrates ten different networks, including those related to drugs, diseases, targets, side-effects, and drug-drug interactions (Zeng et al., 2019). The deepDR approach uses a multi-modal deep autoencoder to learn a low-dimensional representation of drugs and with clinically reported drug-disease pairs from these interaction networks to infer new target diseases for drugs originally approved for other diseases. The authors

Table 2
Studies on machine learning algorithms for various NDDs and tasks.

Study	Disorder	ML algorithm	Task
Hsu & Wang, 2017	SCZ	SVM	Pathogenesis, biomarker detection, drug target discovery
Yang et al., 2019	SCZ	SVM	Identification of target genes
Marunnan et al., 2017	SCZ and depression/anxiety disorder	SVM	Prediction of QSAR models
Zhao & So, 2018	SCZ	SVM, RF, GBM, logistic regression, DNN	Drug discovery or repositioning based on drug expression profile
Obara et al., 2018	ASD	Clustering algorithms (affinity propagation, k-medoids)	Drug response prediction
Ekins et al., 2019	ASD	Bayesian machine learning	Drug repurposing
Heyne et al., 2020	Multiple	GBM, RF, SVM, logistic regression	Predicting functional effects of variants in voltage-based sodium and calcium ion channels
Liu et al., 2018		Kernel machine regression	Predicting the neurodevelopmental toxicity of compounds
Wang et al., 2018	Multiple	DNN	Identification of key genes and pathways associated with the disorder
Nguyen, Jin, & Wang, 2021	SCZ	DNN	Prioritize variants, genes and regulatory linkages
Pan et al., 2023	AD	GNN	Identification of potential repurposed drug therapies for AD
Luo et al., 2017			Drug-target interaction prediction and drug repositioning
Zeng et al., 2019	AD, PD	DAE	Drug-target interaction prediction and drug repositioning
Altae-Tran, Ramsundar, Pappu, & Pande, 2017		One-shot learning, Long short-term memory (LSTM)	Drug discovery with less data

SCZ: Schizophrenia, ASD: Autism spectrum disorder, AD: Alzheimer's disease, PD: Parkinson's disease, SVM: Support vector machines, RF: Random forests, GBM: Gradient boosting machines, DNN: Deep neural networks, GNN: Graph neural network, DAE: Deep auto encoder.

found that deepDR outperforms DTINet and other conventional ML methods for this task (Zeng et al., 2019).

Deep learning methods often require large amounts of training data, but the lack of publicly available datasets often makes it difficult to train such models. To address this issue, (Altae-Tran et al., 2017) showed that one-shot learning (an ML approach that can learn from a single example or a few instances of a class) significantly lowers the amounts of data required for making meaningful predictions in drug discovery applications. The model is part of DeepChem, an open-source framework for deep-learning in drug discovery (https://github.com/deepchem/deepchem). Table 2 lists some of the selected studies where machine learning approaches have been used with regards to various NDDs and tasks.

5. Applications of machine learning in identifying causal mechanisms

While ML has been applied at several stages of the drug discovery and repurposing pipelines as illustrated above, it is not straightforward to connect these approaches directly to GWAS-based candidate genes, because these genes may not code for druggable proteins. Hence it is important to also apply ML to map causal molecular pathways, networks, and processes downstream of GWAS candidate genes to identify potential treatment targets.

One type of ML model that has emerged in this context is the "gray box" (as opposed to black box) neural network whose structure is partially defined by prior biological knowledge. For instance, a Deep Structured Phenotype Network (DSPN) is an artificial neural network (ANN) model for predicting psychiatric phenotypes from genetic variation and gene expression data (Wang et al., 2018). Unlike conventional ANNs where the input data is progressively integrated and processed through hidden layers of artificial neurons in a feedforward manner, DSPN identifies artificial neurons with genes, and embeds a gene regulatory network (GRN) of known transcription factor - target interactions for the brain in the ANN connections. In addition to an improved predictive performance, by tracing important paths in the ANN structure, the model is also able to highlight key genes and pathways associated with the disorder, including immunological, synaptic and metabolic pathways (Wang et al., 2018).

Similarly, a deep neural network model (called Varmole) has been proposed, based on a "biological DropConnect" mechanism for prioritizing disease risk variants and genes (Nguyen et al., 2021). DropConnect is an effective regularization technique for deep neural networks which is based on a random selection of weight connections between the two consecutive layers of the network. Instead of this random selection, the proposed model uses GRNs and eQTLs as prior biological knowledge for selecting the connections, hence the term biological DropConnect. Again, this allows to attribute biological mechanisms to model predictions (Fig. 3).

In a slightly different context, a similar approach called a Visible Neural Network (VNN) has been developed to predict the response of cancer cell lines to drug treatment and, simultaneously, learn biological mechanisms underlying the drug response (Kuenzi et al., 2020). VNNs are artificial neural networks where the hidden layers and their connections are modeled after the hierarchical structure of biological processes in the Gene Ontology database.

Conventional ML algorithms rely on correlations in the data for making accurate predictions. However, in many cases, these correlations do not represent true causal relationships between the variables but may be attributed to confounding variables. Thus, a parallel line of research has sought to expand the use of Mendelian randomization from pairwise analyses between exposure and outcome variables to the reconstruction of large-scale models of causal molecular networks from multi-omics data. These models take the form of Bayesian networks, a type of probabilistic ML models that provide convenient means of expressing prior knowledge about a system, facilitate compact representations of statistical dependences and independences among large numbers of variables, and allow efficient inferences from observational data (Koller & Friedman, 2009; Pearl, 2009). Causal Bayesian network models linking genetic risk variants to gene networks and phenotypes have provided numerous insights into candidate targets and causal mechanisms underlying complex diseases.

For instance, using genotype and liver gene expression data, causal genes at risk loci for type 1 diabetes, coronary artery disease, and plasma low-density lipoprotein cholesterol levels were identified (Schadt et al., 2008). Using genomic, transcriptomic, and proteomic data from postmortem samples from four brain regions of late-onset Alzheimer's disease cases and nondemented individuals, the genes *TYROBP* (Zhang et al., 2013) and *VGF* (Beckmann et al., 2018) were identified as key regulators in multiple AD causal networks and experimentally validated in mouse models. In a study by Talukdar et al., coronary artery disease (CAD) causal Bayesian gene networks were identified by utilizing genomic and transcriptomic data obtained from seven vascular and metabolic tissues of individuals with CAD who underwent surgical

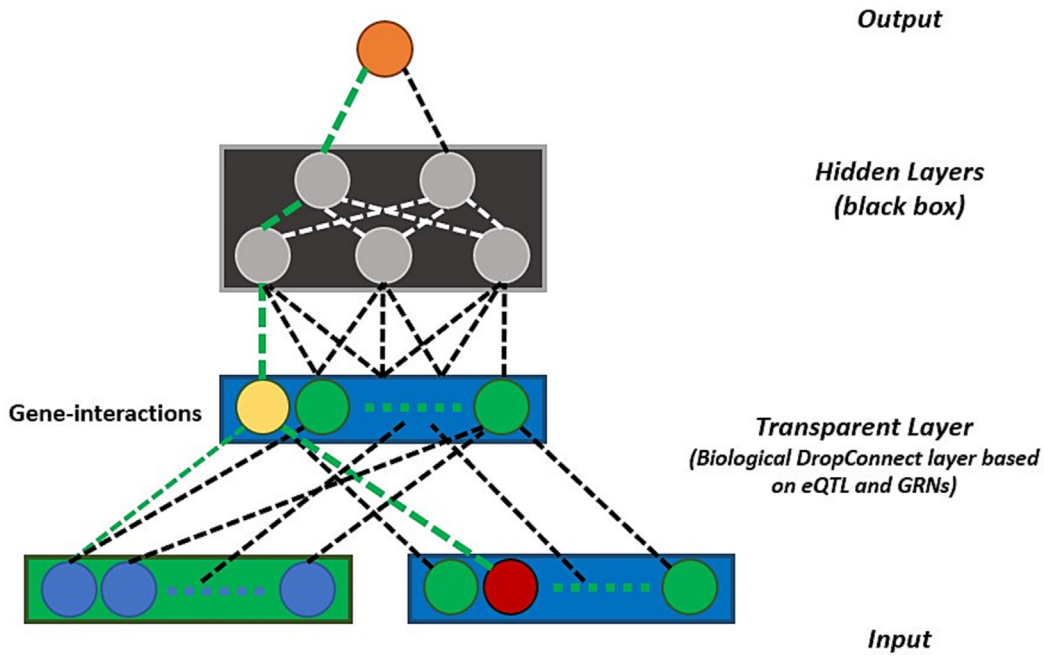

Fig. 3. Illustration of Varmole as proposed by (Nguyen et al., 2021). The model takes genotype and gene expression data as input. The first layer utilizes prior biological knowledge based on eQTL and GRNs. The trained model is interpretable and can reveal information about important inputs and important pathways with regards to output prediction (figure reproduced from Nguyen et al., 2021) under Creative Commons Non-Commercial Attribution License.

intervention (Talukdar et al., 2016). These networks were found to be replicable in corresponding tissues of the Hybrid Mouse Diversity Panel. A similar approach could also be applied to target identification in NDDs, where increasingly larger data sets containing genomic, transcriptomic, and proteomic data are also becoming available (see below).

Despite these successes, it should be noted that Bayesian network reconstruction from high-dimensional data is a challenging task (Wang, Audenaert, & Michoel, 2019). Although it has been found that edge-to-edge reproducibility is strongly dependent on sample size, identification of more highly connected central regulators ("key driver genes") in Bayesian networks can be carried out with high confidence across a range of sample sizes (Cohain et al., 2017).

Such key driver genes directly suggest candidate genes for follow-up. For instance, additional analysis in the context of protein interaction and pharmacological databases can be used to identify established and novel druggable targets and target tissues (Lempiäinen et al., 2018).

Further advances are expected by integrating the principles of causal inference and Bayesian networks with neural networks and other ML methods, and this is an important area of current research in computational biology (Lecca, 2021) and the field of ML more broadly (Peters, Janzing, & Schölkopf, 2017).

6. Selected data resources

ML approaches rely heavily on availability of large amount of good quality labeled data. Therefore, public data resources play a vital role in creating better models. Several large population based research resources, such as the UK Biobank (UKBB) Database and The Norwegian Mother, Father and Child Cohort Study (MoBa), as well as diagnosis specific clinical data bases are available for researchers in this field. A brief summary of selected clinical data resources related to the NDDs is shown in Table 3. The description of each of the resource is as follows:

6.1. Psychiatric Genomics Consortium (PGC)

The Psychiatric Genomics Consortium (PGC) (Sullivan et al., 2018) is a large-scale collaborative initiative for deciphering the genomic basis of

Table 3
Overview of selected NDD-related consortia.

Consortium	No. of samples	Data type	Website
Psychiatric Genomics Consortium (PGC)	~900,000	Genomic Data including GWAS, sequencing data, CNV data, and more	https://www.med.unc.edu/pgc/
Brain eQTL Alamnac (Braineac)	134	Gene expression data	http://www.braineac.org/
Genotype-Tissue Expression (GTEx) Consortium (v7)	80 to 154	Gene expression and genotype data	https://gtexportal.org/home/
CommonMind Consortium	~1000	Genotype, RNA-seq and eQTL data	https://www.synapse.org/#!Synapse:syn2759792/wiki/
BrainSeq	~2000	Genotype, RNA sequence and DNA methylation data	https://eqtl.brainseq.org/
PsychENCODE Consortium	~2000	Transcriptomic and Epigenomic Data measuring gene ex- pression levels and epigenetic modifications	https://resource.psychencode.org/
Simons Foundation Autism Research Initiative (SFARI)	~400,000	Phenotypic and Genomic data	https://www.sfari.org/resource/spark/

psychiatric disorders. PGC has so far revealed the cryptic genetic and biological basis of number of psychiatric diseases by evaluating common single-nucleotide polymorphisms (SNPs), rare variants, gene sets and pathways, and other genetic variations. The PGC was started in 2007 for facilitating large-scale genetic analyses for five major psychiatric disorders (ADHD, autism, bipolar disorder, major depressive disorder and schizophrenia) and has since expanded to many more diagnostic categories and traits. >800 scientists from >40 countries are currently participating in the consortium. One of the major contributions of PGC has been assembly of many of the large-scale GWAS in psychiatry and identification of number of loci associated with psychiatric disorders.

The summary statistics from genomic analyses of different psychiatric disorders are publicly available to download from PGC website.

6.2. Brain eQTL resources

6.2.1. UK Brain Expression Consortium (UKBEC)

UKBEC was launched with the aim to study the regulation and alternative splicing of gene expression in multiple tissues from human brain. A part of the data and its results are freely available at Braineac - The Brain eQTL Almanac webpage (https://braineac.org). As of now the data set consists of data from 10 regions obtained from 134 control individuals (frontal cortex, temporal cortex, occipital cortex, hippocampus, thalamus, putamen, substantia nigra, medulla, cerebellum, and white matter, + mean expression across all 10 regions).

6.2.2. Genotype-Tissue Expression (GTEx) consortium

The GTEx project (Lonsdale et al., 2013) (https://gtexportal.org/home/datasets) started in 2010 with the motive to build a catalog of genetic effects on gene expression on large number of human tissues to reveal the biological mechanisms of genetic associations with complex disease and traits and for improving our understanding of regulatory genetic variation.

The GTEx v7 dataset consists of eQTL data obtained from 80 to 154 samples from 13 brain tissues (cerebellum, caudate, cortex, nucleus accumbens, cerebellar hemisphere, frontal cortex, putamen, hippocampus, anterior cingulate cortex, hypothalamus, amygdala, spinal cord, and substantia nigra).

6.2.3. CommonMind consortium

The CommonMind Consortium (Consortium, C. M, et al., 2017) provides and extensive public resources of processed and quality controlled data with the aim to provide researchers with a resource for applying novel methods and perform integrative analyses. The CommonMind Consortium has generated functional genomic data from multiple regions from 1000 postmortem brain samples from donors with Schizophrenia, Bipolar disease and individuals with no neuropsychiatric disorders. The data is collected from dorsolateral prefrontal cortex (DLPFC), anterior cingulate cortex (ACC), and superior temporal gyrus (STG) tissues in the brain.

6.2.4. BrainSeq

BrainSeq (Schubert et al., 2015), A Human Brain Genomics Consortium is an initiative launched by the Lieber Institute for Brain Development (LIBD) with pharmaceutical industry partners (Astellas Pharma, AstraZeneca, Eli Lilly and Company, F. Hoffmann-La Roche, Johnson and Johnson, Lundbeck and Pfizer). The aim of the consortium is to make use of the emerging genetic knowledge of psychiatric disorders and technical advances for the analysis of gene expression in brain tissue. It includes data from several human postmortem neuropsychiatric disease and control samples. A major aim of BrainSeq is to generate and analyze neurogenomics data (including genotype, RNA sequence and DNA methylation data). The BrainSeq Phase I tissue cohort consists of 746 postmortem samples obtained from the dorsolateral prefrontal cortex (DLPFC) of patients suffering from schizophrenia, mood disorders and other major NPDs. Phase II samples were obtained from mid-hippocampus of 200 patients with schizophrenia and 300 controls.

6.3. PsychENCODE consortium

The PsychENCODE Consortium (Jourdon, Scuderi, Capauto, Abyzov, & Vaccarino, 2021) was launched to improve our understanding of the underlying molecular mechanisms of the strong genetic associations that have been discovered for a number of psychiatric disorders. The PsychENCODE includes data from the adult brains across 1866 individuals. PsychENCODE has generated datasets including bulk transcriptome, chromatin, genotype, and Hi-C datasets and single-cell transcriptomic data from 32,000 cells for major brain regions. It has been shown that embedding the gene regulatory network that links the GWAS variants to genes into an interpretable deep learning model improves disease prediction by 6-fold versus polygenic risk scores (Wang et al., 2018). The deep learning model also helped in the identification of key genes and pathways in psychiatric disorders.

7. Conclusion

NDDs typically affect normal brain development and functions during childhood and impair many aspects of life, including social interactions, communication, productivity, self-regulation and other behaviors. Most NDDs are caused by the accumulation of multiple genetic and environmental risk factors. In addition to sporadic cases, rare Mendelian forms of these disorders are caused by the highly penetrant rare risk variants. Early identification and intervention can be critical in helping individuals with these disorders reach their full potential and lead fulfilling lives. So far, pharmacological treatments for NDDs have been of limited value, due to their moderate efficacy and risk for adverse events. For some disorders, such as ASD, no effective drug treatments have been found yet. This lack of effective pharmacological treatments for NDDs can be attributed to various factors, including their complexity and heterogeneity of symptoms across individuals. As, the underlying neural mechanisms of NDDs are not yet fully understood, it is challenging to develop drugs that target specific aspects of the disorder. However, the introduction of new high-throughput tools for big data 'omics have provided better insights into the underlying biological mechanisms of NDDs.

A possible way to explore the underlying biology of complex disorders like NDDs is to use genetic data. Genes expressed in brain cells play a crucial role in the development and function of the nervous system, and alterations in these genes contribute to the development of NDDs. Identifying genetic loci associated with NDDs that encompass these genes can help in understanding biological mechanisms, and potentially lead to the development of new treatments that target these mechanisms. Large-scale GWASs have identified many genetic loci associated with NDDs, but due to the polygenic nature of these complex disorders, identifying new treatment targets from GWASs has not been very successful. However, the increased availability of information on rare variants and functional annotations of the genome has helped in the development of fine mapping tools that can be used to prioritize genes with common variant association. One important discovery from GWASs of NDDs and other NPDs is that they show a strong genetic overlap. Thus, genetic findings in one condition may be relevant for several disorders across traditional diagnostic boundaries and genetic findings in traits and conditions with large samples sizes may be used to leverage findings in conditions with fewer available samples.

GWASs alone does not discover causal variants for disease. It only implicates genomic loci that harbor such variants. The discovery of causal variants can be addressed by integrating GWAS data with data on the effects of genetic variants on gene and protein expression in relevant tissues and cell types through colocalization and MR analyses. However, these analyses only focus on a single genetic locus and a single candidate gene at a time.

Here we have reviewed recent developments using data-intensive ML approaches that can provide a more comprehensive understanding and identification of downstream pathways by analyzing large-scale human genetic data. Despite the increased use of ML methods, there is still much room for further research in this field. An important recent development is the emergence of more interpretable deep neural network models like the recently proposed Deep Structure Phenotype Network (DSPN) and Varmole, which incorporate prior biological knowledge and provide additional insights into the biological processes underlying these disorders by identifying key genes and pathways. We anticipate that such models will gain further prominence in the coming years.

Traditional ML approaches are mainly reliant on correlations in the data. However, often these correlations are due to the confounding variables and hence do not represent true causal relationship between the variables. A promising avenue to address this issue is the use of causal Bayesian network models that link genetic risk variants to gene networks and phenotypes. We expect that the integration of traditional and deep ML methods with the principles of causal inference in the coming years will play an important role in revealing mechanisms underlying NDDs and providing further insights into candidate targets.

Data availability

No data was used for the research described in the article.

Declaration of Competing Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This study has received funding from the Research Council of Norway (Project No. 331725) and the ADHD Research Network of Norway (NevSom, Project No. 51379).

References

Alam, K. A., Svalastoga, P., Martinez, A., Glennon, J. C., & Haavik, J. (2023). Potassium channels in behavioral brain disorders. Molecular mechanisms and therapeutic potential: A narrative review. *Neuroscience & Biobehavioral Reviews 152*, 105301. https://doi.org/10.1016/j.neubiorev.2023.105301.

Altae-Tran, H., Ramsundar, B., Pappu, A. S., & Pande, V. (2017). Low data drug discovery with one-shot learning. *ACS Central Science 3*(4), 283–293.

Association, A. P (2022). *Diagnostic and statistical manual of mental disorders.* American Psychiatric Association.

Bain, E. E., Shafner, L., Walling, D. P., Othman, A. A., Chuang-Stein, C., Hinkle, J., & Hanina, A. (2017). Use of a novel artificial intelligence platform on mobile devices to assess dosing compliance in a phase 2 clinical trial in subjects with schizophrenia. *JMIR mHealth and uHealth 5*(2), Article e7030.

Barnett, E., Zhang-James, Y., & Faraone, S. V. (2022). Improving machine learning prediction of ADHD using gene set polygenic risk scores and risk scores from genetically correlated phenotypes. *MedRxiv*. https://doi.org/10.1101/2022.01.11.22269027 2022–01.

Beckmann, N. D., Lin, W. -J., Wang, M., Cohain, A. T., Wang, P., Ma, W., ... Comella, P., et al. (2018). Multiscale causal network models of Alzheimer's disease identify VGF as a key regulator of disease. *BioRxiv 458430*.

Chen, B. Y., Bone, W. P., Lorenz, K., Levin, M., Ritchie, M. D., & Voight, B. F. (2022). ColocQuiaL: a QTL-GWAS colocalization pipeline. *Bioinformatics 38*(18), 4409–4411.

Cohain, A., Divaraniya, A. A., Zhu, K., Scarpa, J. R., Kasarskis, A., Zhu, J., ... Schadt, E. E. (2017). Exploring the reproducibility of probabilistic causal molecular network models. *Biocomputing 2017*, 120–131. https://doi.org/10.1142/9789813207813_0013.

Consortium, C. M, et al. (2017). *CommonMind Consortium Knowledge Portal.*

Consortium, C.-D. G. of the P. G, et al. (2013). Identification of risk loci with shared effects on five major psychiatric disorders: A genome-wide analysis. *The Lancet 381*(9875), 1371–1379.

Consortium, T. A. S. D. W. G. of T. P. G (2017). Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24. 32 and a significant overlap with schizophrenia. *Molecular Autism 8*, 1–17.

Consortium*, P (2018). Revealing the brain's molecular architecture. *Science (Vol. 362, Issue 6420, pp. 1262–1263)*. American Association for the Advancement of Science.

Dara, S., Dhamercherla, S., Jadav, S. S., Babu, C., & Ahsan, M. J. (2022). Machine learning in drug discovery: A review. *Artificial Intelligence Review 55*(3), 1947–1999.

Davey Smith, G., & Hemani, G. (2014). Mendelian randomization: Genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics 23*(R1), R89–R98.

Davies, N. M., Holmes, M. V., & Smith, G. D. (2018). Reading Mendelian randomisation studies: A guide, glossary, and checklist for clinicians. *Bmj 362*.

Demontis, D., Walters, G. B., Athanasiadis, G., Walters, R., Therrien, K., Nielsen, T. T., ... Zeng, B., et al. (2023). Genome-wide analyses of ADHD identify 27 risk loci, refine the genetic architecture and implicate several cognitive domains. *Nature Genetics*, 1–11.

Demontis, D., Walters, R. K., Martin, J., Mattheisen, M., Als, T. D., Agerbo, E., ... Bækvad-Hansen, M., et al. (2019). Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nature Genetics 51*(1), 63–75.

Ebrahim, S., & Davey Smith, G. (2008). Mendelian randomization: Can genetic epidemiology help redress the failures of observational epidemiology? *Human Genetics 123*, 15–33.

Ekins, S., Gerlach, J., Zorn, K. M., Antonio, B. M., Lin, Z., & Gerlach, A. (2019). Repurposing approved drugs as inhibitors of Kv7. 1 and Nav1. 8 to treat Pitt Hopkins syndrome. *Pharmaceutical Research 36*(9), 1–10.

Evans, D. M., & Davey Smith, G. (2015). Mendelian randomization: New applications in the coming age of hypothesis-free causality. *Annual Review of Genomics and Human Genetics 16*, 327–350.

Faraone, S. V., & Larsson, H. (2019). Genetics of attention deficit hyperactivity disorder. *Molecular Psychiatry 24*(4), 562–575. https://doi.org/10.1038/s41380-018-0070-0.

Grove, J., Ripke, S., Als, T. D., Mattheisen, M., Walters, R. K., Won, H., ... Anney, R., et al. (2019). Identification of common genetic risk variants for autism spectrum disorder. *Nature Genetics 51*(3), 431–444.

Hegvik, T. -A., Waløen, K., Pandey, S. K., Faraone, S. V., Haavik, J., & Zayats, T. (2021). Druggable genome in attention deficit/hyperactivity disorder and its co-morbid conditions. New avenues for treatment. *Molecular Psychiatry 26*(8), 4004–4015.

Heyne, H. O., Baez-Nieto, D., Iqbal, S., Palmer, D. S., Brunklaus, A., May, P., ... Lemke, J. R., et al. (2020). Predicting functional effects of missense variants in voltage-gated sodium and calcium channels. *Science Translational Medicine 12*(556), Article eaay6848.

Hormozdiari, F., Van De Bunt, M., Segre, A. V., Li, X., Joo, J. W. J., Bilow, M., ... Eskin, E. (2016). Colocalization of GWAS and eQTL signals detects target genes. *The American Journal of Human Genetics 99*(6), 1245–1260.

Hsu, K. -C., & Wang, F. -S. (2017). Model-based optimization approaches for precision medicine: A case study in presynaptic dopamine overactivity. *PLoS One 12*(6), Article e0179575.

Jagadeesh, K. A., Dey, K. K., Montoro, D. T., Mohan, R., Gazal, S., Engreitz, J. M., ... Regev, A. (2022). Identifying disease-critical cell types and cellular processes by integrating single-cell RNA-sequencing and human genetics. *Nature Genetics 54*(10), 1479–1492. https://doi.org/10.1038/s41588-022-01187-9.

Jiang, T. (2013). Brainnetome: A new-ome to understand the brain and its disorders. *Neuroimage 80*, 263–272.

Jourdon, A., Scuderi, S., Capauto, D., Abyzov, A., & Vaccarino, F. M. (2021). PsychENCODE and beyond: Transcriptomics and epigenomics of brain development and organoids. *Neuropsychopharmacology 46*(1), 70–85.

Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD) 6*(4), 1–21.

Kessler, R. C., & Wang, P. S. (2008). The descriptive epidemiology of commonly occurring mental disorders in the United States. *Annual Review of Public Health 29*(1), 115–129.

King, E. A., Davis, J. W., & Degner, J. F. (2019). Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS Genetics 15*(12), Article e1008489.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques.* MIT press.

Kuenzi, B. M., Park, J., Fong, S. H., Sanchez, K. S., Lee, J., Kreisberg, J. F., ... Ideker, T. (2020). Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell 38*(5), 672–684.

Lal, D., May, P., Perez-Palma, E., Samocha, K. E., Kosmicki, J. A., Robinson, E. B., ... Weckhuysen, S., et al. (2020). Gene family information facilitates variant interpretation and identification of disease-associated genes in neurodevelopmental disorders. *Genome Medicine 12*(1), 1–12.

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., ... Ross, K. N., et al. (2006). The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science 313*(5795), 1929–1935.

Lecca, P. (2021). Machine learning for causal inference in biological networks: Perspectives of this challenge. *Frontiers in Bioinformatics 1*, 746712. https://doi.org/10.3389/fbinf.2021.746712.

Lee, P. H., Anttila, V., Won, H., Feng, Y. -C. A., Rosenthal, J., Zhu, Z., ... Posthuma, D., et al. (2019). Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. *Cell 179*(7), 1469–1482.

Lempiäinen, H., Brænne, I., Michoel, T., Tragante, V., Vilne, B., Webb, T. R., ... Willenborg, C., et al. (2018). Network analysis of coronary artery disease risk genes elucidates disease mechanisms and druggable targets. *Scientific Reports 8*(1), 3434.

Liu, J., Cheng, Y., Li, M., Zhang, Z., Li, T., & Luo, X. -J. (2022). Genome-wide Mendelian randomization identifies actionable novel drug targets for psychiatric disorders. *Neuropsychopharmacology*, 1–11.

Liu, S. H., Bobb, J. F., Claus Henn, B., Gennings, C., Schnaas, L., Tellez-Rojo, M., ... Coul, B. A. (2018). Bayesian varying coefficient kernel machine regression to assess neurodevelopmental trajectories associated with exposure to complex mixtures. *Statistics in medicine 37*(30), 4680–4694.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The genotype-tissue expression (GTEx) project. *Nature Genetics 45*(6), 580–585.

Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., Peng, J., Chen, L., & Zeng, J. (2017). A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature Communications 8*(1), 573.

Marunnan, S. M., Pulikkal, B. P., Jabamalairaj, A., Bandaru, S., Yadav, M., Nayarisseri, A., & Doss, V. A. (2017). Development of MLR and SVM aided QSAR models to identify common SAR of GABA uptake herbal inhibitors used in the treatment of schizophrenia. *Current Neuropharmacology 15*(8), 1085–1092.

Mullins, N., Forstner, A. J., O'Connell, K. S., Coombes, B., Coleman, J. R., Qiao, Z., ... Bryois, J., et al. (2021). Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. *Nature Genetics 53*(6), 817–829.

Nelson, M. R., Tipney, H., Painter, J. L., Shen, J., Nicoletti, P., Shen, Y., ... Wang, J., et al. (2015). The support of human genetic evidence for approved drug indications. *Nature Genetics 47*(8), 856–860.

Nguyen, N. D., Jin, T., & Wang, D. (2021). Varmole: A biologically drop-connect deep neural network model for prioritizing disease risk variants and genes. *Bioinformatics 37*(12), 1772–1775.

Obara, T., Ishikuro, M., Tamiya, G., Ueki, M., Yamanaka, C., Mizuno, S., Kikuya, M., Metoki, H., Matsubara, H., Nagai, M., et al. (2018). Potential identification of vitamin B6 responsiveness in autism spectrum disorder utilizing phenotype variables and machine learning methods. *Scientific Reports 8*(1), 1–7.

Pan, X., Yun, J., Akdemir, Z. H. C., Jiang, X., Sahni, N., Wu, E., ... Yi, S. S. (2023). AI-DrugNet: A network-based deep learning model for drug repurposing and combination therapy in neurological disorders. *Computational and Structural Biotechnology Journal 21*, 1533–1542.

Pardiñas, A. F., Holmans, P., Pocklington, A. J., Escott-Price, V., Ripke, S., Carrera, N., ... Hamshere, M. L., et al. (2018). Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nature Genetics 50*(3), 381–389.

Pearl, J. (2009). *Causality: Models, reasoning and inference* (2nd ed.). Cambridge University Press.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. The MIT Press.

Radonjić, N. V., Hess, J. L., Rovira, P., Andreassen, O., Buitelaar, J. K., Ching, C. R., ... McDonald, C., et al. (2021). Structural brain imaging studies offer clues about the effects of the shared genetic etiology among neuropsychiatric disorders. *Molecular Psychiatry 26*(6), 2101–2110.

Reay, W. R., & Cairns, M. J. (2021). Advancing the use of genome-wide association studies for drug repurposing. *Nature Reviews Genetics 22*(10), 658–671.

Sandin, S., Lichtenstein, P., Kuja-Halkola, R., Hultman, C., Larsson, H., & Reichenberg, A. (2017). The heritability of autism spectrum disorder. *JAMA 318*(12), 1182. https://doi.org/10.1001/jama.2017.12141.

Sanislow, C. A., Ferrante, M., Pacheco, J., Rudorfer, M. V., & Morris, S. E. (2019). Advancing translational research using NIMH research domain criteria and computational methods. *Neuron 101*(5), 779–782.

Satterstrom, F. K., Walters, R. K., Singh, T., Wigdor, E. M., Lescai, F., Demontis, D., ... Bybjerg-Grauholm, J., et al. (2019). Autism spectrum disorder and attention deficit hyperactivity disorder have a similar burden of rare protein-truncating variants. *Nature Neuroscience 22*(12), 1961–1965.

Schadt, E. E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P. Y., ... Ulrich, R. (2008). Mapping the genetic architecture of gene expression in human liver. *PLoS Biology 6*, Article e107.

Schubert, C. R., O'Donnell, P., Quan, J., Wendland, J. R., Xi, H. S., Winslow, A. R., ... Airey, D. C., et al. (2015). BrainSeq: Neurogenomics to drive novel target discovery for neuropsychiatric disorders. *Neuron 88*(6), 1078–1083.

Shen, E. H., Overly, C. C., & Jones, A. R. (2012). The Allen human brain atlas: Comprehensive gene expression mapping of the human brain. *Trends in Neurosciences 35*(12), 711–714.

Shendure, J., Findlay, G. M., & Snyder, M. W. (2019). Genomic medicine–Progress, pitfalls, and promise. *Cell 177*(1), 45–57. https://doi.org/10.1016/j.cell.2019.02.003.

Smoller, J. W. (2019). Psychiatric genetics begins to find its footing. *American Journal of Psychiatry 176*(8), 609–614.

de Souza, N. (2012). The ENCODE project. *Nature Methods 9*(11), 1046.

Steel, Z., Marnane, C., Iranpour, C., Chey, T., Jackson, J. W., Patel, V., & Silove, D. (2014). The global prevalence of common mental disorders: A systematic review and meta-analysis 1980–2013. *International Journal of Epidemiology 43*(2), 476–493.

Sullivan, P. F., Agrawal, A., Bulik, C. M., Andreassen, O. A., Børglum, A. D., Breen, G., ... Gelernter, J., et al. (2018). Psychiatric genomics: An update and an agenda. *American Journal of Psychiatry 175*(1), 15–27.

Talukdar, H., Foroughi Asl, H., Jain, R., Ermel, R., Ruusalepp, A., Franzén, O., ... Björkegren, J. L. M. (2016). Cross-tissue regulatory gene networks in coronary artery disease. *Cell Systems 2*, 196–208.

Trubetskoy, V., Pardiñas, A. F., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T. B., ... Hall, L. S., et al. (2022). Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature 604*(7906), 502–508.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., et al. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery 18*(6), 463–477.

Vos, T., Abajobir, A. A., Abate, K. H., Abbafati, C., Abbas, K. M., Abd-Allah, F., ... Abera, S. F., et al. (2017). Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: A systematic analysis for the global burden of disease study 2016. *The Lancet 390*(10100), 1211–1259.

Wallace, C. (2021). A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS Genetics 17*(9), Article e1009440.

Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F. C., ... Yang, Y. T., et al. (2018). Comprehensive functional genomic resource and integrative model for the human brain. *Science 362*(6420), Article eaat8464.

Wang, L., Audenaert, P., & Michoel, T. (2019). High-dimensional Bayesian network inference from systems genetics data using genetic node ordering. *Frontiers in Genetics 10*, 1196. https://doi.org/10.3389/fgene.2019.01196.

Wingo, T. S., Liu, Y., Gerasimov, E. S., Gockley, J., Logsdon, B. A., Duong, D. M., ... Ressler, K. J., et al. (2021). Brain proteome-wide association study implicates novel proteins in depression pathogenesis. *Nature Neuroscience 24*(6), 810–817.

Wu, T. -N., Chen, C. -K., Lee, C. -S., Wu, B. -J., Sun, H. -J., Chang, C. -H., ... Cheng, A. T. -A. (2019). Lithium and GADL1 regulate glycogen synthase kinase-3 activity to modulate KCTD12 expression. *Scientific Reports 9*(1), 1–10.

Wu, Y., Byrne, E. M., Zheng, Z., Kemper, K. E., Yengo, L., Mallett, A. J., ... Wray, N. R. (2019). Genome-wide association study of medication-use and associated disease in the UK biobank. *Nature Communications 10*(1), 1–10.

Yang, Q. -X., Wang, Y. -X., Li, F. -C., Zhang, S., Luo, Y. -C., Li, Y., ... Xue, W. -W., et al. (2019). Identification of the gene signature reflecting schizophrenia's etiology by constructing artificial intelligence-based method of enhanced reproducibility. *CNS Neuroscience & Therapeutics 25*(9), 1054–1063.

Zeng, X., Zhu, S., Liu, X., Zhou, Y., Nussinov, R., & Cheng, F. (2019). DeepDR: A network-based deep learning approach to in silico drug repositioning. *Bioinformatics 35*(24), 5191–5198.

Zhang, B., Gaiteri, C., Bodea, L. G., Wang, Z., McElwee, J., Podtelezhnikov, A. A., ... Emilsson, V. (2013). Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell 153*(3), 707–720.

Zhao, K., & So, H. -C. (2018). Drug repositioning for schizophrenia and depression/anxiety disorders: A machine learning approach leveraging expression data. *IEEE Journal of Biomedical and Health Informatics 23*(3), 1304–1315.

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., ... Visscher, P. M., et al. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics 48*(5), 481–487.