ID N. 33

CAMPUS BIO-MEDICO UNIVERSITY OF ROME

DEPARTMENT OF ENGINEERING

UNIVERSITY OF BARI ALDO MORO

DEPARTMENT OF TRANSLATIONAL BIOMEDICINE AND NEUROSCIENCE

Italian National Ph.D. in Artificial Intelligence

Health and Life Sciences

XXXVII Cycle

Beyond Local Regulation: Network-Based Prediction of Gene Expression and Its Application to Neuropsychiatric Traits

Supervisors

Prof. Giulio Pergola

Prof. Loredana Bellantuono

Prof. Donato Impedovo

Candidate

Fabiana Rossi

May, 2025

We are not alone,
in this world of travelers.

Seek the gaze of those
who share the same dream.

[Eveline's Dust,
Returning Somewhere]

Acknowledgements

It is often said that doing a PhD is one of the toughest stages in an academic career. I can certainly relate—it takes an incredible amount of effort to navigate the world of research as a junior scholar. However, this journey becomes much easier when you have the right people by your side, and I have been lucky in this regard.

First and foremost, I want to thank my supervisor, Prof. Giulio Pergola, for always being available when I needed guidance, for his insightful ideas, and for his practical support throughout this journey.

Special thanks go to the entire Gruppo di Neuroscienze Psichiatriche (GNP), led by Prof. Alessandro Bertolino, whose members have been a valuable source of scientific inspiration and support. Particular mentions go to Leonardo Sportelli, Gianluca C. Kikidis, and Loredana Bellantuono, who contributed to the development of parts of this work. I am also grateful to all my colleagues for their friendly support—and for sharing countless meals, coffee breaks, and moments of encouragement along the way.

I am deeply thankful to the Lieber Institute for Brain Development (LIBD), headed by Dr. Daniel Weinberger, for welcoming me and providing an invaluable opportunity to grow scientifically in an international setting across the ocean.

I am truly grateful for a journey that allowed me to travel across a part of the world and connect with extraordinary people. To all the friends I've made along the way—in Rome, Torino, Marseille, Baltimore, and Bari—thank you: this achievement would not have been possible without your constant encouragement and support.

Last but not least, this thesis is entirely dedicated to my parents, Sandra and Vito, who gave me the opportunity to pursue education and acquire knowledge in a society where access to learning is not always guaranteed.

Contents

Abstract

Understanding how genetic variation shapes brain-related traits with complex heritability requires models capable of capturing the multifaceted regulatory architecture of gene expression. Non-coding variants are thought to influence disease risk primarily by modulating transcriptomic regulation rather than through direct coding changes. Gene-level prediction models, which impute expression from genetic data, offer a framework for linking genetic variation to transcriptional consequences; however, current approaches predominantly rely on *cis*-eQTLs—capturing primarily local regulatory effects—thus limiting their ability to model the distributed architecture of gene regulatory networks.

To address these limitations, the present work develops an integrative framework that incorporates gene co-expression network information into genetically regulated expression (GReX) models. By leveraging the structure of transcriptional networks, the approach guides the detection of *trans*-regulatory (distal) effects and improves transcriptomic prediction.

In the first study, I implemented this framework through the development of two complementary algorithms, INGENE and MODULE, which integrate co-expression module structure into *trans*-eQTL selection and dimensionality reduction for gene-level prediction. Using RNA-seq and genotype data from postmortem brain cohorts (LIBD, CMC, and GTEx), I trained and validated these models across six brain regions (amygdala, caudate nucleus, dorsal/subgenual anterior cingulate cortex, dorsolateral prefrontal cortex, hippocampus). Benchmarking against both an original *cis*-based model and EpiXcan—the leading benchmark for *cis*-model performance on our training dataset—demonstrated that the integration of *cis*- and *trans*-predictions significantly improves gene coverage and predictive accuracy across independent datasets.

In the second study, I applied the co-expression-informed prediction models to large-scale schizophrenia (SCZ) cohorts from the Psychiatric Genomics Consortium wave 3 (PGC3) to evaluate their utility in gene-trait association discovery. By imputing gene expression across brain regions and performing association testing, I identified 1,764 SCZ-associated genes across regions (pFDR $< .01$), including 1,515 novel associations not captured by *cis*-only approaches.

In the third study, I evaluated the generalizability and boundaries of the network-informed prediction framework by applying integrated *cis*- and *trans*-regulatory models to impulsivity, a behavioural proxy for antisocial tendencies, in a high-risk forensic cohort. Using genetically imputed brain expression scores and machine learning models, I evaluated the capacity of GReX predictors to account for individual differences in impulsivity. Despite incorporating *trans*-regulatory information, predictive accuracy remained limited, consistent with the modest heritability and substantial environmental modulation characteristic of the trait.

Together, these studies demonstrate the utility of integrating co-expression network information into GReX models for improving transcriptomic prediction and gene association in brain-related traits. The methodological framework developed in this thesis offers a flexible foundation that can be adapted to diverse datasets, genetic ancestries, and transcriptomic resources, enabling future efforts to more fully capture the regulatory mechanisms linking genetic variation to phenotypes. At the same time, the findings delineate the current boundaries of transcriptomic prediction, particularly for behaviourally complex and environmentally modulated traits, underscoring the need for multimodal approaches that integrate genetic, developmental, and environmental data.

Acronyms

BIS-11 Barratt Impulsiveness Scale

coTWAS Co-expression TWAS

eQTL expression Quantitative Trait Loci

GReX Genetically Regulated Gene Expression

GWAS Genome-Wide Association Study

INGENE Imputed Network Gene Expression *trans*-eQTLs

MODULE MODule qUantitative trait Loci Eigengene

PGC Psychiatric Genomics Consortium

PRS Polygenic Risk Scores

SCZ Schizophrenia

TWAS Transcriptome-Wide Association Study

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

In this thesis, I develop and apply predictive models of gene expression grounded in functional genomics to investigate the genetic architecture of complex traits. By integrating expression quantitative trait loci (eQTL) with transcriptomic datasets and gene co-expression network resources, I aim to improve genotype-based prediction of gene expression. A central focus of this work is on modeling distal (*trans*) regulatory effects—long-range genetic influences that extend beyond local genomic proximity and are often overlooked by conventional *cis*-eQTL–based approaches.

Rather than focusing solely on predictive performance, this work emphasizes mechanistic interpretability and biological coherence, achieved through cross-modal integration of genomic and transcriptomic data. The proposed models are evaluated across two distinct contexts: transcriptome-wide association analyses (TWAS) in schizophrenia (SCZ)—a prototypical polygenic disorder hypothesized to involve widespread regulatory disruption—and individual-level prediction of antisocial behaviour, a behaviourally complex and environmentally sensitive trait with less prominent heritability. Together, these case studies illustrate the versatility and boundaries of co-expression–informed GReX modeling across both association- and prediction-focused applications, and highlight its potential to bridge the gap between statistical genetics and precision medicine. A more detailed overview of the research objectives is provided in Chapter 2.

1.1 Genetic Influence in Human Disease

Diseases with complex heritability—such as several major psychiatric conditions—pose significant challenges in modern genetics. Unlike monogenic disorders, which result from mutations in single genes and follow Mendelian inheritance patterns, these conditions arise from the cumulative effects of widespread genetic variation observed across individuals in a population. In most complex traits, genetic risk is spread across the genome, with numerous variants each exerting small effects—a hallmark of polygenic architecture. This complexity is further compounded by gene–gene interactions, gene–environment interplay, and epigenetic regulation, all of which contribute to the dynamic nature of trait expression and disease susceptibility (Feinberg and Fallin, 2015; Cavalli and Heard, 2019).

Twin studies have been pivotal in demonstrating the substantial role of genetic factors in shaping inter-individual differences. By comparing concordance rates between monozygotic twins, who share nearly identical genomes, and dizygotic twins, who share on average 50% of their genetic material, researchers have consistently estimated high levels of heritability (Polderman et al., 2015). For example, psychiatric disorders such as Schizophrenia (SCZ) show heritability estimates ranging from 40% to 80% (Sullivan et al., 2003; Purcell et al., 2009). However, while twin studies quantify the proportion of phenotypic variance attributable to genetic factors, they do not reveal which specific genetic variants or mechanisms are involved. Addressing this gap has required direct investigation of genomic data at the population level.

Genetic variation across individuals provides the substrate for such investigations. Genetic variation includes single nucleotide polymorphisms (SNPs), insertions and deletions, and structural variants. A major focus has been on common variants, typically defined as alleles with a minor allele frequency (MAF) greater than 1% in the population. Because they are widespread and often evolutionarily older, common variants are well-powered for discovery in large cohorts and have formed the basis of most genetic association studies. Although individual common variants tend to have small effects, their cumulative impact—across thou-

sands of loci—can significantly influence disease susceptibility. Importantly, they do not act in a deterministic way: the same variant may be present in both healthy and affected individuals. This observation aligns with the threshold-liability model (Gottesman and Shields, 1967), which posits that a categorical outcome (such as disease presence) arises from a normally distributed liability-a continuous composite of genetic and environmental risk factors. Once this liability surpasses a certain threshold, the phenotype manifests.

These insights have shifted the field from simple Mendelian models to a nuanced understanding of polygenic inheritance in which numerous loci each contribute a small amount to overall risk. The need to map these variants and interpret their biological consequences has driven the development of high-resolution genomic technologies—most notably DNA sequencing and Genome-Wide Association Study (GWAS)—which allow systematic analysis of common genetic variation and its relationship to human traits.

1.2 From Linkage Studies to GWAS

The completion of the Human Genome Project in 2003 (Collins et al., 2003) marked a pivotal milestone in genetic research, delivering the first complete reference sequence of the human genome and laying the groundwork for systematic studies of genetic variation (International Human Genome Sequencing, 2004). Building on this foundation, large-scale initiatives such as the HapMap Project (Altshuler et al., 2005) and the 1000 Genomes Project (Auton et al., 2015a) cataloged common genetic variants across diverse populations, establishing a critical resource for GWAS.

GWAS rapidly became a transformative tool in human genetics, guiding the identification of thousands of genetic loci associated with disease risk. Unlike traditional linkage analyses, which were limited to familial data, GWAS provided the statistical power to scan the genome at high density across large cohorts of unrelated individuals. This approach allowed for the unbiased interrogation of millions of variants simultaneously, revealing that much of

the heritable signal for phenotypic traits and diseases lies in common non-coding variants, particularly SNPs (Purcell et al., 2009). GWAS significantly accelerated our understanding of genetic architectures and facilitated the identification of potential biological pathways underlying major conditions (Visscher et al., 2017; Uffelmann et al., 2021).

In the field of psychiatric genetics, GWAS has been particularly impactful. Landmark studies by the Psychiatric Genomics Consortium (PGC) have demonstrated the polygenic nature of disorders. For SCZ in particular, the latest GWAS has identified 287 genome-wide significant loci, implicating genes involved in synaptic transmission, calcium signaling, immune response, and neurodevelopment (Trubetskoy et al., 2022). These findings have contributed to a nuanced, system-level understanding of psychiatric disease and have provided foundational targets for transcriptomic follow-up, functional annotation, and drug discovery.

Limitations. Despite their impact, GWAS methodologies face significant limitations. Most notably, GWAS consist of mass univariate statistics, testing each SNP independently for association with a trait. This single-variant framework limits their ability to capture the polygenic complexity of human diseases, particularly where risk is distributed across multiple interacting loci or mediated through subtle regulatory effects.

Interpretation of GWAS findings is further complicated by pleiotropy——where a single genetic locus influences multiple traits. Large-scale analyses have shown that more than 90% of trait-associated loci are pleiotropic (Watanabe et al., 2019), particularly in domains like psychiatry, where disorders such as SCZ, bipolar disorder (BP), and major depressive disorder (MDD) share overlapping genetic architectures (Purcell et al., 2009; of the Psychiatric Genomics Consortium, 2019). While this suggests shared biological underpinnings, pleiotropy at the SNP level often lacks mechanistic resolution. Associations may arise from linkage disequilibrium (LD) with different causal variants, or reflect distinct pathways within the same locus acting on separate traits. Moreover, SNP-level associations can be confounded by ancestry-specific effects or sample-specific biases (Ding et al., 2023). These complexities

limit the biological interpretability of pleiotropy inferred from GWAS and highlight the need for integrative approaches that move beyond isolated variant-trait links to uncover shared regulatory mechanisms.

Furthermore, most GWAS-identified variants have small effect sizes and reside in non-coding regions, where biological interpretation is challenging (Maurano et al., 2012; Edwards et al., 2013). Moreover, GWAS are largely designed around the "*common disease–common variant*" hypothesis, which posits that common disorders arise from the additive effects of multiple common alleles—typically defined as variants with a MAF greater than 5%—each contributing a small increase in risk. While this framework has proven useful for discovering broadly replicable loci, it also imposes important blind spots: GWAS tend to overlook disease heterogeneity, rare variants, and context-specific effects, limiting their utility for understanding individual-level risk or mechanistic heterogeneity (Boyle et al., 2017; Tam et al., 2019; Woodward et al., 2022; Gurdasani et al., 2019). On average, the effect size for candidate SNPs identified in GWAS is around 1.33, underscoring the small and often scattered contributions of each variant (Hindorff et al., 2009).

The "*missing heritability*" problem remains a persistent challenge in complex trait genetics. Although increasing GWAS sample sizes has improved the ability to detect genome-wide significant associations, the incremental gain in explained phenotypic variance has been limited. In the case of SCZ, for instance, early large-scale studies successfully identified dozens of associated loci (Allen et al., 2008; Ripke et al., 2014); however, more recent efforts involving substantially larger cohorts have yielded only modest improvements in the proportion of variance explained (Trubetskoy et al., 2022). This pattern of diminishing returns suggests that simply scaling up GWAS may not be sufficient to fully capture the genetic basis of such disorders.

To address these limitations, Polygenic Risk Scoress (PRSs) have been developed as a way to summarize the cumulative effect of many common variants across the genome (Chatterjee et al., 2016). By aggregating individually weak signals into a single composite score,

PRS aims to enhance predictive power for disease risk stratification. This approach has shown promise in several conditions with well-characterized genetic architectures (Inouye et al., 2018), and continues to improve with methodological refinements and larger reference datasets. However, in highly polygenic and clinically heterogeneous disorders such as SCZ and MDD, current PRS explains only a modest portion of disease liability (Wray et al., 2018; Howard et al., 2019), and their predictive accuracy at the individual level remains limited (Ripke et al., 2014; Sullivan and Geschwind, 2019). Importantly, genetic risk does not act in isolation; it is often intertwined with environmental exposures and socio-demographic factors (Caspi and Moffitt, 2006), suggesting that future models integrating both genetic and environmental data may offer greater utility for understanding and mitigating psychiatric disease risk.

In summary, while GWAS and PRS have been instrumental in identifying statistical associations between genetic variants and disease susceptibility, they often fall short of explaining the biological mechanisms through which these variants exert their effects. This limitation has fueled growing interest in functional genomics approaches, which aim to link genetic variation to intermediate molecular phenotypes—such as gene expression, chromatin accessibility, or protein abundance—that more directly reflect cellular function (Tam et al., 2019; van der Sijde et al., 2014). In this context, expression quantitative trait loci (eQTLs) analysis has emerged as a key framework for connecting regulatory variants to gene expression levels, offering a mechanistic view of how inherited variation influences phenotype at the transcriptomic level.

1.3 Bridging Genotype and Phenotype

As gene expression is intermediate between the DNA sequence and phenotype, mRNA can be considered the proximal functional readout of non-coding genetic variants affecting disease susceptibility (Mostafavi et al., 2023). In psychiatric genetics, integrating GWAS findings

with gene expression profiles from postmortem human brain tissue has proven pivotal for elucidating the functional consequences of risk loci (Gandal et al., 2016). While DNA defines the blueprint of biological potential, it is the spatiotemporal regulation of gene expression that determines how this potential unfolds across cell types, developmental stages, and brain regions. Thus, understanding how genetic variants influence mRNA levels is crucial for studying mechanisms of disease.

1.3.1 How non-coding variants confer susceptibility to diseases

To disentangle how genetic variation contributes to susceptibility, it is essential to understand the mechanisms by which genetic variants influence disease risk and how these variants can be identified.

Genetic variation affects disease susceptibility primarily in two ways: either by altering the protein structure directly (e.g., amino-acid substitution) or by influencing gene expression (e.g., transcription or translation efficiency) (Figure 1). Because the vast majority of disease-associated SNPs are located in non-coding regions of the genome, such as introns, UTRs, and intergenic areas, it is unlikely that their effects are simply due to LD with nearby coding variants. Instead, these non-coding SNPs are thought to directly influence gene regulation, most often by modulating gene expression. Such regulatory effects represent a complex and subtle manifestation of genetic predisposition and may be exerted through various molecular mechanisms. For example, promoter variants can markedly impact the transcriptional activity of a gene by altering transcription factor binding sites, modifying chromatin accessibility, or disrupting core promoter elements that recruit the transcriptional machinery (Greenwood and Kelsoe, 2003; Lemonde et al., 2003). Intronic SNPs can also affect transcription or alter mRNA splicing or stability, i.e. resistence to degradation, and thus the relative abundance and proportions of isoforms (Greenwood and Kelsoe, 2003; Tokuhiro et al., 2003); SNPs in the 3′ UTRs may alter mRNA stability and translation (Mill et al., 2002; Miller and Madras, 2002). Even synonymous exonic SNPs, usually non-functional from the translational point of

view, can influence mRNA structure and translation (Shen et al., 1999; Duan et al., 2003b). Finally, the effects of individual SNPs cannot be interpreted in isolation, as their functional impact is often influenced by the haplotype background (Duan et al., 2003a)—that is, the combination of alleles at adjacent loci that are inherited together on the same chromosome. These linked variants can interact functionally or co-regulate gene activity, implying that the effect of an individual SNP may depend on its broader genomic context. Furthermore, changes in the expression of disease-relevant genes can result from variants that are not themselves directly associated with the trait. Many such genes encode proteins that act as key nodes within molecular signaling networks, making them susceptible to regulation through multiple upstream inputs. As a result, genetic variation in other components of the same pathway can lead to compensatory or downstream changes in expression, further complicating the identification of causal variants in genetic studies.

1.3.2 Expression Quantitative Trait Loci

Transcriptomics provides a comprehensive approach to profiling RNA transcripts generated by the genome under specific cellular or tissue contexts. Advances in high-throughput methodologies such as RNA sequencing have substantially enhanced the capability to measure transcript abundance on a genome-wide scale across diverse biological conditions. These advancements have significantly contributed to the study of how genetic variation influences different patterns of gene expression among individuals.

A critical tool in deciphering these genetic regulatory mechanisms is the study of eQTLs, which tests associations between SNPs and variations in gene expression. In traditional eQTL mapping, individual SNPs are independently tested for their associations with the expression levels of nearby or distant genes within a population (Gilad et al., 2008; Stranger et al., 2007). Crucially, eQTL analysis relies on matched genotype and transcriptomic datasets and is usually conducted in a tissue- or cell-type-specific manner, highlighting the context-dependent nature of gene regulation (GTEx, 2017, 2020).

One significant advantage of eQTL studies is their applicability to tissues that are challenging or impossible to directly profile, such as human brain tissue or tissues at inaccessible developmental stages. In such scenarios, where direct transcriptomic measurements from living individuals are not feasible, eQTLs allow researchers to indirectly infer the regulatory impact of genetic variation from genotyping alone (Nicolae et al., 2010; Fromer et al., 2016). By capturing the heritable components of gene regulation, eQTL analyses offer a complementary approach to GWAS for interpreting non-coding genetic variation in the absence of direct transcriptomic evidence.

eQTL classification. eQTLs can be broadly divided into two categories (Figure 1): *cis*-eQTLs and *trans*-eQTLs. *Cis*-eQTLs are genetic variants located near the genes they regulate, typically within 1 Mb of the transcription start site. These variants often influence gene expression through local regulatory elements such as enhancers and promoters and have received significant attention in large-scale studies due to their relatively strong and readily detectable effects (Liu et al., 2019; Yao et al., 2020). However, *cis*-eQTLs only partially account for gene expression heritability, indicating that distal regulatory mechanisms also play important roles (Umans et al., 2021).

In contrast, *trans*-eQTLs influence the expression of genes located at distant loci, often on different chromosomes, typically through intermediaries such as transcription factors, chromatin remodelers, or noncoding RNAs (Battle et al., 2014; Liu et al., 2022; Pierce et al., 2014). Although these long-range regulatory interactions are critical for elucidating gene regulatory networks, they are considerably more challenging to detect due to their smaller effect sizes—often an order of magnitude weaker than those of *cis*-eQTLs—and their broader genomic dispersion (Liu et al., 2019). For instance, while over 90% of expressed genes exhibit at least one significant *cis*-eQTL, fewer than 10% have reliably detectable *trans*-eQTLs at typical sample sizes (Liu et al., 2019). Consequently, accurate identification of *trans*-eQTLs requires substantially larger cohorts, stringent multiple testing correction, and integrative

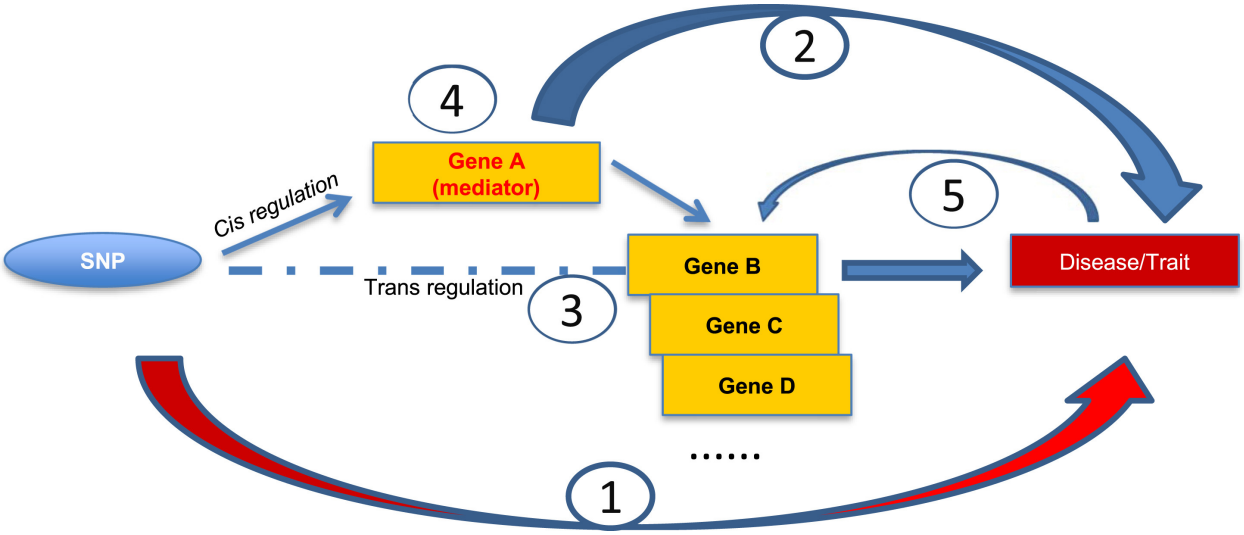

Figure 1: **Mediation Mechanisms of eQTLs**. Genetic variants can affect traits through the following mechanisms: (1) missense SNP affects protein structure/function; (2) non-coding SNP affects gene expression (*cis*); (3) non-coding SNP affects remote (*trans*) gene expression directly or by (4) *cis*-eGene mediation of the *trans*-eQTL-*trans*-eGene association; or (5) reverse causality (trait has feedback effect on gene expression). Figure from Yao et al. (2017).

modeling strategies. While *cis*-eQTLs have proven valuable for linking genetic variation to gene regulation, they account for only a fraction of expression variance. *Trans*-eQTLs, though individually weaker, may collectively explain a large component of gene expression variability and contribute meaningfully to complex trait heritability.

Tissue and Cellular Context of eQTLs. eQTL studies also indicate that the regulatory impact of genetic variation is not static but is fundamentally shaped by tissue- and cell-specific factors (GTEx, 2017, 2015, 2020). The Genotype-Tissue Expression (GTEx) Project has been designed to advance our understanding of tissue-specific genetic regulation of gene expression. Their foundational work mapped genetic effects across 49 human tissues, revealing thousands of *cis*-eQTLs and demonstrating the extensive tissue-specificity of regulatory variation (GTEx, 2017). These studies demonstrate how even subtle genetic variation can affect gene expression in tissue-specific ways, forming a critical resource for interpreting the functional implications of such variants.

Both *cis*- and *trans*-eQTL discovery strongly depend on tissue and cell type. While *cis*-eQTLs are generally robust and more consistently detected across diverse tissues, they predominantly capture local regulatory effects that often lack fine-grained specificity for particular biological contexts (GTEx, 2020). In contrast, *trans*-eQTLs are typically much more context-dependent, with effects that can vary dramatically across tissues, developmental stages, and cellular states. Because *trans*-eQTLs often act through intermediaries such as transcription factors or signaling pathways, they are especially sensitive to cellular type and physiological conditions. As a result, while *cis*-eQTLs provide important baseline information about genetic regulation, they may miss the dynamic, higher-order regulatory interactions that shape tissue- and cell-specific phenotypes. Recent single-cell studies further stress this point, revealing that many eQTL effects are masked in bulk tissue analyses and become detectable only when examined at the level of specific cell types (van der Wijst et al., 2018).

Despite eQTL advances, significant challenges persist in translating findings into clear biological insights, especially concerning diseases with complex heritability. A large proportion of GWAS-identified risk variants reside in non-coding regions without clear overlap with known eQTLs (Chun et al., 2017; Umans et al., 2021; Mostafavi et al., 2023). This gap is likely attributable to incomplete representation of relevant tissues, cell types, and developmental stages in current datasets, as well as the difficulty of detecting rare or *trans*-acting regulatory variants. Therefore, *trans*-eQTL models should capture subtle, context-dependent regulatory effects comprehensively.

Genetically Regulated Gene Expression. To overcome the limitations of traditional eQTL approaches—particularly their fragmented treatment of regulatory variation and limited biological interpretability—gene-level predictive models have been developed to estimate genetically regulated gene expression (GReX) directly from genotype data (Gamazon et al., 2015). These models aggregate the effects of multiple *cis*-acting SNPs, weighting each by its

contribution to expression variance, to generate a single predictive score per gene. By capturing weak, distributed regulatory signals that may not achieve genome-wide significance individually (Zhang et al., 2019; Huckins et al., 2019), they enhance statistical power, reduce environmental and technical noise, and improve signal-to-noise ratios in downstream analyses. Crucially, their gene-centric output facilitates biological interpretation and functional prioritization of disease-associated loci, providing a more coherent framework for linking genetic variation to molecular phenotypes.

Among these methodologies, the **Predixcan** framework (Gamazon et al., 2015) is a foundational innovation. It uses regularized regression models (e.g., elastic net) trained on large-scale eQTL reference datasets to impute GReX from genotype data alone (Figure 2A). In one benchmarking study, PrediXcan successfully imputed expression levels for **6,695 genes** using GTEx whole blood models; of these, **6,127 genes** produced valid performance metrics when evaluated in the testing cohort from the 1000 Genomes Project (Li et al., 2018a).

MetaXcan (Barbeira et al., 2018) extends the utility of these models by enabling gene-based association tests using GWAS summary statistics rather than individual-level genotype data. In doing so, it allows broader applications in large-scale consortia where only summary data are available, and integrates expression prediction models across tissues and populations to enhance statistical power and generalizability.

EpiXcan (Zhang et al., 2019) extends the PrediXcan framework by integrating epigenomic annotations—such as histone modifications, chromatin accessibility, and transcription factor binding sites—into the gene expression prediction pipeline. This multi-layered approach improves model performance, particularly for genes with weak *cis*-eQTL signals or in tissues where epigenetic regulation is more prominent. By incorporating regulatory context, EpiXcan enhances variant prioritization and facilitates the identification of functional elements within noncoding regions. In their foundational study, Zhang et al. (2019) trained models across 49 GTEx tissues, achieving significant cross-validated prediction accuracy (R^2

> 0.01) for **9,259 out of 14,961 genes** across tissues. These enhanced models enabled the discovery of novel, tissue-specific gene–trait associations that were previously undetectable using earlier approaches such as PrediXcan.

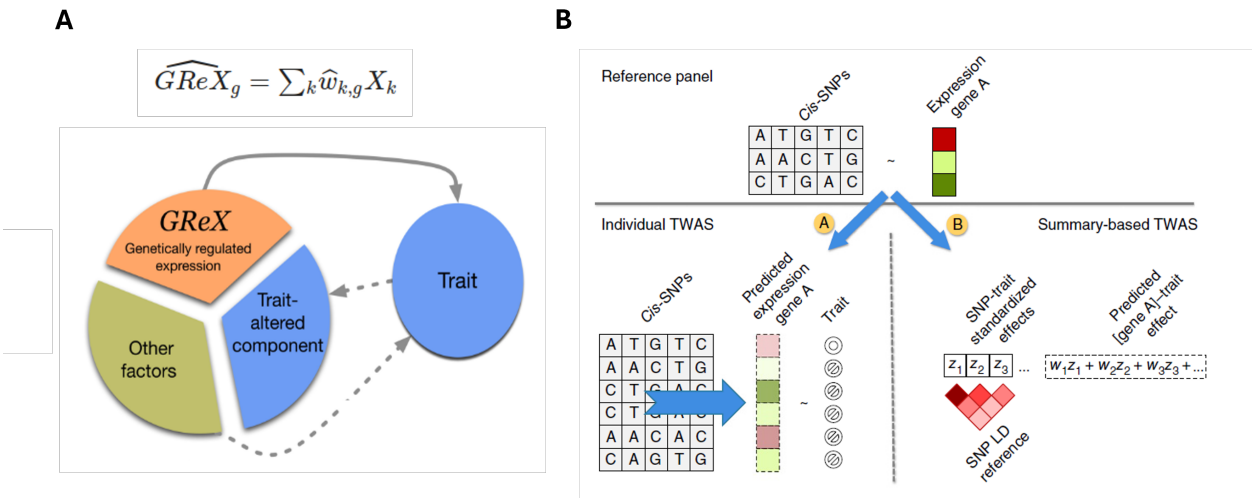

Figure 2: **Concept of GReX and overview of TWAS methodology.A)** The left panel illustrates the composition of gene expression variance, highlighting the GReX component as distinct from trait-altered expression and other non-genetic factors. **B)** The right panel outlines the general framework of TWAS, emphasizing two analytical strategies: (A) Individual-level TWAS, where gene expression levels predicted from *cis*-genotypes in reference panels are directly tested for association with traits in individual datasets; and (B) Summary-based TWAS, which utilizes precomputed SNP-trait associations information to infer gene-trait associations without requiring individual-level genotype data. Figure A from Gamazon et al. (2015); Figure B from Gusev et al. (2016).

Limitations. Despite their utility, current gene-level prediction models are constrained by a critical limitation: their reliance on local *cis*-regulatory variation. Because these models typically restrict the training to SNPs within a narrow window around each gene, they fail to capture the influence of distal *trans*-regulatory elements that contribute to shaping the transcriptional landscape. As a consequence, a substantial proportion of genes—especially those lacking strong *cis*-heritability—cannot be reliably predicted. The restricted scope of these models not only reduces the number of genes available for downstream analyses but also introduces systematic bias in phenotypic association studies, where statistical power and discovery are directly linked to the number of imputable genes and the accuracy of their

predicted expression. Even in well-characterized tissues, many genes yield low cross-validated performance or fall below predictive thresholds, limiting their inclusion in association models and complicating the interpretation of negative findings (Zhang et al., 2019; Huckins et al., 2019).

1.3.3 Genetic Association Studies via TWAS

Building on gene-level predictive models, Transcriptome-Wide Association Studies (TWAS) integrate genetically predicted gene expression with complex trait association analyses (Gusev et al., 2016). Rather than testing millions of individual SNPs as in GWAS, TWAS shift the focus to a smaller, biologically informed set of gene-level predictors, reducing the multiple-testing burden and enhancing interpretability (Figure 2B). By imputing GReX across large cohorts, TWAS enables systematic evaluation of gene-trait associations, offering mechanistic insights that are often unavailable in variant-level analyses.

However, as discussed above, the power and resolution of TWAS are fundamentally limited by the scope and accuracy of the underlying expression prediction models. Only genes with reliably imputable GReX—typically those with strong *cis*-regulatory architecture—are eligible for association testing. This dependency introduces both statistical and biological constraints: poor model performance reduces discovery power, while genes predominantly influenced by *trans*-regulatory elements or context-specific mechanisms are systematically underrepresented. Consequently, the accuracy of GReX models, in terms of performance and number of imputable genes, critically determines the extent and interpretability of associations uncovered by TWAS.

Limitations and Challenges. Despite methodological advances, TWAS face persistent limitations related to data availability, predictive accuracy, interpretability, and causal inference. The scarcity of tissue-specific eQTL datasets—particularly for critical tissues such as the brain and during key developmental stages—undermines the reliability and generaliz-

ability of TWAS findings (Li and Ritchie, 2021; Mai et al., 2023; Wainberg et al., 2019). For many genes, predictive accuracy remains low, constrained by modest eQTL effect sizes and inherent model uncertainties. Moreover, conducting TWAS across multiple tissues inflates the multiple-testing burden, requiring stringent corrections that can reduce discovery power. Classical TWAS frameworks, primarily focused on common *cis*-eQTLs, often miss rare variants and context-specific regulatory effects, leading to a systematic bias against distal and noncanonical regulatory mechanisms (Li and Ritchie, 2021; Luningham et al., 2020).

Interpreting TWAS results in complex disorders is further complicated by the intricate architecture of gene regulatory networks and the confounding effects of linkage disequilibrium, which can obscure the distinction between causal genes and correlated neighbours. In multifactorial conditions, where gene regulation is highly context-dependent and often nonlinear, conventional TWAS models—built on linear additive assumptions—risk overlooking critical *trans*-acting regulation and co-regulatory modules central to disease pathogenesis.

1.4 Network-Based Strategies: Capturing Polygenic Architecture via Gene Co-expression

As discussed in previous sections, both traditional eQTL models and TWAS-based approaches are limited in their ability to capture the full scope of gene regulation—particularly for *trans*-acting, context-specific, and polygenic mechanisms. Gene co-expression network analyses offer a complementary, system-level strategy to address these gaps. By clustering genes with correlated expression patterns across individuals, co-expression networks reveal biologically coherent modules that often reflect shared regulation, cell-type specificity, and functional pathways (Gandal et al., 2018a; Pergola et al., 2023c).

From Polygenic Risk to Co-regulated Modules. In contrast to single-gene models, co-expression approaches enable the aggregation of weak, distributed regulatory signals into

interpretable functional units. This is particularly valuable for complex traits like SCZ, where genome-wide risk loci—although scattered across the genome—often converge within specific co-regulated gene modules (Fromer et al., 2016; Gandal et al., 2018a; Hartl et al., 2021). These modules are not static: their structure and functional relevance vary across brain regions, developmental stages, and cellular environments (Panagiotakos and Pasca, 2022; Cameron et al., 2023).

For example, Hartl et al. (2021) constructed a brain-wide co-expression map from RNA-seq data across 12 regions, identifying modules enriched for neurodevelopmental and activity-dependent processes—including synaptic signaling and splicing—that overlap genetic risk for SCZ and autism (Hartl et al., 2021). In more focused analyses of the prefrontal cortex, Pergola et al. (2019) demonstrated that SCZ risk genes cluster within regulatory modules associated with clinical treatment response (Pergola et al., 2019a). Their follow-up study extended this analysis across developmental windows and brain regions, identifying a core set of 28 co-expressed SCZ-risk genes—23 of which had not been previously linked to the disorder (Pergola et al., 2023b).

Functional and Mechanistic Insights. Co-expression modules not only reveal convergence of genetic risk but also guide downstream biological validation. Fromer et al. (2016) identified SCZ-associated genes such as FURIN and SNAP91 via co-expression-informed analysis, then validated their effects on neuronal development in zebrafish and human neural progenitors (Fromer et al., 2016). Pergola et al. (2017) further demonstrated that co-expression-based polygenic scores could predict neuropsychological traits relevant to SCZ, linking molecular patterns to behavioural phenotypes (Pergola et al., 2017). More recently, Pergola et al. (2023) identified microRNA hubs and transcriptional regulators of disease-enriched modules, highlighting upstream mechanisms that modulate neurodevelopmental risk (Pergola et al., 2023c). Complementing these findings, Sportelli et al. (2024) identified a striatal dopamine-related module differentially expressed in SCZ, which not only tracked

polygenic risk scores but also predicted functional neuroimaging phenotypes.

Together, these studies highlight the power of co-expression modules to bridge genetic risk with molecular function, cellular phenotypes, and clinically relevant outcomes, making them a critical tool for mechanistic insight and translational research in psychiatric disorders.

Methodological Advantages for *trans*-eQTL Mapping Co-expression network analysis provides a critical solution to one of the most persistent challenges in human genetics: the detection of *trans*-eQTLs. Traditional genome-wide *trans*-eQTL mapping suffers from a massive multiple-testing burden, involving millions of SNP–gene combinations across the genome. This makes it severely underpowered for identifying weak and dispersed effects, which are common in complex traits. Early efforts to infer *trans*-regulatory effects—such as those by Gamazon et al. (2015) using PrediXcan—produced limited results, likely because true signals were diluted by noise in the absence of biological priors.

Co-expression analysis addresses this limitation directly. By clustering genes into modules of co-regulated expression, methods like WGCNA allow researchers to prioritize subsets of genes that are not only co-expressed, but also likely co-regulated (Fromer et al., 2016; Hartl et al., 2021). This targeted approach drastically reduces the number of tests, allowing for focused, biologically plausible hypotheses about *trans*-regulatory relationships. Instead of performing a full GWAS for every individual gene, researchers can map variants to entire modules—effectively compressing the problem from millions of tests to thousands, while increasing the interpretability of the results.

This shift from single-gene to module-based analysis not only enhances statistical power but also reflects the true architecture of gene regulation, which is modular and hierarchical. By identifying module-QTLs—variants associated with entire gene networks—co-expression frameworks expose regulatory loci that would be invisible to standard approaches. Furthermore, these networks enable the construction of polygenic expression scores that integrate weak signals across genes, tissues, and variants—making them ideally suited for modeling

polygenic risk in psychiatric and neurological conditions (Fazio et al., 2018; Borcuk et al., 2024).

In summary, co-expression is not merely a complement to *trans*-eQTL mapping—it is a prerequisite for making it tractable and biologically meaningful. In the context of complex brain disorders, where regulation is distributed and noisy, co-expression-based strategies offer a statistically coherent and mechanistically grounded alternative to conventional, genome-wide *trans*-mapping.

Chapter 2

Research Proposal

Building on the context presented in Chapter 1, this chapter outlines the research proposal of this thesis and elaborates on the conceptual motivations, methodological framework, and scientific objectives that guide the work. The proposal is situated at the intersection of statistical genetics, transcriptomics, and neuropsychiatric research, aiming to improve the predictive power of GReX models. Central to this effort is the integration of co-expression networks—data-driven maps of gene-gene relationships—into gene expression imputation models. This strategy is designed to better capture the distributed and context-dependent architecture of gene regulation, particularly in the brain.

The core motivation stems from the limitations of traditional TWAS, which primarily rely on *cis*-eQTLs to impute gene expression and test for trait associations. While effective in some contexts, these approaches often neglect the broader regulatory landscape, including *trans*-eQTLs and the coordinated behaviour of genes across co-expression modules. As a result, they may miss critical components of the genetic architecture underlying complex traits, especially those involving long-range regulatory effects and functionally interconnected pathways.

To address these challenges, this thesis proposes a novel framework that integrates co-expression networks into gene expression prediction and downstream association analyses.

By doing so, it seeks to (1) improve the detection of *trans*-regulatory effects, (2) enhance the biological relevance of genetically imputed expression scores, and (3) explore the applicability of TWAS-style analyses to phenotypes influenced by high heritability and gene-environment interactions, while acknowledging the challenges these traits pose for genetically anchored prediction models. Detailed descriptions of the selected phenotypes are provided in Section 2.2.

2.1 Thesis Objectives and Structure

This work pursues three interconnected objectives (Figure 3):

Objective 1: Develop and validate co-expression-informed *trans*-eQTL models to complement *cis*-based approaches and improve gene expression prediction across the brain transcriptomes. The first objective focuses on building integrative gene expression prediction models that incorporate co-expression network structures to better capture the distributed nature of gene regulation—particularly distal (*trans*) effects often missed by conventional *cis*-eQTL approaches. By embedding genes within biologically coherent co-expression modules, the framework models gene expression as an emergent property of regulatory networks rather than as an isolated, variant-to-gene local signal.

To achieve this objective, elastic net–based predictive models were trained using post-mortem transcriptomic and genotype data from large-scale resources including the Lieber Institute for Brain Development (LIBD), GTEx, and the CommonMind Consortium (CMC). Co-expression networks derived from prior studies guided SNP selection and dimensionality reduction, giving rise to two complementary *trans*-identification modeling strategies: "*Imputed Network Gene Expression trans-eQTLs (INGENE)*", which infers expression based on the *cis*-predicted expression of co-expressed partners, and "*MODule qUantitative trait Loci Eigengene (MODULE)*", which uses eigengene-linked *trans*-eQTLs to predict gene-level expression within modules.

These models were benchmarked against both a self-trained *cis*-based model (CIS) and the publicly available EpiXcan framework (Zhang et al., 2019) to assess improvements in transcriptome-wide coverage, predictive accuracy, and cross-cohort generalizability. By validating performance across independent datasets, this objective establishes a biologically grounded and scalable framework for improving GReX models in human brain tissue.

Objective 2: Apply prediction models to SCZ cohorts to identify novel genetic associations — SCZ as a case study. Building on the predictive models developed in Objective 1, this phase applies co-expression-informed scores to large-scale SCZ cohorts from the PGC3. SCZ serves as a powerful case study given its high heritability, extreme polygenicity, and rich catalog of transcriptomic alterations across brain regions and cell types.

The main aim is to assess whether integrating *trans*-regulatory architecture via co-expression networks improves the detection of gene-trait associations beyond conventional *cis*-based TWAS approaches. By imputing expression across diverse brain regions and applying these models to over 100,000 individuals, this objective seeks to identify novel risk genes—many of which are invisible to proximity-based or *cis*-only frameworks.

Analyses will also evaluate the functional relevance of the identified genes, including their enrichment in neurodevelopmental and immune-related pathways and their specificity across brain cell types. More broadly, this case study illustrates how network-informed transcriptomic models can complement traditional GWAS approaches by refining variant-to-gene attribution. By incorporating regulatory context and gene co-expression, these models help uncover distal targets of risk variants that might be missed by proximity-based methods. In doing so, they offer a pathway toward more biologically grounded interpretations of genetic associations in psychiatric disorders.

Objective 3: Evaluate the generalizability of prediction models in behaviourally complex phenotypes—Antisocial Behavior as a case study. In this final objective,

the focus shifts toward evaluating the boundaries of transcriptomic prediction for behavioural traits, using antisocial behaviour—proxied by impulsivity—as a case study. Leveraging genetically imputed brain expression scores derived from combined *cis*- and *trans*-regulatory models, this objective explores whether these predictors can account for individual variation in behavioural outcomes within a high-risk forensic cohort. The long-term aim is to employ these genetic predictions in a gene-environment interaction framework (e.g., parenting, trauma, socioeconomic adversity). However, the current results highlight substantial limitations in predictive power. Indeed, the GReX scores showed minimal utility in forecasting impulsivity, likely due to the trait's modest heritability, environmental plasticity, and measurement noise. This study, therefore, provides a proof of concept for future refinement of scores aimed at testing gene-environment interplay.

Rather than demonstrating successful prediction, this objective contributes a critical boundary analysis: clarifying the current limits of GReX-based models for behavioural phenotypes and identifying key barriers—statistical, biological, and conceptual—that future studies must address. It underscores the need for richer, multimodal frameworks that incorporate dynamic environmental measures, developmental timing, and brain-based intermediate phenotypes.

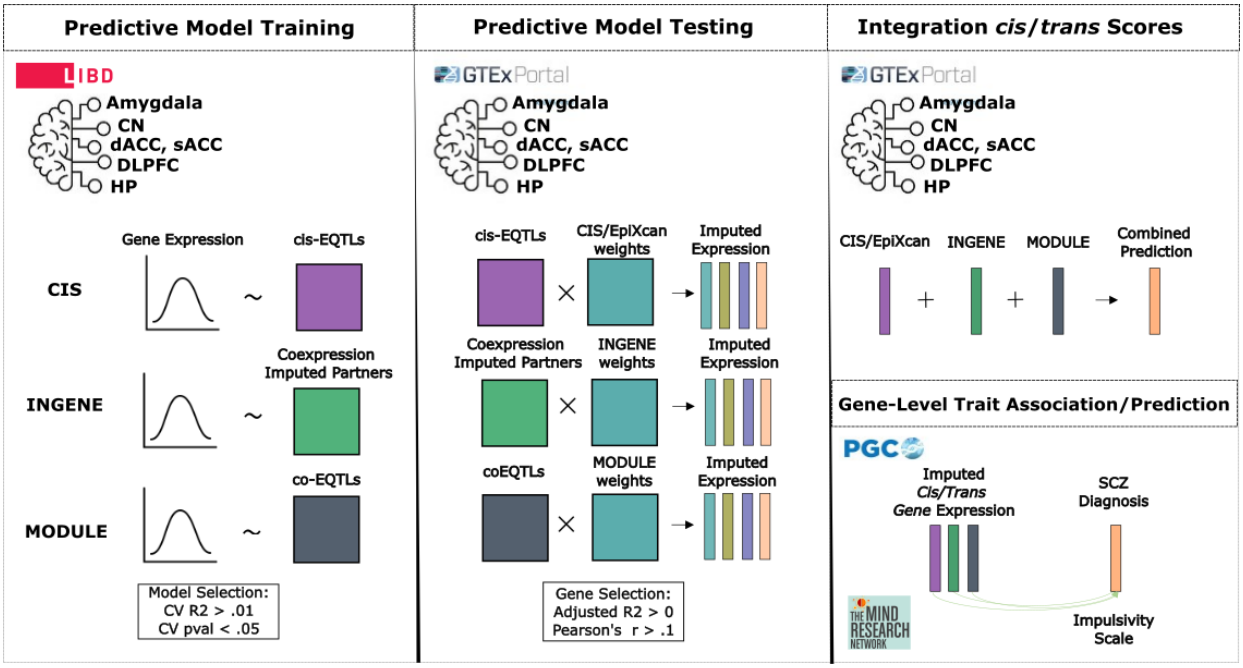

Figure 3: **Project Overview.** The analytical pipeline consists of five main stages. **Step 1: Model Training.** CIS, INGENE, and MODULE gene expression prediction models are developed using Elastic Net regularized regression applied to postmortem brain transcriptomic data from the LIBD dataset (Chapter 3, Section 3.3.1). **Step 2: Model Testing.** *cis*- (CIS, EpiXcan) and *trans*-based (INGENE, MODULE) models are used to impute gene expression in an independent genotype dataset (GTEx) (Chapter 3, Section 3.4.2). **Step 3: Score Integration.** For genes with multiple predictors, we combine *cis*- and *trans*-derived scores following the integration strategy described in Chapter 3, Section 3.3.5. **Step 4/5: Association Testing and Prediction.** The final integrated gene expression scores are evaluated in two primary applications: (i) association with SCZ diagnosis using individual-level genotype data from 62 PGC3 cohorts, as detailed in Chapter 4; and (ii) prediction of impulsivity in a forensic cohort of 468 adult inmates from the Mind Research Network, as presented in Chapter 5.

2.2 Case Studies

To ground the methodological framework of this thesis in real-world applications, the following case studies illustrate how co-expression-informed genetic models can enhance our understanding of complex psychiatric phenotypes. By applying the proposed integrative approach to SCZ and impulsivity as a proxy of antisocial behaviour—two conditions marked by polygenic architectures—we study the utility of combining network-based prediction models

with clinical, neurobiological, and environmental data. These case studies serve as both validation and extension of the proposed methodology, highlighting its potential to reveal biologically meaningful patterns that traditional approaches may overlook.

2.2.1 Schizophrenia as a Window into Polygenic Regulation

SCZ exemplifies the complexity of polygenic disorders with multifactorial transmission, where both genetic and environmental factors contribute to disease risk (Davis et al., 2016). Affecting approximately 1% of the global population and associated with a 15-year reduction in life expectancy, SCZ highlights the challenges of linking high heritability estimates (60–80%; Gottesman and Shields 1967) to specific biological mechanisms.

Large-scale GWAS, as those by the PGC consortium, have identified hundreds of common variants of small effect, collectively explaining ≈25% of SCZ heritability. Yet, PRS built from these variants capture only about 7% of phenotypic variance in case-control comparisons (Trubetskoy et al., 2022). This gap emphasizes a fundamental limitation: GWAS-identified loci, dispersed across the genome, often lack functional cohesion when aggregated additively, obscuring their potential convergence within biological networks and regulatory circuits (Arnedo et al., 2015; Pergola et al., 2019a, 2023b,c).

TWAS have emerged as a promising strategy to bridge this gap (Gusev et al., 2016, 2018). By leveraging GReX as an intermediate phenotype, TWAS can prioritize disease-relevant genes and provide more interpretable biological insights than SNP-based analyses alone.

Early applications of TWAS to SCZ, as summarized in ST8 in the Appendix A.1, include the study by Gusev et al. (2018), which integrated GWAS data from 79,845 individuals with gene expression profiles from brain, blood, and adipose tissues. This large-scale analysis identified 157 genes significantly associated with SCZ through TWAS, including 35 located outside previously established GWAS loci. Notably, 42 of these genes were associated with specific chromatin features, suggesting regulatory mechanisms that merit experimental

follow-up. The study also demonstrated that brain-derived expression and splicing patterns captured the majority of TWAS signals, underscoring the central role of brain-specific regulation in SCZ risk.

Building on this foundation, Gandal et al. (2018a) further demonstrated the utility of TWAS in prioritizing SCZ-associated genes. By imputing *cis*-regulated gene expression in the brain and integrating it with GWAS data, they identified 193 significant genes (164 outside the MHC) at Bonferroni-corrected thresholds, with 107 showing conditionally independent signals.

Collado-Torres et al. (2019) further expanded this approach by integrating GWAS and eQTL data across four expression feature types (gene, exon, junction, and transcript) from the dorsolateral prefrontal cortex (DLPFC) and hippocampus (HP). Their TWAS analyses uncovered 1,656 features spanning 624 genes significantly associated with SCZ in both regions, including many beyond known GWAS loci, highlighting novel risk associations.

Another major advancement came from **Huckins et al. (2019)**, who applied TWAS across 12 brain tissues, identifying 413 significant gene–tissue associations with SCZ, including 67 independent signals and 19 genes not previously linked to GWAS loci, thereby highlighting the power of transcriptome-based approaches to uncover novel risk genes.

Building on this foundation, **Hall et al. (2020)** further advanced biological specificity by employing TWAS models based on UK brain bank datasets, identifying 89 genes significantly associated with SCZ, 20 of which had not been previously reported. Their work prioritized genes involved in presynaptic and postsynaptic signaling, reinforcing synaptic dysfunction—particularly in neurotransmitter release and vesicle cycling pathways—as a key component of SCZ pathophysiology.

Most recently, **Bhattacharya et al. (2023)** introduced isoTWAS, a multivariate framework that integrates genetic data with isoform-level expression to enhance the discovery of trait-associated genes. Applying isoTWAS to SCZ, they identified multiple associations at the isoform level that were undetectable when analyzing total gene expression alone. No-

tably, the study prioritized specific isoforms of genes such as *AKT3*, *CUL3*, and *HSPD1*, highlighting the critical importance of incorporating isoform-level resolution into integrative genomic approaches, particularly for brain-related traits where alternative splicing plays a major role.

Despite these advances, most TWAS frameworks,including those cited above, remain *cis*-focused, relying on local genetic regulation within 1 Mb of gene loci. However, *cis*-eQTLs typically explain only 10–20% of gene expression variance, leaving much of the regulatory architecture uncharted. Critically, *trans*-eQTLs—variants that regulate gene expression at distant or even interchromosomal loci—are increasingly recognized as essential for understanding the diffuse, network-based nature of SCZ risk. Yet their detection remains statistically challenging due to modest effect sizes and the severe multiple testing burden.

To address these limitations, emerging approaches have begun to incorporate co-expression network structures, such as those identified by WGCNA, to organize genes into biologically coherent modules. Studies have shown that SCZ risk variants are non-randomly distributed across these transcriptional modules, often converging on hub genes that coordinate complex regulatory programs (Borcuk et al., 2024; Rodriguez-López et al., 2020).

Integrating co-expression architecture and *trans*-eQTL information into gene mapping frameworks offers a way to capture additional heritable signals while improving biological interpretability through regulatory and network context. Rather than treating genes in isolation or relying solely on local *cis*-effects, this approach leverages the modular structure of gene regulation to prioritize those genes most likely to be functionally impacted by genetic variation. This recognition motivates the central methodological innovation of this thesis: using biologically grounded, network-informed strategies to refine gene-level discovery in complex disorders like SCZ, moving from co-regulated modules toward identifying specific genes that mediate genetic risk.

2.2.2 Impulsivity as a Test Case for Transcriptomic Prediction

Antisocial behaviour encompasses a spectrum of actions that violate societal norms or infringe upon the rights of others, ranging from overt aggression and property destruction to more subtle traits such as impulsivity, deceitfulness, and social manipulation (Tuvblad and Beaver, 2013). Although clinically and behaviourally heterogeneous, antisocial behaviours share common neurodevelopmental and neuropsychological substrates, particularly deficits in attention, emotional regulation, and executive function. Among these dimensions, impulsivity is considered a fundamental trait that predisposes individuals to a variety of antisocial outcomes, including aggression, risk-taking, and noncompliance with social norms.

Heritability estimates for antisocial behaviour range from 40% to 50% (Rhee and Waldman, 2002; Burt, 2009). The genetic influence parallels those observed in core personality traits such as impulsivity and emotional reactivity, which not only correlate with antisocial behaviour but also mediate the social consequences of genetic predispositions (McAdams et al., 2013). Notably, the impact of personality traits on interpersonal dynamics suggests a gene–environment correlation mechanism: individuals with greater genetic similarity tend to experience more comparable life events, particularly within the social domain (Kendler and Baker, 2007).

However, antisocial behaviour does not arise solely from genetic predisposition; rather, it emerges from dynamic gene–environment interactions (G×E) across development. One of the earliest demonstrations of a genetic contribution to social reactions involved the association between rule-breaking behaviour in adolescents and a genetic variant affecting serotonin neurotransmission (Burt, 2009). This form of gene–environment interplay has substantial relevance for mental health, given its long-term consequences into adulthood. For instance, Pergola et al. (2019b) showed that genetic risk for SCZ was associated with adverse social experiences during early adolescence, which in turn predicted later psychosis. Further independent studies have supported this link between genetic risk, adverse social experiences

in adolescence, and adult-onset psychiatric disorders (Guloksuz et al., 2019; Schoeler et al., 2019). This emerging literature is consistent with well-established evidence that childhood maltreatment interacts with genetic vulnerability to shape unfavorable mental health outcomes (Caspi et al., 2002). Collectively, these studies provide empirical support for the idea that molecular sensitivity to environmental contexts plays a crucial role in behavioural trajectories.

Although early research primarily focused on candidate genes, subsequent GWAS have demonstrated that antisocial traits may be highly polygenic (Sanchez-Roige et al., 2018). Tielbeek et al. (2017) estimated SNP-based heritability for impulsivity and antisocial behaviour at 8–16% in a cohort of over 16,000 individuals, although individual loci exhibited only small effect sizes. Further studies have shown that personality traits genetically correlated with antisocial behaviour—such as impulsivity ($h^2 = 0.25$–0.36) and risk-taking—are associated with multiple loci implicated in serotonergic and dopaminergic signaling pathways (Lo et al., 2017).

Beyond methodological challenges, an additional biological constraint limits the feasibility of genetic prediction: impulsivity, a key dimension underlying antisocial behaviour, is, for instance, itself only modestly heritable. Meta-analytic studies estimate the heritability of impulsivity-related traits for about 45% (Congdon and Canli, 2008), while self-reported measures such as the Barratt Impulsiveness Scale (BIS-11) often capture even lower genetically driven variance due to strong environmental modulation and situational factors (Sharma et al., 2014). Consequently, even under ideal modeling conditions, the theoretical ceiling for genetically based prediction of impulsivity remains constrained.

In summary, this biological limitation motivated the selection of impulsivity as the focus of the present case study. Rather than targeting a highly heritable trait to maximize predictive success, the study sought to test whether improving GReX models—through the integration of *trans*-eQTL information—could expand the explainable variance for a behaviour characterized by modest genetic influence. In this context, impulsivity served as a

stringent and informative benchmark: if enhanced GReX models could recover significant predictive power for impulsivity, it would suggest broader applicability even to complex, environmentally sensitive traits. Conversely, failure to achieve meaningful prediction would reveal intrinsic boundaries in the capacity of transcriptomic proxy models to capture individual behavioural differences.

Chapter 3

Study 1: Training & Testing of *trans*-eQTL Algorithms

A slightly edited version of this chapter is currently under revision for publication in *Nature Genetics*: Rossi F., et al. *Co-expression-based models improve eQTL predictions and highlight many novel transcriptome-wide genes associated with schizophrenia.*

3.1 Introduction

The role of genetic variation in modulating gene expression and contributing to traits with complex heritability—through mechanisms such as *cis*- and *trans*-eQTLs—has been outlined in Chapter 1. The following sections focus on the specific challenges of identifying *trans*-eQTLs and explore methods aimed at improving their detection.

Challenges in *trans*-eQTL Discovery. Despite their biological significance, *trans*-eQTLs remain among the most difficult regulatory elements to map and interpret (Yao et al., 2017; Battle et al., 2014; Umans et al., 2021). Unlike *cis*-eQTLs, whose effects are often stronger and more localized, *trans*-eQTLs typically exhibit modest effect sizes and are scattered across the genome, often separated from their target genes by megabases or even located on different

chromosomes (Liu et al., 2022). This weak signal amplifies the statistical burden of testing, which can involve billions of SNP-gene pairs, dramatically increasing the false discovery rate if not adequately controlled (Wang et al., 2018). Nevertheless, stringent multiple testing correction, while necessary, drastically reduces the detection rate of true positives—especially when sample sizes are limited.

These issues are further amplified in postmortem brain tissue studies. Critical confounders such as RNA integrity (RIN), postmortem interval, batch effects, and cell-type heterogeneity can mask or mimic true *trans* effects (Jaffe et al., 2018; GTEx, 2017). For example, expression differences driven by cell-type proportions can lead to spurious correlations (GTEx, 2017; Mostafavi et al., 2018), and technical variability across brain regions and sequencing protocols can introduce additional noise, entangling replication efforts across cohorts (Fromer et al., 2016; GTEx, 2020).

Projects like GTEx (GTEx, 2015, 2017, 2020) and the CMC (Fromer et al., 2016) have made significant strides by increasing sample sizes and developing sophisticated normalization pipelines. However, even in these landmark datasets, *trans*-eQTL discovery remains difficult, with most studies reporting relatively few replicable associations outside of *cis* loci (Gamazon et al., 2015; Battle et al., 2014; Võsa et al., 2021).

Value of Postmortem Brain Datasets. Despite the inherent complexities of working with postmortem tissues, these datasets offer an irreplaceable window into the transcriptional and regulatory architecture of the human brain (Hawrylycz et al., 2012; Fromer et al., 2016), which likely is the most complex and least accessible organ in human genetics. Unlike peripheral tissues, the brain is highly heterogeneous in structure and function, with gene expression patterns that vary dramatically across regions, cell types, and developmental stages (GTEx, 2020; Hartl et al., 2021). Postmortem brain datasets provide a rare opportunity to study these differences by delivering matched genotype and transcriptome data from neuroanatomically defined areas, enabling researchers to explore region-specific regu-

latory mechanisms with unprecedented resolution (Hawrylycz et al., 2012; Collado-Torres et al., 2019; Benjamin et al., 2022; Sportelli et al., 2024; Pergola et al., 2023a; Huckins et al., 2019).

Moreover, these resources often include deep phenotypic characterization of donors, covering variables such as age, sex, psychiatric diagnosis, medication status, and neuropathology, allowing for integrative analyses that link genetic regulation to neurobiological traits and disorders. Projects like the Brainseq,CMC, and GTEx have demonstrated that such datasets are essential for uncovering brain-specific eQTLs, many of which are not detectable in blood or other accessible tissues (GTEx, 2020).

Importantly, postmortem brain studies allow researchers to move beyond proxy tissues and generate gene expression prediction models that are biologically and anatomically relevant (Hall et al., 2021). This is especially critical in psychiatric and neurodevelopmental disorders, where pathogenic mechanisms are expected to unfold within specific circuits or developmental windows (Hall et al., 2021; Pergola et al., 2023a; Gandal et al., 2018a). The spatial and molecular precision afforded by postmortem resources enables modeling efforts to incorporate transcriptional features that would otherwise be missed, facilitating the discovery of new risk mechanisms and therapeutic targets.

A central strength of this work is the use of the LIBD resource, part of which is publicy available in the Brainseq and PsychENCODE projects——one of the most comprehensive brain tissues–specific postmortem resources available. LIBD provides high-depth RNA-sequencing and genotype data across six anatomically and functionally distinct brain regions: amygdala (Jaffe et al., 2022; Zandi et al., 2022), caudate nucleus (CN) (Benjamin et al., 2022), dorsal anterior cingulate cortex (dACC) (Jaffe et al., 2022), dorsolateral prefrontal cortex (DLPFC) (Jaffe et al., 2018; Collado-Torres et al., 2019; Jaffe et al., 2022), hippocampus (HP) (Collado-Torres et al., 2019; Daskalakis et al., 2024), and subgenual anterior cingulate cortex (sACC) (Zandi et al., 2022). This anatomical granularity supports brain-region–specific modeling of gene regulation, enabling the development of tailored pre-

diction models with high biological fidelity.

A co-expression-Guided Framework for *Trans* Modeling. While prior work has used co-expression networks for pathway enrichment, functional annotation, and gene prioritization (Radulescu et al., 2020; Walker et al., 2019; Werling et al., 2020; Pergola et al., 2017, 2019a; Fromer et al., 2016; Li et al., 2018b; Gandal et al., 2018a; Hartl et al., 2021), the framework presented in this thesis introduces a novel strategy: integrating co-expression relationships directly into the modeling of *trans*-regulatory effects. Specifically, we use gene co-expression modules to constrain the *trans*-eQTL search space, based on the hypothesis that co-expressed genes are more likely to share upstream regulators or operate within the same functional circuits (Pergola et al., 2023c; Borcuk et al., 2024; Pergola et al., 2023b).

There is strong empirical support for this hypothesis. Co-expression-based clustering has repeatedly revealed biologically meaningful gene modules that reflect shared transcriptional regulation, often uncovering regulatory hotspots where single genetic variants influence large groups of co-expressed genes (Pergola et al., 2023c; Fagny et al., 2017; Fromer et al., 2016). These modules can increase power by reducing the dimensionality of *trans*-eQTL searches and by aggregating weak signals that would be missed in single-gene analyses.

In the brain, this modular structure is particularly relevant. Pergola and colleagues (Pergola et al., 2017, 2019a) have demonstrated that SCZ risk genes are embedded within tightly regulated co-expression modules in the prefrontal cortex, and that these modules predict both clinical response and brain function (Fromer et al., 2016; Radulescu et al., 2020; Gandal et al., 2018a). These findings underscore the utility of co-expression not only for defining biologically coherent gene sets, but also for linking genetic risk to neurobiological processes in a developmentally and regionally specific manner (Pergola et al., 2023c,a).

Modeling Strategy and Statistical Approach. Building on prior work, the framework presented in this thesis leverages co-expression information not just as an annotation layer, but as a core structural prior in *trans*-eQTL modeling. In particular, two algorithms were

developed: INGENE, which infers *trans* predictors via the *cis*-eQTLs of co-expressed partner genes; and MODULE, which directly links *trans*-regulators to gene modules. These co-expression-informed *trans*-predictive strategies enable generalizable modeling across the transcriptome by reducing dimensionality and improving interpretability. Rather than conducting exhaustive genome-wide *trans* scans, both approaches use data-driven priors to guide SNP inclusion, thereby preserving statistical power and biological relevance.

To estimate the relationship between genetic variation and gene expression, we employed elastic net regression (Zou and Hastie, 2005), a regularized linear modeling approach that balances the sparsity of LASSO with the stability of ridge regression. This method is particularly well-suited for genomic settings characterized by high dimensionality and correlated predictors, such as SNPs in LD. For each gene, we trained predictive models using both proximal *cis*-SNPs and distal *trans*-SNPs identified through our INGENE and MODULE pipelines.

In addition to developing a custom *cis*-based model (CIS), we incorporated predictions from the publicly available EpiXcan framework (Zhang et al., 2019), which enhances *cis*-eQTL-based expression modeling by integrating tissue-specific epigenomic annotations. While *cis* and *trans* regulatory mechanisms are biologically distinct, their integration has the potential to yield complementary improvements in prediction accuracy and transcriptome coverage. However, to our knowledge, no prior studies have incorporated *trans*-eQTL signals into unified predictive models alongside *cis* signals to enhance gene-level prediction accuracy. Our framework addresses this gap by generating composite models that merge four layers of regulatory information: *cis*-based signals from both EpiXcan and CIS models, and *trans*-based signals from INGENE and MODULE. This integrated strategy allows for a more complete representation of gene regulation, capturing both local and long-range effects, and enhances the potential for accurate transcriptomic imputation in downstream analyses.

Preview of Results. This chapter presents the development, training, and evaluation of a co-expression-informed *trans*-eQTL modeling framework (Figure 4). We begin by outlining the datasets and quality control procedures, using the LIBD dataset as the primary training resource. We then describe the methodological pipeline for co-expression network integration, SNP prioritization, model training, and performance benchmarking.

We show that the INGENE and MODULE approaches successfully predict expression for over 20,000 genes across brain regions with cross-validated $R^2 > 0.01$, nearly doubling the transcriptome coverage achieved by traditional *cis*-only models. When evaluated in independent cohorts (GTEx and CMC), prediction performance remained robust across diverse brain regions, confirming cross-cohort generalizability. Furthermore, integrating *cis* and *trans* predictors substantially improved gene-level prediction accuracy. These results establish a strong foundation for the downstream application of these models to gene-trait association and prediction analyses, which will be presented in Chapter 4 and 5.

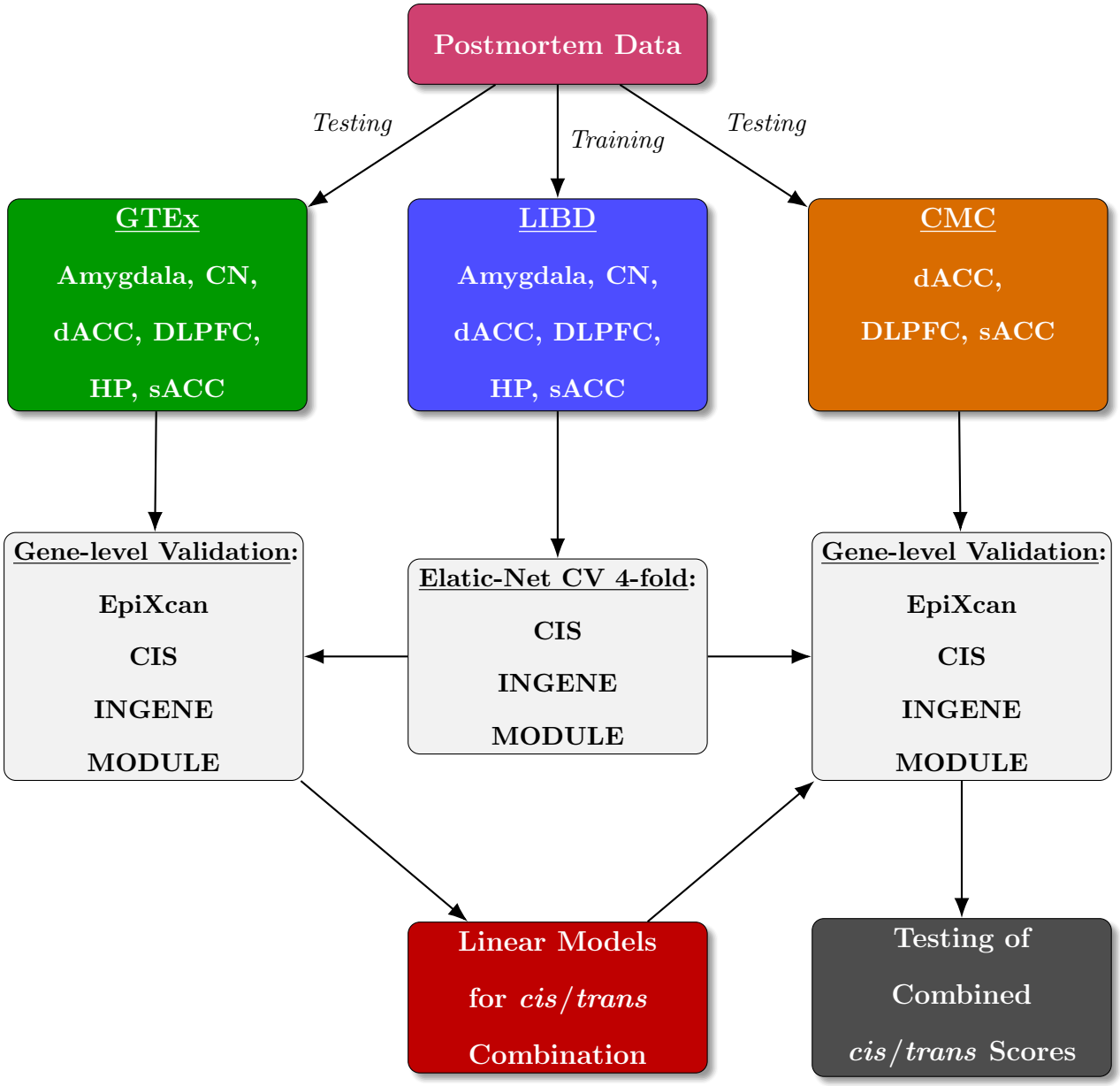

Figure 4: **Integrative pipeline for genetically regulated gene expression modeling across brain transcriptomic datasets**. Postmortem transcriptomic and genotype data from the LIBD, GTEx, and CMC brain repositories were used to train and evaluate elastic-net models of GReX. Models were trained on LIBD data using *cis*-eQTLs (CIS model) and co-expression-informed *trans* features (INGENE and MODULE models) across six brain regions. Independent testing was performed in the GTEx and CMC datasets. Predictions from the *cis* and *trans* models were combined using linear modeling in the GTEx dataset and then evaluated for predictive power in the CMC cohort across available brain regions.

3.2 Data

3.2.1 Overview of Postmortem Datasets (LIBD, GTEx, CMC)

To develop, train, and evaluate predictive models of gene expression incorporating both *cis* and co-expression-informed *trans*-eQTLs, we leveraged three large-scale postmortem human brain transcriptomic resources: LIBD (https://www.libd.org/), GTEx (GTEx, 2015, 2017, 2020), and CMC data (Fromer et al., 2016). These datasets offer high-quality genotype and RNA-sequencing data from neuroanatomically defined brain regions (Table 1) and represent independent cohorts. To maximize robustness and minimize bias in our predictive modeling framework, we structured dataset usage based on their respective characteristics. The LIBD dataset was selected for training because it provides the largest sample size and includes six brain regions (Table 1), offering a strong substrate for building stable prediction models. GTEx was employed as an independent validation resource: although it has a smaller sample size, it includes overlapping brain regions, which allowed us to assess predictor performance in an external cohort while guarding against overfitting. Finally, the CMC dataset, which includes two regions (Table 1), was reserved for replication, ensuring that findings generalized to an additional independent cohort. Collectively, they provide a robust substrate for modeling genetically regulated expression in the human brain.

Demographic Summary and Selection criteria. The discovery dataset of this study (LIBD) included postmortem brain specimens and genotype data from both neurotypical controls (NC) as well as individuals diagnosed with SCZ, bipolar disorder (BP), and major depressive disorder (MDD). To keep consistency with our primary postmortem replication dataset (GTEx) and enhance the statistical power of our predictive pipeline, we included only individuals of European ancestry aged 17 years or older. Table 1 summarizes the demographic characteristics of subjects included in the postmortem datasets across brain regions.

Table 1: **Postmortem data demographics across brain regions and datasets.** *Abbreviations*: CN = caudate nucleus; dACC = dorsal anterior cingulate cortex; sACC = subgenual anterior cingulate cortex; DLPFC = dorsolateral prefrontal cortex; HP = hippocampus.

Region	**LIBD**			**GTEx**			**CMC**		
	N	**Age (mean ± SD)**	**% Female**	**N**	**Age (mean ± SD)**	**% Female**	**N**	**Age (mean ± SD)**	**% Female**
AMY	461	47±16	29.0%	114	58±10	30.0%	—	—	—
CN	211	50±16	25.1%	199	59±10	25.0%	—	—	—
dACC	176	47±15	35.3%	141	60±10	28.4%	159	61±18	33%
sACC	508	47±16	31.0%	—	—	—	—	—	—
DLPFC	584	46±15	31.5%	160	59±9	27.0%	405	70±17	40%
HP	236	46±16	23.0%	145	59±11	26.2%	—	—	—

3.2.2 Gene Expression and Genotype Data Processing

Postmortem RNA-seq Data. LIBD samples included homogenate RNA-seq from the following brain regions: amygdala (Jaffe et al., 2022; Zandi et al., 2022), CN (Benjamin et al., 2022), dACC (Jaffe et al., 2022), DLPFC (Jaffe et al., 2018; Collado-Torres et al., 2019; Jaffe et al., 2022), HP (Collado-Torres et al., 2019; Daskalakis et al., 2024), and sACC (Zandi et al., 2022). GTEx RNA-seq data (v8, dbGaP phs000424.v8.p2) were downloaded from the following brain regions: ACC, amygdala, CN, DLPFC (BA9), and HP. CMC RNA-seq data were obtained from the CommonMind Knowledge Portal for DLPFC (release 3.0, syn18097439) and ACC (version 6.0, syn29442240).

For each dataset and brain region, gene-level quantification, filtering, and normalization were performed using standard pipelines, including expression filtering, TPM conversion, log-transformation, and outlier removal. Expression residualization was conducted through linear modeling to adjust for technical and biological confounders, followed by Blom normalization to mitigate deviations from the normal distribution (Pergola et al., 2017). Full methodological details, including quality control steps, covariate specifications, and data processing workflows, are provided in the Appendix section B.1.

Genotype Data LIBD, GTEx and CMC genotype data were generated as previously described (Jaffe et al., 2018; Benjamin et al., 2022; GTEx, 2015; Fromer et al., 2016) and as detailed in the Appendix section B.2. Post-imputation quality control was conducted uniformly with PLINK toolkit version 1.07 (Purcell et al., 2007). SNPs were excluded based on MAF ≤ 0.01, Hardy-Weinberg equilibrium p-value $< 10^{-6}$, or $> 5\%$ missingness. Individuals were excluded for $> 2\%$ missing genotypes or relatedness $\hat{\pi} > 0.125$. Population structure was inferred using principal components aligned to HapMap3 (Altshuler et al., 2010), retaining only individuals with $> 90\%$ overlap with European reference clusters.

The final number of genotypes after processing was as follows: LIBD = 7,521,829; GTEx = 8,623,182; CMC = 5,859,752. Additionally, we subset LIBD genotypes with overlapping SNPs in the GTEx cohort to maximize the power of the analyses, resulting in a final number of 6,819,569 SNPs.

3.2.3 Source of Co-Expression Networks

A defining feature of the modeling framework presented in this thesis is the incorporation of *trans*-regulatory information through biologically informed co-expression networks.

Critically, co-expression modules were not constructed *de novo* in this study. Instead, we used co-expression networks published in prior large-scale transcriptomic studies (Hartl et al., 2021; Pergola et al., 2019a, 2023b; Radulescu et al., 2020; Gandal et al., 2018b,a;

Fromer et al., 2016; Werling et al., 2020; Li et al., 2018b; Walker et al., 2019), which had already identified robust modules of co-expressed genes in brain tissues using the WGCNA method (Langfelder and Horvath, 2008) (see Table ST3 for details). These networks group genes based on consistent expression correlation patterns across individuals and the gene clusters within them (called modules) are often enriched for biological functions, pathways, or cell-type-specific signatures. By leveraging previously validated modules, we ensured biological relevance while minimizing the influence of specific methodological choices—such as dataset characteristics or parameter settings—that can introduce variability in *de novo* module detection.

For each target gene, the corresponding co-expression module was identified from these reference networks, and its member genes were leveraged to guide the selection of putative *trans*-regulatory variants as described in the Methods section 3.3.1.

3.3 Methods

3.3.1 Predictive Algorithms: CIS, INGENE, and MODULE

To model genetically regulated gene expression and capture both *cis*- and *trans*-regulatory effects, we developed three complementary predictive frameworks: (i) **CIS**, a baseline model relying solely on local *cis*-regulatory variants; (ii) **INGENE** (*Imputed Network Gene Expression trans-eQTLs*), which leverages co-expression networks to predict a gene's expression from its distal partners' genetically imputed profiles; and (iii) **MODULE** (*MODule qUantitative trait Loci Eigengene*), which models gene expression via *trans*-acting SNPs associated with co-expression module eigengenes.

All models were trained using matched genotype and expression data from the LIBD dataset. While the CIS model relied solely on local genetic variation, INGENE and MODULE further incorporated network structure by leveraging 48 co-expression modules (Table ST3). Model fitting was performed using elastic net (EN) regularization with nested

cross-validation to optimize predictive performance (see Appendix B.3 for EN algorithmic details).

The following sections provide detailed descriptions of CIS, INGENE, and MODULE algorithms, including training procedures, feature selection strategies, and performance benchmarks.

CIS and EpiXcan Models: Baseline *cis*-Regulatory Predictors

The CIS model serves as a baseline framework for predicting gene expression based solely on local *cis*-regulatory variation. Following the widely adopted PrediXcan approach (Gamazon et al., 2015), each gene's expression was modeled from SNPs located within ±1 Mb of its transcription start site (Figure 5).

Model training used EN regression implemented via the `cv.glmnet` R package, with nested 4-fold cross-validation to optimize the penalty parameter λ. Models were retained if they met minimal predictive performance thresholds: $R^2_{\text{CV}} \geq 0.01$, $p_{\text{CV}} < 0.05$, and Pearson's $r \geq 0.1$.

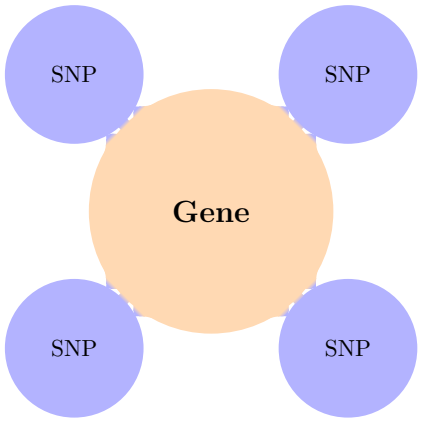

Figure 5: **CIS model graphical representation.** Gene expression is modeled from local *cis*-acting genetic variation. A single EN model is trained per gene, independent of co-expression structure.

EpiXcan Integration. To benchmark and complement our in-house CIS models, we incorporated external prediction weights from the EpiXcan resource (Zhang et al., 2019),

publicly available at https://predictdb.org/. Specifically, we downloaded the DLPFC models, which integrate epigenetic priors to enhance prediction accuracy. EpiXcan models served two purposes: (1) independent benchmarking against our CIS models, and (2) expanding the pool of *cis*-based predictors available for downstream network-based frameworks such as INGENE.

Together, the CIS and EpiXcan models establish a consistent *cis*-regulatory prediction layer, against which our co-expression-informed *trans*-models (INGENE and MODULE) are compared.

INGENE: Imputed Network Gene-Expression *trans*-eQTL

INGENE predicts the expression of a target gene by aggregating the *cis*-predicted expression of its co-expressed partners within a given module—thus it does not use *cis*-SNPs for the target gene directly, but rather builds on *cis*-based predictions of genes within the same co-regulatory context (Figure 6A). This indirect approach allows it to capture *trans*-regulatory signals encoded within the broader network architecture.

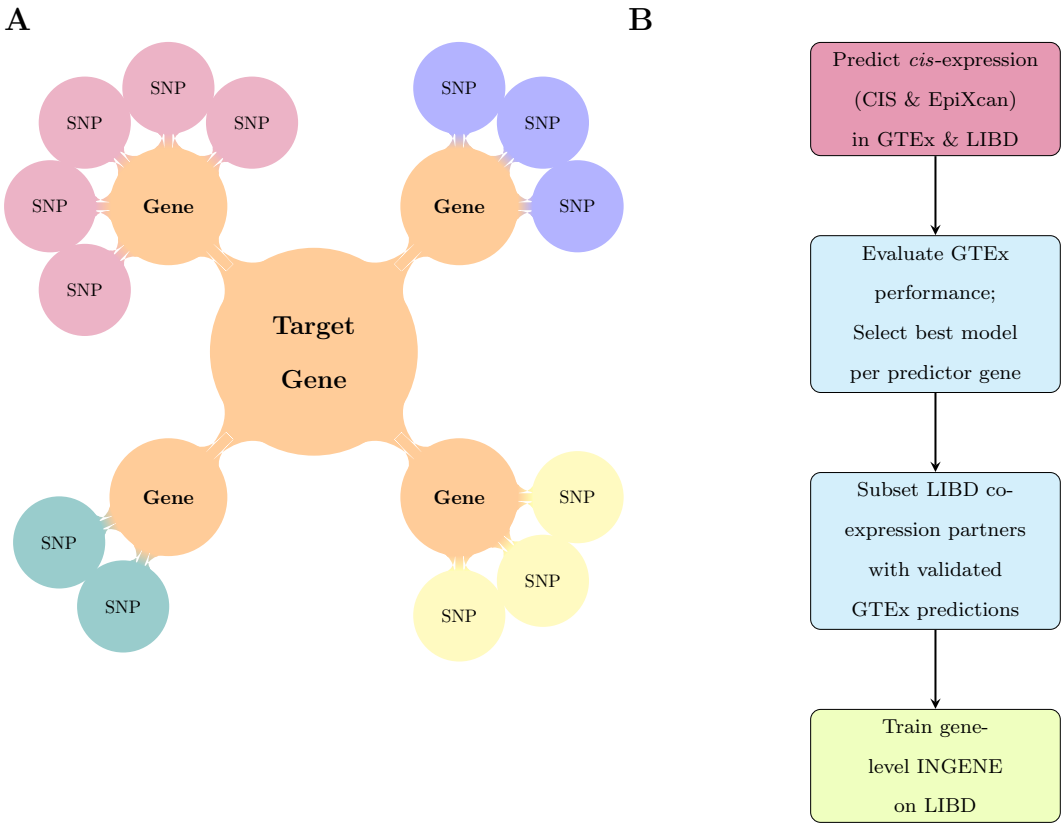

Figure 6: **Overview of the INGENE framework. A)** Target gene expression is modeled using the *cis*-predicted expression of co-expressed genes. **B)** INGENE training pipeline: candidate predictors are imputed in LIBD, their performance is benchmarked in GTEx to retain only robust predictors and to choose between CIS and EpiXcan models, and final target gene models are trained exclusively in LIBD. GTEx is thus used solely as an external predictor-validation step, while CMC serves as an independent replication dataset.

Mathematically, INGENE models the expression of a target gene g weighted sum of the genetically predicted expression levels of its co-expressed partner genes:

$$\hat{Y}_g = \sum_{p=1}^{P} w_{p,g} \cdot \hat{Y}_{g_p} \tag{3.1}$$

where $\hat{Y}_{g_p}$ is the *cis*-predicted expression of co-expression partner g_p, and $w_{p,g}$ is the EN-derived weight for that partner. **Training procedure.** To implement INGENE, we designed a multi-step pipeline (Figure 6B) to generate gene-level prediction models from co-expression networks. Below, we outline the main stages.

1. Expression Imputation and Predictor Generation. We first computed *cis*-predicted gene expression values using two models: (1) the in-house CIS model, and (2) the EpiXcan model (Zhang et al., 2019). These models were applied to LIBD genotypes to generate candidate predictors (co-expression partners) for each target gene. Predictors were retained only if they showed reliable out-of-sample performance when benchmarked against observed expression in GTEx, ensuring that weak or spurious predictors were excluded. At this stage, GTEx functioned solely as an external filter, while all subsequent target gene training was performed in LIBD.

2. Model Selection. For each predictor that passed Step 1, we compared the performance of the CIS and EpiXcan models in GTEx using adjusted R^2, and selected the better-performing model as the source of predicted expression. This ensured that the final INGENE feature set was composed of robust, GTEx-validated predictors, without involving GTEx in target gene training.

3. Co-Expression Partner Assembly and Trans-Filtering. For each target gene, we then identified its co-expression partners within each of the 48 network modules. We filtered this list to retain only partners with validated *cis*-predicted expression. To preserve the *trans*-only nature of INGENE, we additionally removed any co-expression partner located within ± 1 Mbp of the transcription start site of the target gene. This genomic distance filter prevents potential contamination by shared *cis*-regulatory SNPs between partner and target genes.

4. Model Training in LIBD. Using the selected and filtered co-expression predictors, we trained EN regression models to predict expression of each target gene within each co-expression module. Importantly, all training was performed exclusively in LIBD, using LIBD genotype-predicted expression values as input. We adapted the original PrediXcan training pipeline (Gamazon et al., 2015), applying the `cv.glmnet` R function with nested 4-fold cross-validation to tune the regularization parameters λ. This process was repeated for every gene

across all modules and tissues, resulting in multiple models per gene, each reflecting a unique co-expression context.

5. Model Filtering and Final Selection. After training, we retained only models that passed performance thresholds (adjusted $R^2_{\text{CV}} \geq 0.01$, $p < 0.05$, Pearson's $r \geq 0.1$). For each gene, we selected the best-performing model across all 48 networks based on these metrics. This selection yielded a robust, biologically informed atlas of INGENE models, each capturing distal, co-regulation-mediated expression patterns. Non-zero weights and associated summary statistics were extracted.

Algorithm 1 INGENE Training Pipeline

Require: Co-expression networks $\{\mathcal{M}^{(n)}\}_{n=1}^{48}$, predicted expressions $\hat{Y}_{g_p}$ from LIBD, validation metrics from GTEx

for each gene g **do**
  for each co-expression network $n = 1$ to 48 **do**
    Identify co-expression module $\mathcal{M}_g^{(n)}$ containing g
    Select co-expression partners $g_p \in \mathcal{M}_g^{(n)}$
    Filter out partners within ± 1 Mbp of g's TSS
    Retain only partners with validated cis-predictions (from GTEx)
    if fewer than 2 predictors remain **then**
      continue
    end if
    Train elastic net model: predict Y_g on $\{\hat{Y}_{g_p}\}$ using LIBD
    Tune λ via nested 4-fold cross-validation
    Store model weights $w_{p,g}^{(n)}$ and performance metrics
  end for
  Select best model for gene g across 48 networks based on cross-validated R^2, p-value, and Pearson r
  Retain model if $R^2 \geq 0.01$, $p < 0.05$, $r \geq 0.1$
end for

MODULE: MODule qUantitative trait Loci Eigengene

MODULE is a co-expression-informed model designed to predict gene-level expression using *trans*-eQTLs identified through their association with the eigengene of a co-expression module. The module eigengene (ME)——defined as the first principal component of the expression matrix for each module——is used as a proxy for coordinated gene activity (Figure 7A). This model assumes that a gene g with P co-expressed partners and a set of markers

K associated with the ME can be predicted through *trans*-acting regulatory effects. Figure 7B illustrates MODULE pipeline flow.

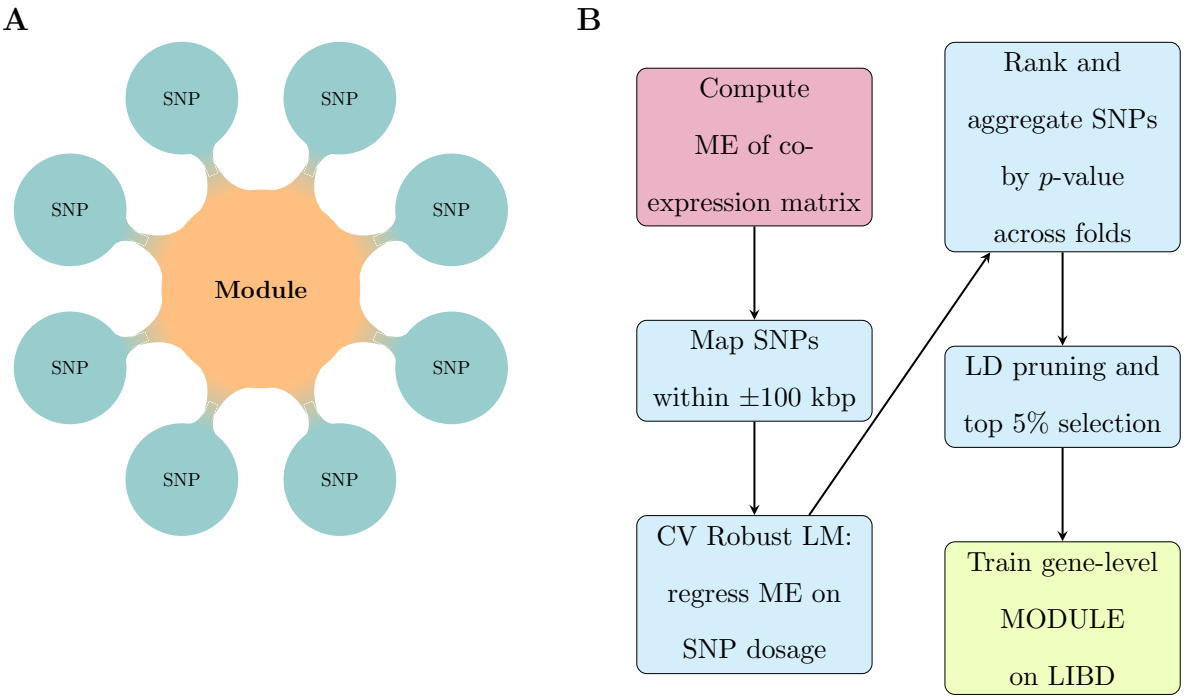

Figure 7: **Overview of the MODULE framework. (A)** Genetic regulation is modeled at the module level using SNPs associated with the module eigengene (PC1 of expression). **(B)** MODULE training pipeline: SNP-to-ME associations are identified using cross-validated robust regression, prioritized via rank product, LD-pruned, and used to train gene-level elastic net models.

Model and Assumption. For each gene g, MODULE models expression as a linear combination of co(expression)-eQTLs associated with its module's eigengene. Importantly, to preserve a strictly *trans*-acting regulatory architecture, we excluded all SNPs within the *cis*-window (±1 Mbp) of the gene. The resulting *trans*-prediction is formulated as:

$$\hat{Y}_g = \sum_{k \in K_{-cis}} w_{k,g} \cdot X_k \tag{3.2}$$

where $\hat{Y}_g$ is the predicted expression of gene g, K_{-cis} is the set of co-eQTLs excluding SNPs in the cis-region of g, $w_{k,g}$ is the learned weight of SNP k, and X_k is the genotype dosage of the reference allele for SNP k.

1. Co-eQTL Discovery. To identify co-eQTLs for each module, we first computed the ME for each co-expression module independently within each brain region and network. ME was calculated as the first principal component of the expression matrix of the module.

We then mapped candidate SNPs using MAGMA (v1.09b) (de Leeuw et al., 2015), selecting all 1000 Genomes Phase 3 European SNPs (de Leeuw et al., 2015) within ±100 kb of genes in the module. To reduce false positives, we implemented a stratified 4-fold cross-validation framework. In each fold, we recomputed the ME from RNA-seq data and tested SNP-to-ME associations via robust linear modeling. Specifically, we used genotype dosage as the predictor and ME as the outcome, extracting p-values via the `f.robftest` function from the R `sfsmisc` package.

SNPs were ranked by ascending p-value within each fold, and ranks were aggregated across folds using the rank product method. To ensure independence, we calculated pairwise R^2 within ±250 kb of the top-ranked SNP and performed iterative LD pruning, retaining only those SNPs with $R^2 < 0.1$ (Pergola et al., 2019a). Finally, we selected the top 5% of ranked SNPs for each module, yielding a compact and cross-validated set of ME-associated co-eQTLs.

2. Gene-Level Model Training. For each gene g belonging to a given module, we subset LIBD genotypes to derive a set of genetic predictors. To rule out that predictions were driven by residual *cis*-effects, we removed any SNPs located within ±1 Mbp of the gene's start and end coordinates. Using these filtered SNPs as predictors, we trained elastic net regression models to predict the expression of g in LIBD, with hyperparameter λ optimized via nested 4-fold cross-validation (see section B.4 in Appendix for further details).

3. Model Filtering and Database Construction. We retained only models meeting quality control thresholds: $R^2_{\text{CV}} \geq 0.01$, $p_{\text{CV}} < 0.05$, and Pearson correlation $r \geq 0.1$. Non-zero SNP weights and associated summary statistics were extracted for downstream use.

Algorithm 2 MODULE Model Training

Require: Module eigengene ME_M, genotype matrix X, expression vector Y_g

Identify candidate co-eQTLs K: SNPs within $\pm$100 kb of module genes
In each fold, regress ME_M on X_k and rank SNPs by p-value
Aggregate ranks across folds using rank product
Apply LD pruning ($R^2 < 0.1$, $\pm$250 kb) and retain top 5% $\Rightarrow K_{\text{trans}}$
for each gene g in module M **do**
    Remove SNPs in $\pm$1 Mbp cis-window of g from K_{trans}
    Train elastic net model: regress Y_g on $X_{K_{\text{trans}}}$
    Tune λ, α via nested 4-fold cross-validation
    Store weights $w_{k,g}$ and performance metrics
end for

3.3.2 Model Application to Independent Genotypes

After training, all models (CIS, INGENE, MODULE, along with EpiXcan) were applied to independent genotype data from the GTEx and CMC cohorts using the `predict.py` script from the MetaXcan framework (Barbeira et al., 2018). This script reads the trained EN weights for each gene and imputes GReX by applying them to genotype dosage data.

For the CIS and EpiXcan models, a single gene-level prediction $\hat{Y}_g$ was obtained directly from the weights. In contrast, the INGENE and MODULE models each yielded a set of predictions $\hat{Y}_g^{(n)}$ per gene, where $n = 1, \ldots, 48$ indexes the distinct co-expression networks.

3.3.3 INGENE & MODULE Network-Averaged Prediction

The gene-level expression predictions produced by INGENE and MODULE across the 48 co-expression networks were aggregated to generate a single consensus value per gene. This step ensures that the final imputed expression reflects robust and generalizable patterns across diverse co-regulatory contexts.

For each gene g, the final predicted expression $\hat{Y}_g$ was computed as the mean of all successfully trained network-specific models:

$$\hat{Y}_g = \frac{1}{N_g} \sum_{n=1}^{N_g} \hat{Y}_g^{(n)} \tag{3.3}$$

where $N_g \leq 48$ is the number of networks that yielded a valid model for gene g.

Algorithm 3 Network-Averaged Prediction in Testing Dataset

Require: Trained models $\{\hat{Y}_g^{(1)}, \ldots, \hat{Y}_g^{(N)}\}$, testing genotype matrix X

for each gene g **do**
  for each network $n = 1$ to N **do**
    if model $\hat{Y}_g^{(n)}$ is available **then**
      Apply weights to genotypes $\Rightarrow$ compute $\hat{Y}_g^{(n)}$
    end if
  end for
  Average predictions: $\hat{Y}_g = \frac{1}{N_g} \sum_n \hat{Y}_g^{(n)}$
end for
return Final predicted expression matrix $\hat{Y}$

Evaluation Metrics in Independent Testing Dataset

To assess the generalizability and out-of-sample performance of the predictive models we used two performance metrics:

- **Pearson Correlation Coefficient (r).** For each gene, the Pearson correlation between observed expression values $Y = (y_1, \ldots, y_n)$ and predicted values $\hat{Y} = (\hat{y}_1, \ldots, \hat{y}_n)$. This metric quantifies the linear association between predicted and observed expression levels.

- **Adjusted Coefficient of Determination (R^2_{adj}).** To account for the number of predictors and sample size, we used the adjusted R^2 from the LM function in R of observed

expression on predicted expression. The adjusted R^2 penalizes model complexity and provides a more conservative estimate of model fit.

A gene was considered successfully predicted in a given tissue if $r > 0$ and $R^2_{\text{adj}} > 0$. This dual-threshold criterion filtered out unstable or degenerate predictions.

3.3.4 Maximum Likelihood Evaluation of Trans-Predictive Enhancement

To assess the enhanced explanatory power of our combined *cis*- and *trans*-predictions on the observed gene expression, we performed a maximum likelihood estimation (MLE) comparison between full and null models using the `anova` function in R. This analysis was conducted independently within each brain region using observed GTEx gene expression (logTPM) as the dependent variable.

The full model included the following covariates: sex, mean age, RNA integrity number (RIN), RNA sequencing rate, overall mapping rate, postmortem interval (PMI), the first three PCs, and the first five GEs. Expression predictions (*cis* and/or *trans*) were then added as explanatory terms depending on model availability. The null model included only the covariates.

For genes uniquely predicted by one method (CIS [C], EpiXcan [E], INGENE [I], or MODULE [M]), we used the following model comparison:

$$H_0 : \quad y \sim \text{covariates}$$
$$H_1 : \quad y \sim y_{C,E,I,M} + \text{covariates}$$

For genes with both *cis* and *trans* predictions available, the MLE framework compared models with and without the *trans* component:

$$H_0: \quad y \sim y_{\text{cis}} + \text{covariates}$$
$$H_1: \quad y \sim y_{\text{cis}} + y_{\text{trans}} + \text{covariates}$$

If only *trans*-predicted values were available for a gene (i.e., no valid *cis* model), the null and full models were:

$$H_0: \quad y \sim \text{covariates}$$
$$H_1: \quad y \sim y_{\text{trans}} + \text{covariates}$$

For each gene, the model with the most significant increase in likelihood (at $\alpha = 0.05$) was selected as the best explanatory model—categorized as *cis*-only, *cis*+*trans*, or *trans*-only.

3.3.5 Combination of *cis* and *trans* Prediction Scores

Integration of *cis*/*trans* predictive models was performed during the testing phase using the GTEx dataset. For each gene g with at least two available predictions among the set of models $\mathcal{M} = \{\text{CIS}, \text{EpiXcan}, \text{INGENE}, \text{MODULE}\}$, we fitted a multiple linear regression model with observed gene expression as the outcome:

$$Y_g = \beta_0 + \sum_{m \in \mathcal{M}_g} \beta_m \cdot \hat{Y}_g^{(m)} + \varepsilon \tag{3.4}$$

where Y_g is the observed expression of gene g in GTEx, $\hat{Y}_g^{(m)}$ is the predicted expression from model $m \in \mathcal{M}_g$, β_m is the learned regression coefficient for model m, and ε is the residual error term. The subset $\mathcal{M}_g \subseteq \mathcal{M}$ includes only models with valid predictions for gene g.

Model performance was evaluated by applying the fitted regression weights to the CMC

cohort. Predicted expression values were compared against observed expression levels, and gene-level R^2 was computed to assess accuracy. To determine the benefit of multi-model integration, we compared mean R^2 values from the combined model against those of each individual predictor using Mann–Whitney U tests.

3.4 Results

3.4.1 Evaluation of *cis*- and *trans*-Model Training Performance

We used the LIBD dataset for training as it featured larger sample sizes than GTEx for the brain regions we considered. We performed a comparative analysis between our *cis*-predictive model (CIS) and our *trans*-models (INGENE and MODULE) to predict gene expression levels.

Transcriptomic coverage across models. Within the brain regions considered, the models demonstrated different predictive capabilities for a variable number of genes per region: CIS predicted between 5,579 and 9,354 genes per region, INGENE between 18,495 and 20,254 genes, and MODULE between 17,938 and 21,672 genes per region (Figure 8A). When pooling data across all regions, unique gene counts were 18,819 genes for CIS, 23,863 genes for INGENE, and 24,395 genes for MODULE (Figure 8B). Notably, the CIS model yielded the highest number of unique predictions (1,975 genes), i.e., not shared with *trans*-models, while INGENE and MODULE showed considerable overlap, predicting a common set of 23,384 genes, representing $> 95\%$ of their respective predictions.

Biological coherence of *trans*-based predictions. To further explore the regulatory landscape of these *trans*-shared predicted genes, we examined whether they shared regulatory components. Specifically, we considered instances where a SNP acted as a *trans*-eQTL for a MODULE-predicted gene and simultaneously served as a *cis*-eQTL of the co-expression

partners that INGENE uses to predict that same gene. We found that 15%-24% of genes predicted by MODULE across the six brain regions exhibited regulatory relationships (Table ST1).

Benchmarking against EpiXcan *cis*-epigenomic-informed models. Considering the relatively limited sample sizes for certain training regions (Table 1), we integrated the CIS model with the EpiXcan model (Zhang et al., 2019), which demonstrated superior performance in our LIBD training dataset over other PrediXcan family models (Table ST2). Focusing on the DLPFC, where EpiXcan was specifically trained, we compared its training performance with that of our CIS model. EpiXcan identified 13,529 genes that met the filtering criteria, in contrast to 9,354 genes by CIS. Among these, 5,857 genes were commonly predicted by both models, accounting for 43% of EpiXcan and 63% of CIS predictions (Figure 8C). While EpiXcan explained on average 14% variance in gene expression for these genes, CIS accounted for a lower, yet notable mean variance of ≈10% (Figure 8C, top-right panel). Interestingly, the CIS model outperformed EpiXcan in terms of variance explained for 30% of these predictions. Figure 8C (bottom) includes the top 50 genes with the largest differences in R^2 values between the two models. Considering the complementarity of CIS and EpiXcan predictions, we used both in subsequent analyses to enhance the identification of *cis*-predicted genes. Figure S1 presents a comparative overview of EpiXcan against all other models assessed.

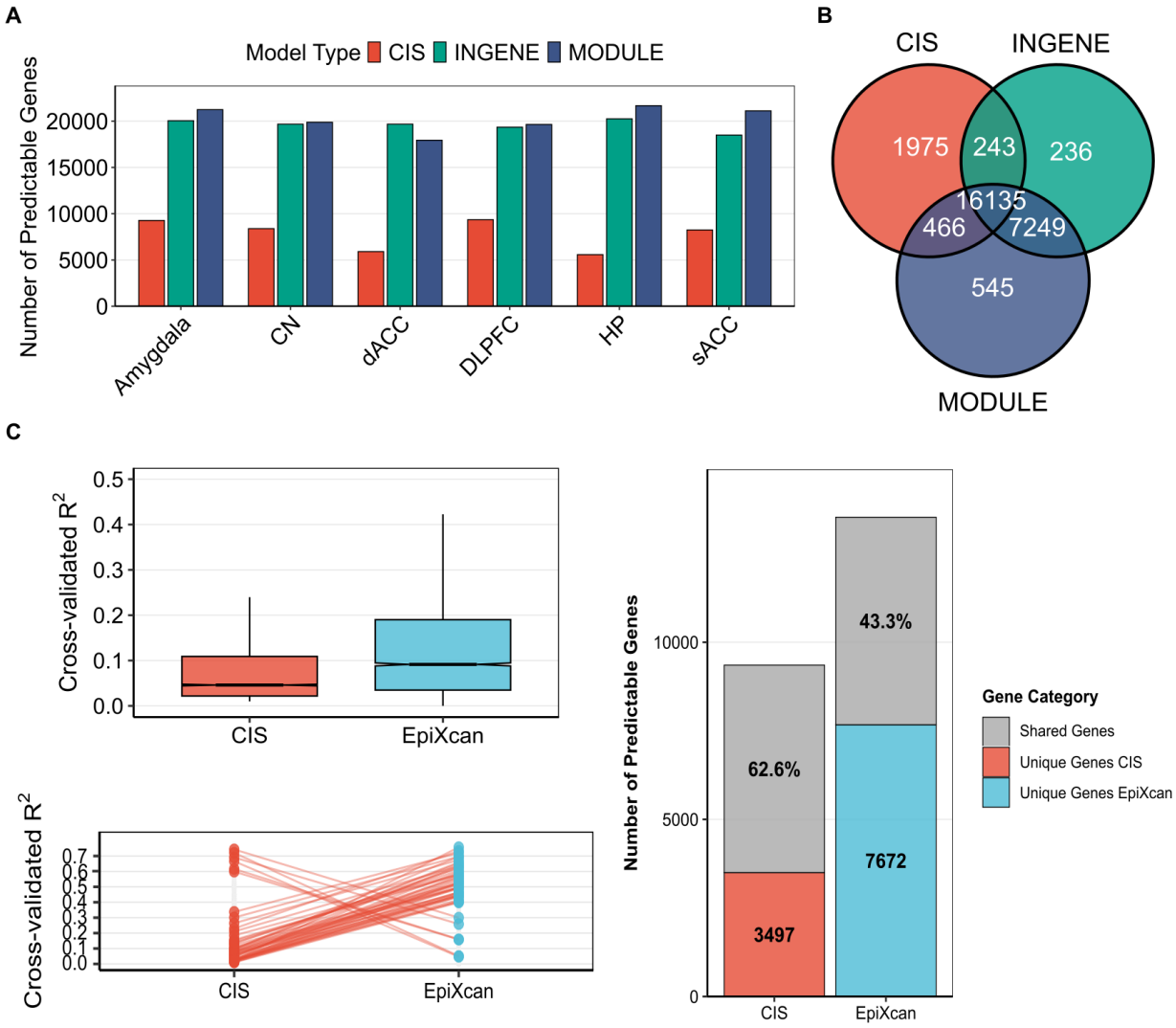

Figure 8: **Model performance in LIBD training data**. (A) Number of genes with cross-validated $R^2 \geq 0.01$ for CIS (red), INGENE (green), and MODULE (blue) across brain regions. (B) Overlap of all predicted genes across models. (C) Left: overlap and exclusivity of DLPFC-predicted genes for CIS (red) and EpiXcan (light blue). Right: R^2 comparison for the top 50 genes with the largest performance differences.

Model comparison summary. In summary, INGENE and MODULE exhibited higher prediction capabilities in terms of the number of imputable genes compared to CIS (Figure 8A-B) and EpiXcan (Figure S1), and both outperformed the CIS model when evaluating the variance explained for commonly predicted genes in the training step. MODULE emerged as the most effective model among those evaluated, featuring the highest number of imputable

genes and the most substantial variance explained during cross-validation within the training dataset (refer to Table 2 for comparison among models). It is important to note, however, that comparisons based solely on the training set might be susceptible to overfitting.

Model	Total Genes	Region	Mean CV R^2	Key Features
CIS	18,819	Amygdala CN	0.10	Baseline *cis*-eQTL model
INGENE	23,863	dACC DLPFC	0.035	*cis*-eQTLs of co-expressed partners
MODULE	24,395	HP sACC	0.040	co-eQTLs of co-expression module PC1
EpiXcan	13,529	DLPFC	0.14	*cis*-eQTLs informed with epigenomic priors

Table 2: **Comparison of gene expression prediction models**. In-house models (CIS, INGENE, MODULE) were trained across six brain regions in the LIBD dataset. EpiXcan was trained externally (Zhang et al., 2019) and evaluated in DLPFC only. *Note*: "Total Genes" reflects the number of genes predicted with $R^2_{\text{CV}} \geq 0.01$. Mean CV R^2 is averaged across all predicted genes in the training data.

3.4.2 Evaluation of *cis*- and *trans*-Models in Independent Testing Datasets

Validation framework. Given the different training origins—CIS, INGENE and MODULE were trained on the LIBD datasets, whereas EpiXcan was originally trained in CMC data—we assessed the replication of all four models in the GTEx-independent cohort, encompassing the same six brain tissues (Table 1) we used to train CIS, INGENE and MODULE. We applied each model to GTEx genotype data to generate predictions, this time ruling out any potential overfitting as they were applied to the testing dataset. Our assessment of predictions was based on two criteria: (i) the number of genes with positive Pearson correlation ($r > 0$) between predicted and observed expression, and (ii) the number of genes with adjusted $R^2 > 0$, indicating significant explained variance.

Transcriptome-wide prediction performance. In the GTEx dataset, the *trans*-based models consistently predicted a higher number of genes compared to the *cis*-based models. CIS predicted between 1,775 and 4,962 genes across brain regions, and EpiXcan predicted between 5,704 and 6,840 genes. In contrast, INGENE predicted 15,496 to 16,870 genes, and MODULE predicted between 12,564 and 14,690 (Figure 9A). The relative gain in coverage was substantial: INGENE predicted 3.23–9.50 fold more genes than CIS and 2.44–2.90 fold more than EpiXcan (Figure S2).

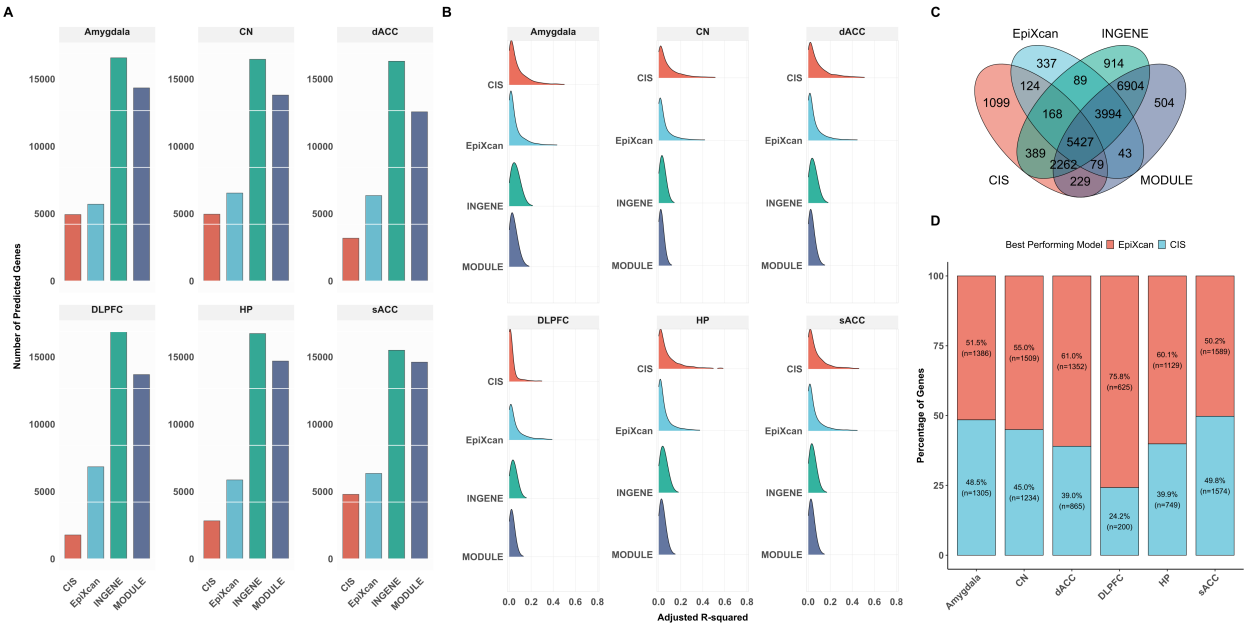

Figure 9: **Performance of gene expression prediction models in the independent GTEx dataset.** (A) Number of predicted genes per model across brain regions. (B) Distribution of adjusted R^2 values per model. (C) Venn diagram showing overlap in predicted genes across models. (D) Relative performance of CIS and EpiXcan across shared genes in each brain region.

Variance explained across models. As expected, due to the generally larger impact of *cis* over *trans* effects on gene expression variance, the CIS and EpiXcan models explained a higher percentage of variance for the fewer genes they predicted compared to the INGENE and MODULE models. Mean adjusted R^2 values for CIS ranged from 3% to 10%, while EpiXcan achieved between 6% and 7%, in line with prior studies (Zhang et al., 2019). In comparison, INGENE explained between 3% and 5% of variance, and MODULE between

2.4% and 4% (Figure 9B). This trade-off between precision and breadth was consistent across all brain regions.

Wilcoxon rank-sum tests comparing adjusted R^2 distributions across models confirmed EpiXcan's significantly greater predictive power relative to both CIS and *trans*-based models (all $p < 2.2 \times 10^{-16}$). This trend also held within gene sets commonly predicted by all models (Figure S3).

Model complementarity and *cis*-model selection. Each model contributed unique predictions. Notably, CIS captured the largest number of genes not predicted by other models (Figure 9C). For genes jointly predicted by CIS and EpiXcan, we compared per-gene adjusted R^2 values to identify the better-performing *cis*-model across tissues. EpiXcan generally outperformed CIS in DLPFC, which is consistent with the origin of its training data (Figure 9D). For downstream analyses, we retained the best-performing model (CIS or EpiXcan) on a gene-by-gene basis to represent *cis*-regulatory effects.

Replication in the CMC. To assess the replicability of our *cis*- and *trans*-predictions beyond the GTEx cohort, we extended our analysis by applying our models to postmortem datasets of DLPFC and ACC from the CMC (Fromer et al., 2016). We observed consistent gene-level prediction patterns between CMC and GTEx for commonly imputed genes. The CIS model predicted between 1,400 and 2,320 genes across both cohorts, with median Pearson correlation coefficients ranging from $r = 0.33$ (ACC) to $r = 0.43$ (DLPFC). The INGENE model yielded between 2,318 and 3,106 common gene predictions, with median correlations from $r = 0.10$ (ACC) to $r = 0.47$ (DLPFC). Finally, the MODULE model predicted 11,043 to 11,729 genes across tissues, achieving a median correlation of $\approx r = 0.18$ (Figure S4).

Summary. Overall, these results demonstrate that *trans*-based models (INGENE and MODULE) considerably expand the number of replicable gene expression predictions in

external datasets. While *cis*-based models yield higher per-gene accuracy, the *trans*-based approaches offer broader transcriptome-wide generalizability—a feature especially valuable in downstream association studies.

3.4.3 Functional Genetics of co(expression)-eQTLs

To further investigate SNPs exhibiting both *cis*- and *trans*-regulatory effects among validated predictions in GTEx (refer to Results section 3.4.2), we focused on MODULE-derived *trans*-SNPs that were also significant *cis*-eQTLs across 49 tissues in the GTEx v8 dataset (GTEx, 2020) (Table 3). Our analysis revealed that ≈40% of MODULE *trans*-SNPs were also GTEx *cis*-eQTLs, associated with between 5,821 and 19,276 predicted *cis*-regulated genes (eGenes) across tissues.

Region	**No. *trans*-SNPs as GTEx *cis*-eQTLs**	**Percentage of *trans*-SNPs as GTEx *cis*-eQTLs**	**N° GTEx *cis*-eGenes**
Amygdala	21,549	29%	10,314
CN	22,791	30%	11,347
dACC	32,027	42%	19,276
DLPFC	11,595	25%	5,821
HP	26,306	36%	12,834
sACC	25,778	44%	17,896

Table 3: **Summary of MODULE-derived *trans*-SNPs that also act as GTEx *cis*-eQTLs across brain regions.** Values reflect the number and proportion of overlapping SNPs and their associated GTEx *cis*-regulated genes (eGenes).

Gene Ontology enrichment analysis of GTEx-imputed *cis*-eGenes revealed significant overrepresentation of molecular functions related to ATP-dependent activity, catalytic and electron transferase functions, and protein binding activities, including MHC class II receptor binding and cadherin binding (false discovery rate, FDR < 0.05; Figure S5).

To assess transcription factor (TF) enrichment, we performed a regulomic analysis. This revealed 252 TFs that were significantly overrepresented among GTEx *cis*-eGenes linked to

MODULE *trans*-SNPs (Bonferroni-adjusted $p < 0.05$; Figure 10).

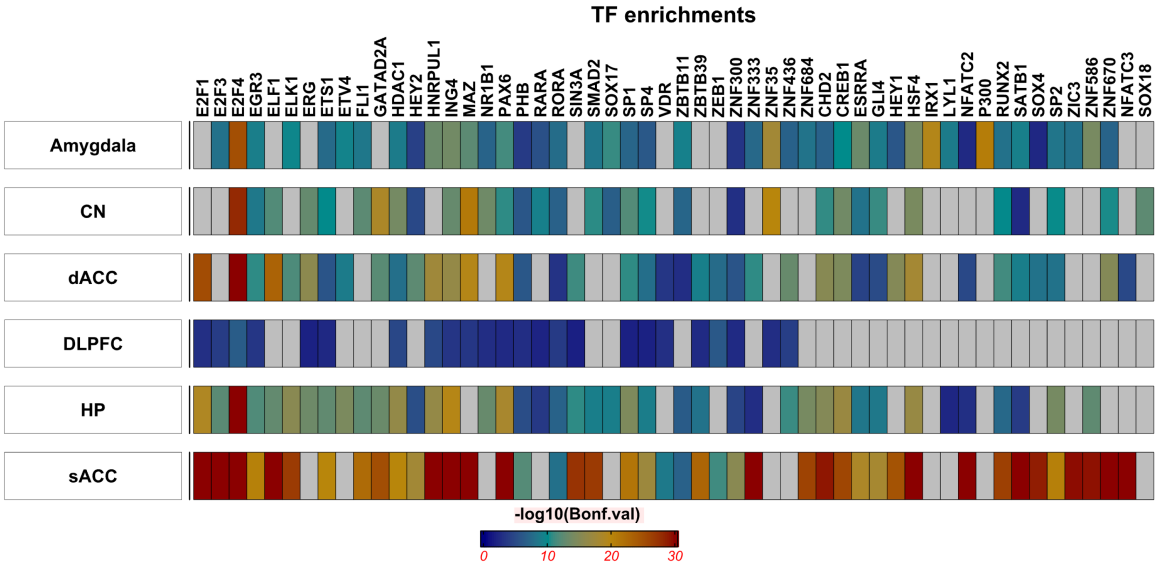

Figure 10: **Regulome Enrichment Analysis of GTEx *cis*-eGenes for MODULE *trans*-eQTLs.** Enrichment for TFs across brain regions. To generate this visualization, we identified the top 20 most significant TFs for each brain region and assessed their overrepresentation. A grey block in the figure denotes that TF is not significantly overrepresented in that region.

3.4.4 Combining *cis* and *trans* Predictions Enhances Gene Expression Modeling

Testing *cis*-*trans* integration in GTEx. We hypothesized that integrating *cis*- and *trans*-based predictors would maximize both the number of predictable genes and the variance explained in gene expression. To evaluate this, we performed gene-level significance testing in GTEx using MLE at a significance threshold of $\alpha = 0.05$ (see Methods section 3.3.4 for details). Specifically, we compared a full model— including covariates and both *cis* and *trans* predictors— with a reduced model containing only covariates and *cis* predictors. Genes for which the inclusion of *trans* predictors led to a significant improvement in adjusted R^2 were classified as *trans*-enhanced.

Our analysis revealed a substantial increase in gene predictability when combining *cis*,

cis + *trans*, and *trans*-only models across brain regions (Figure 11A). In particular, the addition of the *trans* component in the *cis* + *trans* model consistently improved the variance explained relative to the *cis*-only model, as indicated by one-tailed Wilcoxon signed-rank tests ($p < 0.001$; Figure 11A). When aggregating predictions across all brain regions, we identified a total of 19,802 genes that were significantly predictable by all three model types (Figure 11A).

Unified modeling approach. Having established that the inclusion of *trans* components can enhance *cis*-based gene expression predictions for a subset of genes, we next evaluated whether integrating *cis* and *trans* predictors into a unified model would further improve gene-level prediction performance. The goal was to leverage the complementary strengths of each predictor type. Importantly, *cis* SNPs were excluded when generating *trans* predictions, ensuring that the predictor sets were non-overlapping and statistically independent.

For genes with at least two available prediction models, we constructed a linear model in the GTEx dataset (serving as the testing cohort), using observed gene expression as the dependent variable and a combination of CIS, EpiXcan, INGENE, and MODULE predictions as independent variables. The performance of this combined model was subsequently evaluated in the CMC dataset, focusing on the dACC and sACC brain regions. The DLPFC region was excluded to prevent potential information leakage, given that EpiXcan was originally trained on CMC DLPFC data.

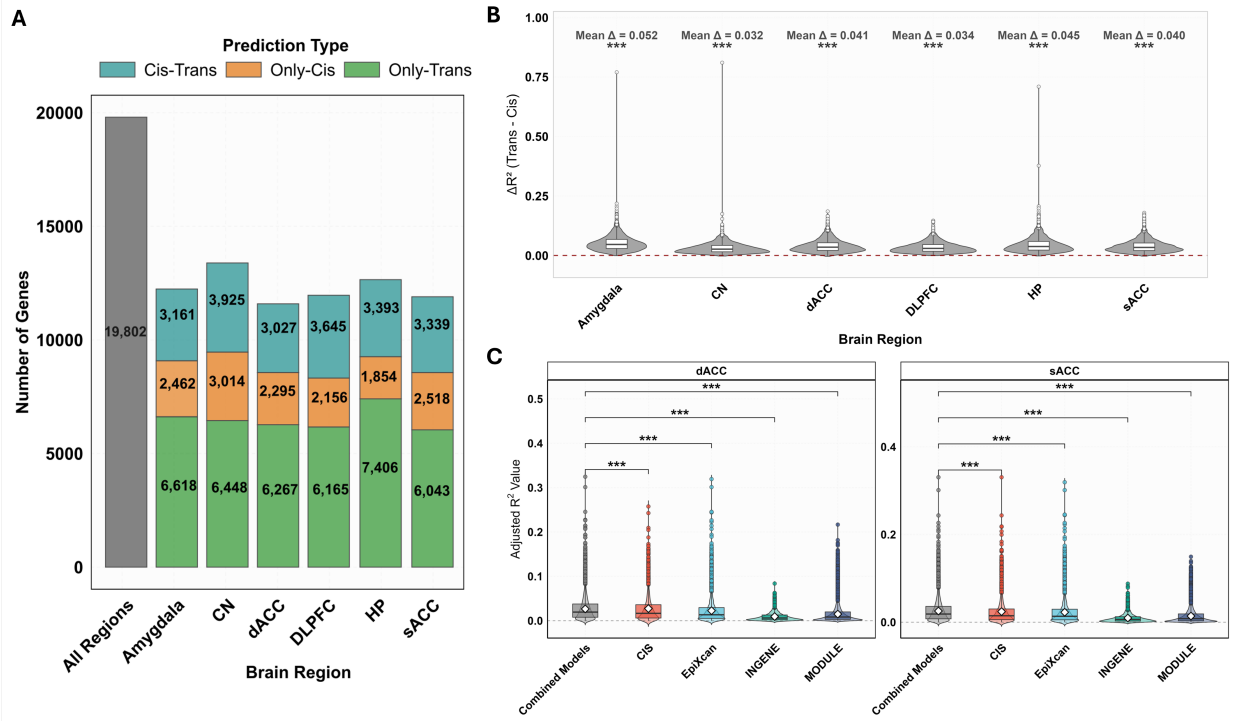

Figure 11: **Cis-trans integration improves gene prediction performance.** (A) Number of genes significantly predicted across brain regions using only-cis (orange), only-trans (green), and combined cis-trans (turquoise) models. Grey bars indicate pooled predictions across regions. (B) Distribution of adjusted R^2 improvements (ΔR^2) from cis-trans versus cis-only models in GTEx. Asterisks (***) denote $p \leq 0.001$. (C) Comparison of model performance in CMC dACC and sACC datasets. Combined models (grey) outperform CIS (red), EpiXcan (light blue), INGENE (green), and MODULE (blue). Asterisks (***) denote $p \leq 0.001$ by Mann–Whitney tests.

Combined models outperform individual predictors in CMC. The resulting "Combined Models" (Figure 11C) explained ≈2.7% of the variance in gene expression across 10,748–11,729 genes in dACC and sACC. In comparison, the CIS model accounted for 2.4–2.8% of variance in 2,238–3,192 genes, EpiXcan explained 2.3% across 5,038 genes, INGENE explained ≈1.1% in 2,318–2,498 genes, and MODULE explained 1.5% in ≈12,000 genes. Mann–Whitney rank-sum tests confirmed that, based on mean R^2, the Combined Models significantly outperformed each of the individual models in both dACC and sACC ($p \leq 0.001$; Figure 11C).

Summary. Together, these results support the integration of *trans*-regulatory information alongside *cis*-prediction scores to enhance gene expression modeling. Combining models not only increases transcriptome-wide prediction coverage but also improves prediction accuracy across cohorts.

3.5 Discussion

This study developed gene expression prediction models across six brain regions by leveraging both *cis*- and *trans*-regulatory information embedded within gene co-expression networks. Rather than relying solely on local regulatory variation, we demonstrated that incorporating structured *trans*-variants substantially expands the scope and biological interpretability of genetically regulated expression models.

Integrating *Cis* and *Trans* Predictors Enhances Transcriptomic Coverage. Our findings reveal that traditional *cis*-only models, while effective at capturing proximal regulatory signals, substantially underestimate the heritable component of gene expression variance (GTEx, 2017; Yao et al., 2020). Through co-expression-guided strategies—INGENE and MODULE—we increased the number of accurately predictable genes by ≈50% compared to *cis*-only frameworks such as EpiXcan (Zhang et al., 2019) and the CIS model. This expansion underscores the importance of distal regulatory mechanisms in shaping the human brain transcriptome, consistent with observations from large-scale transcriptomic efforts (Battle et al., 2014; Hartl et al., 2021).

Importantly, while *cis*-based models naturally exhibited higher predictive effect sizes, the integration of *trans*-signals provided critical breadth without sacrificing overall model accuracy. This broader coverage enhances the downstream potential for association studies, especially for polygenic traits influenced by distributed regulatory architectures.

Mechanistic Insights into Brain Gene Regulation. The structure of our *trans*-predictive models offers new biological insights. Approximately 24% of MODULE-predictable genes were explainable through indirect *cis*-mediated pathways identified via INGENE predictions, supporting prior hypotheses that co-expression hubs mediate distal regulatory effects (Borcuk et al., 2024). Furthermore, the enrichment of known transcriptional regulators—such as *GATAD2A*, *RERE*, and *SP4*—among MODULE-derived *trans*-eGenes aligns with established findings that transcription factors serve as critical mediators of complex gene regulation in the brain (Rodriguez-López et al., 2020; Võsa et al., 2021).

However, the majority of MODULE-predicted genes lacked identifiable *cis*-indirect paths, suggesting the existence of diffuse, multilayered regulation potentially driven by uncharacterized factors, developmental context, or feedback loops (GTEx, 2020; van der Wijst et al., 2018). These observations highlight the current limits of available co-expression resources and the need for dynamic, condition-specific regulatory maps.

Systems-Level Approaches Improve Biological Interpretability. Our findings align with growing evidence that network-based strategies offer essential advantages over single-variant or single-gene models for mapping complex traits (Borcuk et al., 2024; Pergola et al., 2023c,b). By leveraging co-expression architecture, we reduced the statistical noise inherent in *trans*-eQTL discovery, captured distributed regulatory effects, and improved biological interpretability. This systems-level approach is particularly valuable for studying disorders like schizophrenia, where risk loci often converge within co-regulated modules rather than isolated genes (Hartl et al., 2021; Pergola et al., 2023c).

Nevertheless, variability in model performance across brain regions highlights the critical role of tissue specificity in regulatory architecture (GTEx, 2017; Panagiotakos and Pasca, 2022). This variability underscores the need for region-specific modeling frameworks when targeting developmental or neuropsychiatric phenotypes.

Strengths and Limitations A major strength of our modeling approach lies in its ability to integrate distal regulatory effects while preserving biological coherence. Unlike genome-wide *trans*-eQTL mapping, which suffers from an extreme multiple-testing burden, our co-expression-informed strategies focus on biologically plausible variant sets, improving both statistical power and interpretability.

However, limitations persist. Our reliance on static co-expression networks, derived from postmortem tissue, may fail to capture developmental-stage-specific or cell-type-specific regulation critical for neurodevelopmental processes (Gandal et al., 2018a; van der Wijst et al., 2018; Pergola et al., 2023a). Additionally, although we residualized gene expression data for age and clinical status to mitigate confounding in the LIBD training set, the demographic composition of LIBD—predominantly individuals of EUR ancestry and limited age ranges—may restrict the generalizability of the trained models to more diverse populations.

Future research should prioritize integrating single-cell transcriptomic and epigenomic data to refine *trans*-regulatory models further, with particular attention to developmental and cellular context. Combining genetically regulated expression with multi-omic regulatory features may yield richer, more dynamic predictions of disease-relevant gene expression changes.

Finally, it is important to acknowledge that current resources remain heavily Eurocentric, and our models are trained exclusively on European-ancestry cohorts. While this reduces population heterogeneity, it limits portability across ancestries. Dedicated multi-ancestry resources and empirical validation will be required to address this limitation, a point revisited in the General Discussion (Section 6.4).

Future Directions: Cross-Dataset Training and Model Generalizability. A natural continuation of this work will be to expand the current modeling framework toward **cross-dataset training and validation** of the INGENE and MODULE algorithms. While the present study focused on models trained on the LIBD dataset and externally validated in

GTEx and CMC, future analyses will implement a **multi-cohort training and replication design** encompassing all available postmortem transcriptomic resources. This extension will enable a systematic assessment of the **stability, generalizability, and transferability** of co-expression–guided models across independently collected brain datasets.

In practice, the extended pipeline will begin by **harmonizing RNA-seq and genotype preprocessing** across datasets to ensure comparable alignment, normalization, and covariate correction. Predictive models will then be **trained independently within each dataset** (LIBD, GTEx, and CMC) and **evaluated reciprocally across all possible dataset pairs**. For each gene, the *cross-validated adjusted* R^2 obtained during training will guide model inclusion, while performance will be evaluated using complementary metrics: (i) the number of genes successfully predicted across datasets, (ii) the overlap of predicted gene sets between cohorts, and (iii) the variation in performance as a function of training sample size. Together, these measures will provide a quantitative assessment of model reproducibility and help identify potential sources of false positives or dataset-specific effects.

The second component of the pipeline will examine **cross-cohort consistency in model-derived effect sizes**. For each gene, expression values predicted by models trained on LIBD, GTEx, and CMC will be **correlated within the same set of LIBD individuals**, thereby isolating true biological concordance from sample-size or technical biases. Genes displaying strong and positive cross-dataset correlations will be prioritized as **replicable *trans*-regulated targets**.

By completing this multi-cohort training and validation framework, future work will provide a rigorous assessment of the extent to which co-expression–based prediction models generalize across brain transcriptomic resources. These planned extensions represent an important methodological advancement toward **standardized, cross-cohort evaluation of trans-regulatory gene prediction** and will substantially enhance the interpretability, reproducibility, and biological reliability of network-based GReX frameworks in subsequent research.

Conclusion

By integrating *cis*- and *trans*-regulatory information through co-expression-informed models, this study substantially advances the predictive landscape of brain gene expression. Our findings highlight the necessity of system-level approaches for understanding the genetic architecture of complex brain traits and lay the foundation for more powerful and interpretable transcriptome-wide association studies.

Chapter 4

Study 2: Co-expression TWAS in PGC3 SCZ Cohorts

A revised version of this chapter is currently under revision for publication in *Nature Genetics*: Rossi F., et al. *Co-expression-based models improve eQTL predictions and highlight many novel transcriptome-wide genes associated with schizophrenia.*

4.1 Introduction

SCZ is a severe, chronic psychiatric syndrome with an average lifetime prevalence of almost 1% and an average life expectancy that can be reduced by as much as 15 years compared to the general population (Perälä et al., 2007). Treatment response in SCZ is another variable aspect of the disorder and a critical piece of evidence of plausible neurobiological heterogeneity reflected at the clinical level. Currently, antipsychotics are the gold standard for the treatment of SCZ, acting as modulators of D2 dopamine receptor activity via antagonism or partial agonism (Miyamoto et al., 2012). However, one-third of patients do not satisfactorily respond to currently available therapies, and even though many patients achieve symptomatic remission after their first psychotic episode, up to 80% of patients experience a psychotic relapse within five years after remission of their first episode (Buckley and Miller,

2017; Leucht et al., 2022). This motivates the development of translational approaches that connect statistical genetic findings with underlying biological mechanisms.

Despite high heritability estimates (60–80%) (Sullivan et al., 2003; Purcell et al., 2009), much of the genetic architecture remains functionally unexplained. While GWAS have provided critical insights into the polygenic nature of SCZ, they offer limited functional interpretability. Larger sample sizes have incrementally increased discovery power (Allen et al., 2008; Ripke et al., 2014), but the proportion of explained variance remains modest (Trubetskoy et al., 2022). Thus, advancing beyond GWAS requires novel methods that can translate statistical signals into biological mechanisms (Tam et al., 2019; Mostafavi et al., 2023).

Gene Expression as a Functional Intermediate. Gene expression serves as a functional intermediary between DNA sequence and disease phenotype, making mRNA a proximal readout of non-coding genetic risk (Mostafavi et al., 2023). Integrating GWAS summary statistics with transcriptomic data from postmortem human brain tissue has emerged as a powerful approach in functional genomics (Gandal et al., 2016). eQTLs are key to this integration (Nicolae et al., 2010; Gamazon et al., 2015; Zhu et al., 2016; Huckins et al., 2019; Zhang et al., 2019), yet most GWAS SCZ hits do not colocalize with known eQTLs (Chun et al., 2017; **?**; Connally et al., 2022; Mostafavi et al., 2023). This unexplored variance attributable to genetics suggests that much of the genetic risk uncovered by GWAS may act through regulatory mechanisms not captured by current eQTL datasets, underscoring the need to explore alternative, noncanonical pathways of gene regulation.

Limitations of *cis*-Based TWAS Models. TWAS prioritize disease-associated genes by leveraging GReX models (Gusev et al., 2016). Conventional TWAS approaches, such as PrediXcan (Gamazon et al., 2015), are largely confined to *cis*-eQTLs—variants located within 1 Mb of their target gene. While valuable, *cis*-eQTLs explain only a fraction of expression heritability (Purcell et al., 2009; Yao et al., 2020), limiting the scope of gene discovery (Liu et al., 2019). Subtle and distal regulatory effects may arise indirectly through

the *cis*-eQTLs of co-expressed partners or directly via long-range interactions with regulatory elements located at great genomic distances (Friston et al., 2016; Boyle et al., 2017; Mostafavi et al., 2023), reflecting a more distributed architecture of gene expression regulation. These genetic factors are the so-called *trans*-eQTLs and are estimated to account for up to 70% of the heritable variability in SCZ expression (Battle et al., 2014; Pierce et al., 2014; Liu et al., 2019, 2022). Despite their importance, their detection remains statistically challenging due to weaker effect sizes and the heavy multiple-testing burden (Gamazon et al., 2015; Umans et al., 2021).

The Role of Gene Networks in SCZ Risk. Rather than acting in isolation, genetic risk variants for SCZ tend to converge within biologically coherent gene networks. Across multiple studies, SCZ-associated loci have been shown to cluster within co-expressed modules active during specific developmental windows, brain regions, and cell types (Fromer et al., 2016; Gandal et al., 2018a; Radulescu et al., 2020; Ramaswami et al., 2020; Hartl et al., 2021; Panagiotakos and Pasca, 2022; Pergola et al., 2023a; Jaffe et al., 2020; Cameron et al., 2023). This non-random aggregation suggests that the genetic architecture of SCZ reflects coordinated perturbations of regulatory programs rather than independent variant effects (Borcuk et al., 2024).

Co-expression Networks as a Framework for Mapping *trans*-Regulatory Architecture. Co-expression networks provide a biologically motivated framework to capture this distributed risk structure and to constrain the otherwise vast search space for *trans*-eQTL discovery. By grouping genes with correlated expression patterns, these networks model shared regulatory influences that often transcend local genomic proximity. This is particularly critical for SCZ, where many risk loci may exert distal regulatory effects that are invisible to conventional *cis*-eQTL-based approaches (Liu et al., 2019). Aggregating genes into modules enhances statistical power and enables the detection of subtle, network-mediated genetic effects (Pergola et al., 2023b).

From Networks to Predictive Models. Building on these observations, co-expression-based models have been developed to infer upstream regulators—including transcription factors and microRNAs—that shape network behaviour (Chen et al., 2018; Pergola et al., 2017, 2023c). Validation from *in vitro* studies further supports the functional relevance of these network-driven predictions (Torretta et al., 2020). Several studies have leveraged module structure to generate co-eQTL-informed polygenic scores and predict complex brain phenotypes, such as PET imaging responses and functional connectivity patterns (Nath et al., 2017; Fazio et al., 2018; Kolberg et al., 2020; Sportelli et al., 2024). These approaches capitalize on the hypothesis that co-expression modules encode latent regulatory programs, offering a scalable path toward integrating *trans*-regulatory information into trait mapping.

The Power of Large-Scale Cohorts. Despite their promise, mapping *trans*-regulatory architecture and validating network-based models requires very large cohorts. As mentioned above, distal regulatory effects typically have small effect sizes and are sensitive to context, making them statistically underpowered in modest samples. Unlocking the full potential of co-expression-informed mapping thus demands access to harmonized, large-scale genetic resources that can power the discovery of distributed and subtle genetic signals.

The PGC is an international collaboration established to advance the understanding of the genetic basis of psychiatric disorders.The third wave of PGC3 provides an unprecedented opportunity to address the challenges discussed above. With genotype data from 168,431 individuals across 90 independent SCZ cohorts (Trubetskoy et al., 2022), PGC3 offers both the scale and heterogeneity across individuals needed to implement network-informed transcriptomic prediction at a population level. Integrating predictive models trained on postmortem brain data with this extensive living cohort allows for systematic testing of how co-expression architecture and *trans*-regulatory effects contribute to clinical phenotypes.

Preview of Results. Whereas Study 1 (Chapter 3) focused on the development and validation of *cis* and *trans*-prediction models using postmortem brain transcriptomic data, the

present study applies these models to large-scale genetic datasets from living individuals from the PGC3. This approach, which we define as Co-expression TWAS (coTWAS), integrates both local and distal regulatory information into transcriptomic imputation. Specifically, we imputed gene expression using CIS, EpiXcan, INGENE, and MODULE predictors across 62 SCZ cohorts of EUR Ancestry of the PGC3 (Trubetskoy et al., 2022), generating both *cis*- and *trans*-based expression estimates. These predictions were then used to perform gene-wise association testing with SCZ diagnosis across 102,613 individuals (see Table ST5 in the appendix). Figure 12 summarizes the key analytical steps of this study.

The coTWAS approach identified 1,764 SCZ-associated genes at FDR < 0.01, of which 1,515 were novel relative to prior TWAS efforts (Gusev et al., 2018; Gandal et al., 2018a; Collado-Torres et al., 2019; Huckins et al., 2019; Hall et al., 2020) (refer to Table 8 in the appendix for TWAS details). Notably, across brain regions, 50% to 58% of the associations were supported exclusively by *trans*-based predictions, highlighting the unique discovery power of co-expression-informed models. When restricting the analysis to genes with a *cis*-predictor component, we identified 619 genes across all regions, of which 437 represent novel associations not previously captured by *cis*-focused TWAS approaches. Enrichment analyses of significant coTWAS genes pointed to critical biological pathways, including synaptic transmission, immune signaling, and cell adhesion.

The findings in this chapter illustrate that integrating *trans*-regulatory architecture through co-expression modeling expands the transcriptomic landscape of SCZ and enhances the mechanistic resolution of gene-trait mapping. As a generalizable framework, the coTWAS expands the methodological toolkit for functional genomic discovery in psychiatric research.

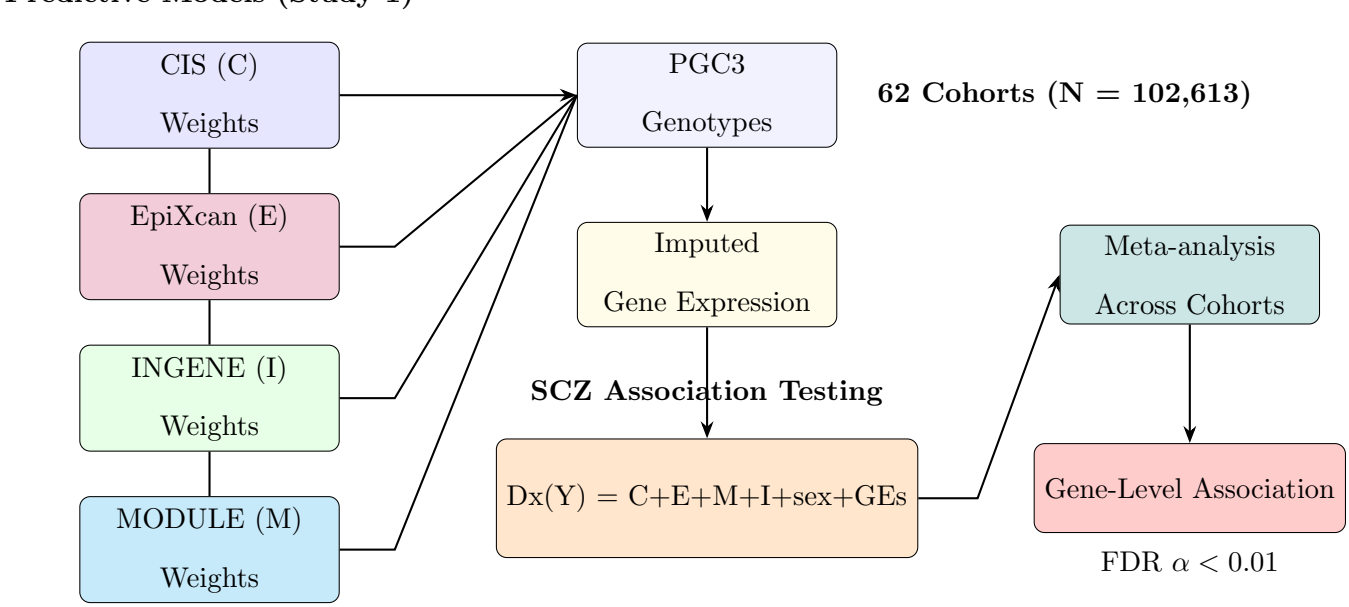

Figure 12: **Study 2 pipeline: From predictive modeling to gene-level SCZ association.** Predictive weights from four models—CIS, EpiXcan, INGENE, and MODULE—trained and validated on postmortem brain data in Study 1 (Chapter 3) were applied to genotype data from 62 PGC3 cohorts. Gene expression was imputed and combined according to the integration strategy described in Chapter 3, Section 3.3.5. Logistic regression analyses were conducted separately within each cohort, adjusting for sex and genomic eigenvectors (GEs). A meta-analysis across cohorts was then performed to estimate gene-level associations with SCZ diagnosis, with statistical significance defined at a Benjamini-Hochberg FDR threshold of 0.01.

4.2 Data

4.2.1 PGC3 SCZ Cohorts

In this study, we analyzed individual-level data from 62 cohorts of European ancestry, totaling 102,613 participants (Table ST5, Figure 13). These cohorts included both SCZ cases and healthy controls, allowing a large-scale and well-powered investigation of genetic associations with SCZ risk. The genotype and phenotype data from the PGC3 were accessed under controlled conditions. Data access was granted following approval by the PGC Data Access Committee, and all researchers were required to adhere to the consortium's data use

agreement. Full information on access procedures and requirements is available at: https://pgc.unc.edu/for-researchers/data-access-committee/data-access-information/.

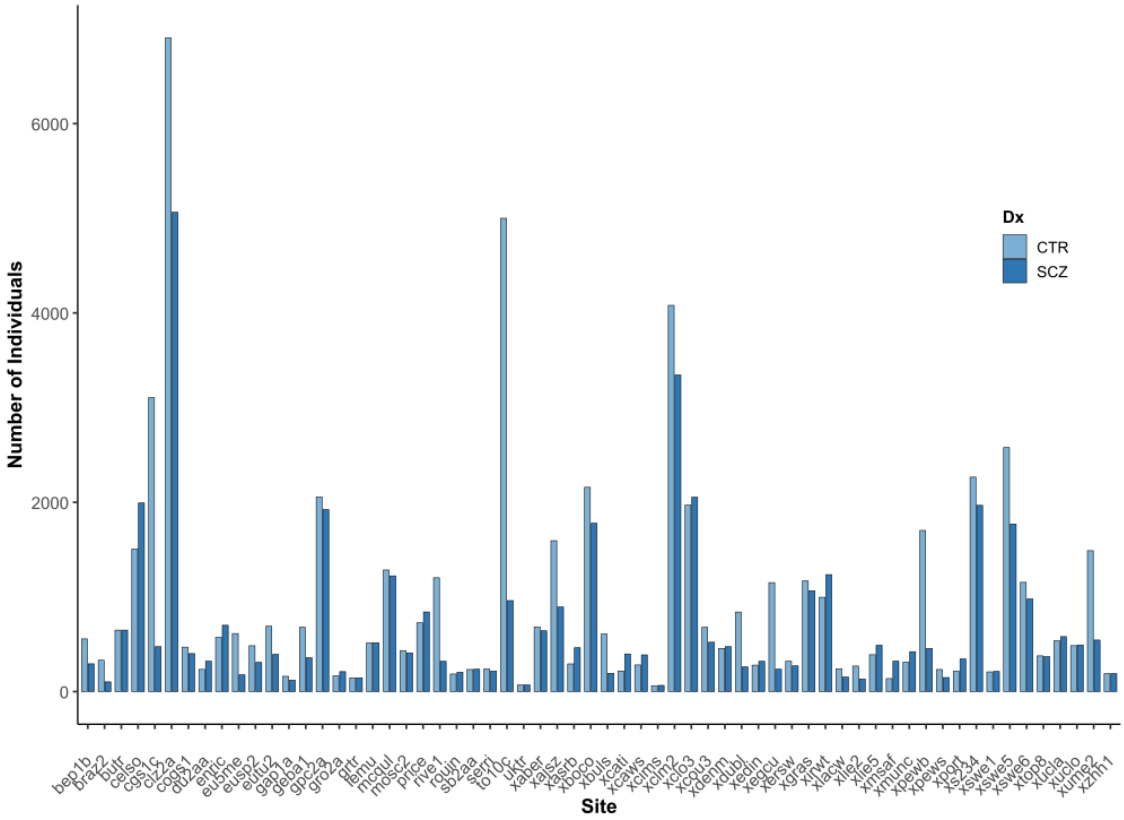

Figure 13: **Distribution of Diagnosis across PGC3 SCZ Cohorts**

Ethical Considerations. All participating cohorts adhered to ethical guidelines and obtained necessary approvals from their respective institutional review boards. Informed consent was secured from all participants, ensuring compliance with ethical standards for human subjects' research. The collaborative nature of the PGC emphasizes transparency and ethical responsibility in genetic research. More details of ethical approval protocols are described in the Supplementary Cohort Descriptions from Trubetskoy et al. (2022) (Trubetskoy et al., 2022).

4.3 Methods

4.3.1 PGC3 Cohort Genotype Preprocessing

Genotype data acquisition, quality control, imputation, and computation of genomic eigenvariates (GEs) for population stratification were performed independently for each cohort separately as previously reported by the PGC (Trubetskoy et al., 2022).

SNPs were filtered within each cohort using the following criteria: imputation quality score INFO > 0.9, missingness $< 1\%$, minor allele frequency (MAF) > 0.01, and Hardy-Weinberg equilibrium $P > 10^{-6}$. SNP filtering was performed using PLINK v1.09 (Purcell et al., 2007).

4.3.2 Correlation Between SNP Weights and PGC3 Odd Ratios

We downloaded GWAS summary statistics from the PGC for SCZ (https://figshare.com/articles/dataset/scz2022/19426775?file=34865091) to obtain SNP-level odds ratios (ORs). SNP weights were extracted from each of the predictive models—CIS, EpiXcan (*cis*-models), and MODULE (*trans*-model)—that were developed, trained, and evaluated in Study 1 (Chapter 3) using postmortem brain transcriptomes. For each SNP, we computed the mean absolute weight across all genes that were previously validated in the GTEx according to the criteria established in Study 1 (refer to Chapter 3 section 3.4.2). This ensured that our downstream analyses were grounded in models with robust predictive performance, anchored to biologically meaningful expression profiles derived from human brain tissue.

We subset the PGC summary statistics to only include SNPs with a nominally significant association with the diagnosis (p <0.05). We further selected SNPs present in the *cis*- or *trans*-model independently (Table ST4). We evaluated the correlation of the weights with log(OR) values using Pearson correlation coefficients, with statistical significance determined by two-tailed tests. To stabilize the variance of correlation coefficients, we employed Fisher's

Z-transformation (Fisher, 1915). First, we calculated the Fisher's Z-value for CIS, EpiXcan, and MODULE models. Next, to account for different numbers of SNPs in the models, we computed the standard error of each Fisher's Z-value and quantified the difference computing the test statistic Z as:

$$Z = \frac{|Z_1 - Z_2|}{\sqrt{SE_1^2 + SE_2^2}}$$

where Z_1 and Z_2 are Fisher-transformed correlation coefficients, and SE denotes the standard error. Finally, we derived the two-tailed p-value from the standard normal distribution to assess the statistical significance of the difference. This approach allowed us to robustly compare correlation coefficients, accounting for inherent variability in sample sizes and ensuring the reliability of our findings.

4.3.3 Connectivity trend analysis across PGC-weight quintiles with permutation-based null

We quantified whether genes more heavily influenced by PGC3 SNPs exhibit greater connectivity to SCZ risk genes. For CIS, EpiXcan, MODULE models we:

- Computed a PGC-weight metric per gene as the combined ratio of absolute model weights carried by PGC3 SNPs ($p < .05$) relative to all SNPs, scaled by the square root of the PGC-SNP proportion;
- Binned genes into quintiles of this metric (equal-sized bins);
- Calculated mean connectivity to PGC risk genes within each quintile, producing a 5-point trajectory per modelregion;

We assessed monotonicity using two statistics across quintile index: a linear trend slope and Spearman's ρ. To obtain a permutation-based null, we performed gene-level permutations (n=1,000) by randomly reassigning genes to quintiles, re-aggregating mean connectivity

per quintile, and recomputing both statistics. Empirical two-sided p-values were defined as the proportion of permuted statistics whose absolute value equaled or exceeded the observed statistic.

4.3.4 coTWAS Analysis in PGC3 Cohorts

Gene Expression Imputation and Selection

We imputed gene expression in each PGC3 site using CIS (C), EpiXcan (E), INGENE (I), and MODULE (M) models—all developed and benchmarked in Study 1 (Chapter 3). When multiple co-expression networks contributed to the prediction of the same gene within the INGENE and MODULE models, we first averaged predictions across networks to derive a model-specific expression estimate (see Chapter 3 section 3.3.3).

For downstream analysis, we retained only those genes that passed predefined validation criteria in the GTEx dataset, namely adjusted $R^2 > 0$ and Pearson's $r > 0$, as established in Study 1 (Section 3.4.2). Genes predicted by a single model were defined as *unimodal*. For genes predicted by multiple models, we applied the linear combination strategy introduced in Study 1 (Section 3.3.5) to generate a unified expression prediction. These were designated as *multimodal* genes and represent a biologically-informed integration of *cis*- and *trans*-regulatory signals.

Logistic Regression Analysis and Meta-analysis

For each site, we performed logistic regression to associate predicted gene expression with SCZ diagnosis. Covariates included sex and genomic eigenvariates (GEs) previously associated with diagnosis (Dx) (Trubetskoy et al., 2022). The regression models were defined as:

$$\mathrm{Dx}(Y) = \mathrm{Unimodal}(C|E|M|I) + \mathrm{sex} + \mathrm{GEs} \tag{4.1}$$

$$\mathrm{Dx}(Y) = \mathrm{Multimodal}(C + E + M + I) + \mathrm{sex} + \mathrm{GEs} \tag{4.2}$$

Because the 62 PGC3 cohorts differ in SNP coverage, not all genes could be imputed in every cohort. To ensure robust and comparable gene-level predictions, we retained only genes that were successfully imputed in at least 32 cohorts. This threshold balanced gene retention with cross-cohort consistency, avoiding biases introduced by missing coverage in a subset of cohorts. By requiring presence across multiple cohorts, we also ensured that predictive SNPs were sufficiently represented in the population, increasing the reliability of the retained set of genes. Meta-analysis of β coefficients across cohorts was performed using Stouffer's method, with sample size-based weighting (Zaykin, 2011), implemented in the `metap` R package.

Heterogeneity across sites was assessed using Cochran's Q test. Genes with heterogeneity p-values $> 1 \times 10^{-3}$ were retained. Multiple testing correction was applied across all 114,954 tissue-gene pairs using the Benjamini-Hochberg FDR procedure at $\alpha = 0.01$.

Assessment of Statistical Inflation

To evaluate the robustness of our coTWAS findings and assess the potential presence of false positives, we performed a systematic analysis of p-value distributions across all gene–region associations. Specifically, we calculated the genomic inflation factor (λ) as well as the scaled metric λ_{1000}, which normalizes λ to a fixed number of 1,000 tests, to quantify potential deviations from the expected null distribution.

We further compared the observed versus expected number of associations across pre-defined p-value bins, generated histograms of adjusted p-values, and constructed quantile–quantile (Q–Q) plots to visualize enrichment relative to the null. To assess regional variation, λ values were also estimated separately for each brain region. Finally, we applied formal goodness-of-fit tests, including the Kolmogorov–Smirnov and chi-square tests against the uniform distribution, to evaluate whether the empirical distribution of p-values significantly

deviated from expectation under the null hypothesis.

4.3.5 Cell-Type Specificity Analysis

To assess the cellular specificity of coTWAS-significant genes, we used specificity indices derived from single-nucleus RNA sequencing of human brain tissue (Habib et al., 2017), covering ten neuronal and glial cell types. We conducted enrichment testing using the Mean-rank Gene Set Test implemented in the `limma` R package (v3.46) (Ritchie et al., 2015). Multiple comparisons across genes and cell types were corrected using FDR ($\alpha = 0.05$).

4.4 Results

4.4.1 Functional Enrichment of SCZ Risk Variant-Associated Genes in Predictive Models

We analyzed genetic associations and functional implications of SNPs regulated by both *cis*- and *trans*-eQTL models in psychiatric disorders,focusing on the potential enrichment for SCZ GWAS-significant SNPs. We hypothesized that SCZ would show relatively high enrichment for *trans*-eQTLs, as evidence supports the role of co-expression networks in channeling genetic risk (Liu et al., 2019; Pergola et al., 2019a; Borcuk et al., 2024).

We evaluated the correlation between the log odds ratio (log(OR)) of SCZ-associated SNPs from GWAS ($p < 0.05$) (Trubetskoy et al., 2022) and SNP predictive weights from CIS, EpiXcan, and MODULE models. SNP weights were computed as the mean absolute weight across all genes each SNP regulates. Across brain regions, the CIS model included between 2,065 and 6,118 SNPs overlapping with PGC SNPs (≈6% of the model's SNPs), the EpiXcan model included 100,282 to 139,932 SNPs (≈13%), and the MODULE model included 47,262 to 86,200 PGC SNPs (≈11%) (Table ST4).

Correlation of Model Weights with SCZ Risk. In all brain regions investigated, we observed a significant positive Pearson correlation between model SNP weights and SCZ GWAS log(OR) values. Notably, the MODULE model consistently exhibited the strongest correlations (Figure 14A, top panel). After removing outliers (±3 SD from the mean of weights and logOR variable, within each model–pathology combination), CIS and EpiXcan models showed weaker but stable correlation values of ≈ 0.15–0.19 (all ($p < 0.01$), whereas MODULE correlations remained substantially higher, ranging from $r = 0.28$ to 0.42 ($p < 0.001$ across disorders; Table ST4). These results indicate a stronger association between SNPs used in MODULE predictions and SCZ genetic risk.

Since models have different numbers of SNPs, we employed Fisher's Z-transformation (Fisher, 1915) to test whether the observed differences in correlation coefficients were statistically significant. This analysis confirmed that the MODULE model significantly outperformed both CIS and EpiXcan across all brain regions (all $p < 2 \times 10^{-16}$; Figure 14B), suggesting a stronger link of *trans*- than *cis*-eQTLs with common variant pathogenicity. A similar trend was also found in other disorders, such as major depressive (MDD) and bipolar (BIP) disorders, with significant differences in correlation coefficients between predictive models (Figure 14A middle and bottom panels). The SCZ pattern thus extends to other psychiatric disorders showing some evidence of *trans*-heritability (Borcuk et al., 2024).

Network Connectivity with PGC3-Prioritized Genes. To further evaluate the functional relevance of *trans*-regulated genes, we prioritized MODULE-predicted genes based on a PGC-weight ratio. This ratio was defined as the average weight of SCZ-associated SNPs ($p < 0.05$) from PGC3 summary statistics divided by the average weight of all predictive SNPs regulating each gene. Genes with a ratio > 0.5 were selected under the hypothesis that prioritizing genes more heavily influenced by SCZ risk variants would improve detection of biologically relevant associations.

We evaluated the connectivity of these genes with the 120 PGC3-prioritized SCZ genes (Tru-

betskoy et al., 2022). This analysis was informed by a recent study from Borcuk et al. (2024), which demonstrated that genes more connected to network modules enriched for risk loci also accumulate genetic risk. The authors also identified genes most strongly connected to the set of PGC3 prioritized genes, to the extent that CRISPR activation of PGC3 prioritized genes correlated with their connectivity measure. Using the published gene-wise connectivity scores across five brain regions (amygdala, DLPFC, CN, HP, and sACC), we assessed the relationship between PGC-weight ratios and connectivity by binning genes into quintiles of their PGC-weight ratio and quantifying mean connectivity per quintile (Fig. S6) (see Methods 4.3.3 for details). Using classical linear regression and Spearman's correlation, we observed significant nominal trends in two model–region combinations: MODULE in sACC (slope $p = 0.0107$) and CIS in HP (slope $p = 0.0255$).

To rigorously evaluate these patterns, we performed a gene-level permutation test ($n = 1000$ permutations) (Methods 4.3.3). This revealed robust connectivity increases for MODULE models in four regions: DLPFC, CN, HP, and sACC (all permutation slope $p < 0.001$; in sACC also Spearman $p_{\text{perm}} = 0.018$). In addition, a weaker but nominally significant effect was observed for CIS in Amygdala (slope $p_{\text{perm}} = 0.037$). No significant monotonicity was detected for EpiXcan in any region (Fig. S7).

We note that in some cases permutation tests yielded significant results while the raw linear slope test did not. This discrepancy reflects differences in sensitivity: the raw linear test, based on only 5 quintile means, has low power and assumes a strictly linear relationship. In contrast, the permutation test directly evaluates whether the observed ordering of quintile means departs from chance, without assuming linearity. The fact that permutation tests identified robust signals in MODULE models suggests that the effect is real but not strictly linear, consistent with enrichment being concentrated in the upper quintiles.

Collectively, these findings refine our initial descriptive observation. The enrichment of connectivity with SCZ risk variants is not a global property of all models, but rather a robust feature of co-expression–based MODULE models across several regions, with only weak or

region-specific evidence for CIS and none for EpiXcan.

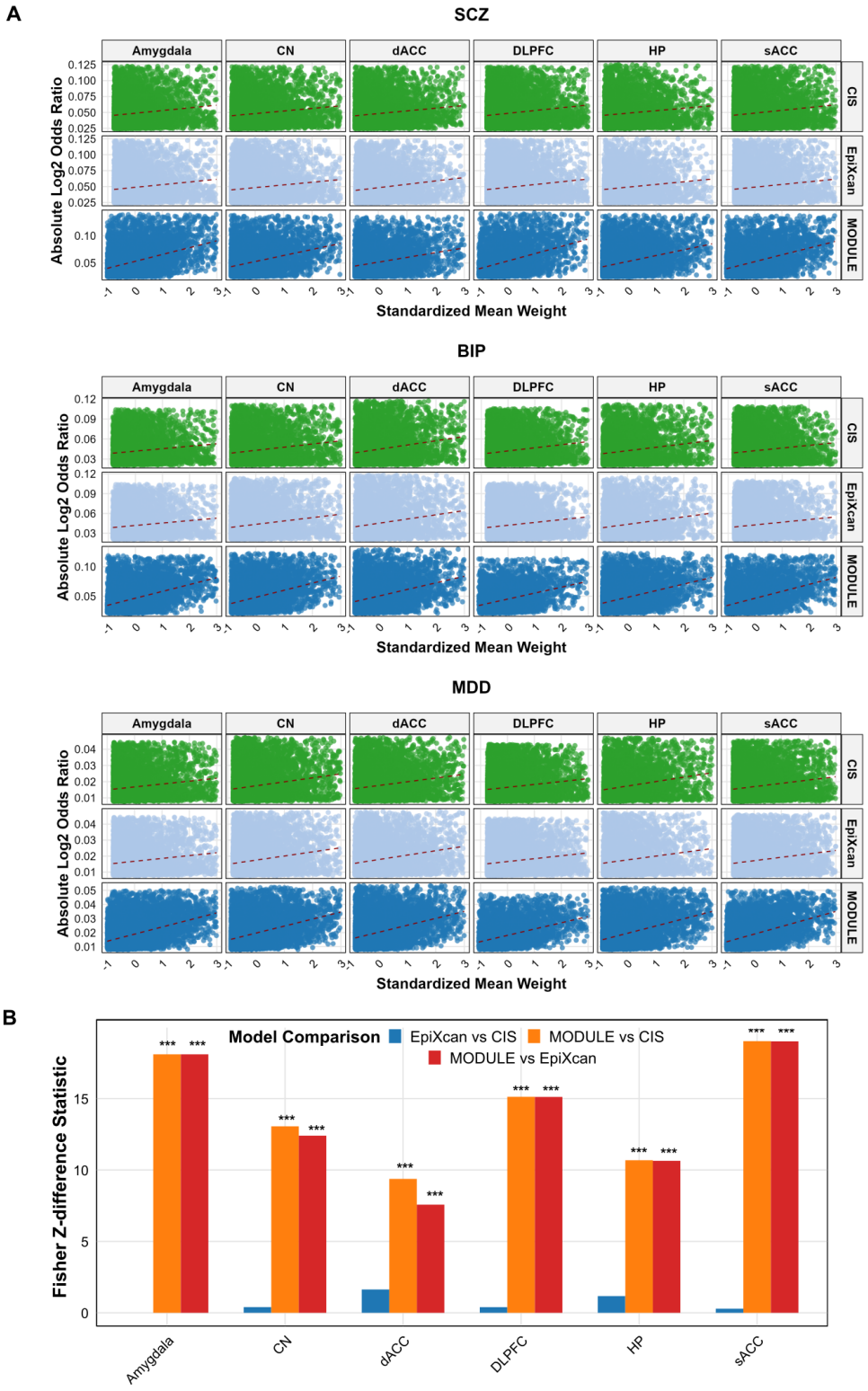

Figure 14: **Comparison of SNP Effect Sizes and Model Weights. A)** Scatterplots of absolute PGC3 log(OR) values vs. mean SNP weight Z-scores for CIS (green), EpiXcan (light blue), and MODULE (blue) across brain regions and disorders (SCZ, MDD, BIP). **B)** Barplots showing Fisher's Z tests comparing correlation strengths between models. Significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Summary. Taken together, these results underscore the biological relevance of interactions among genes predicted by *trans*-eQTLs and the functional role of the SNPs identified, potentially implicating pathways and mechanisms relevant to SCZ etiopathology that act via the mediation of gene co-expression.

4.4.2 Identification of SCZ-Associated Genes via coTWAS

To identify genetic associations with SCZ, we applied *cis*- and *trans*- models developed in Study 1 to impute gene expression and perform trait mapping using 62 independent genotype cohorts from the PGC3 (Trubetskoy et al., 2022), including 102,613 individuals (refer to ST5 for the distribution of PGC3 cohorts). Within each cohort and brain region, we selectively retained model-specific genes that met the postmortem selection criteria from Study 1 (see chapter 3 section 3.4.2 for gene selection details). Figure S8A reports the total number of predicted genes in each brain region when combining predictions across sites. We used logistic regression to associate predicted gene expression with SCZ diagnosis (coTWAS) in each PGC3 cohort separately. We corrected results for multiple comparisons using the Benjamini-Hochberg false discovery rate (FDR) with $\alpha = 0.01$, as subject overlap between brain regions violates Bonferroni's assumption of test independence. For details on the number of genes meeting various significance thresholds, see Figure S8B.

Transcriptome-wide identification of coTWAS hits. Out of 114,954 total tests performed across all brain regions, we identified 2,232 (2,030 non-MHC) significant associations, corresponding to 1,764 (1,683 non-MHC) unique genes (Table 4; Figure 15; Figure S8). The direction of gene effect size was derived from the β coefficient of the logistic regression; we found 1,042 genes genetically up-regulated and 1,190 down-regulated in SCZ patients compared to controls across regions (Figure 15). To quantify cross-region consistency in gene effects, we examined the genes found significant in $\geq$2 regions and computed pairwise correlations of TWAS β across brain regions. Effects were generally concordant (mean r

≈ 0.85; range $\approx$ 0.23-0.97), but several region pairs showed lower concordance, indicating region-specific directionality for a subset of genes (Supplementary Fig.S9). These findings suggest that while many risk-related signals are shared across regions, a measurable fraction is region-dependent.

Brain Region	**Total Tests**	**Significant Genes (% MHC)**	$\beta > 0$**;** $\beta < 0$
Amygdala	19,661	331 (10%)	185; 146
CN	19,517	341 (10%)	167; 174
dACC	18,576	401 (9.4%)	226; 175
DLPFC	19,352	327 (9.5%)	167; 160
HP	19,000	340 (8.2%)	151; 189
sACC	18,848	492 (7.7%)	146; 346
All Regions	114,954	1,764 (4.4%)	N/A

Table 4: **Summary of gene-level association testing with SCZ across brain regions (FDR $\alpha = 0.01$).** "Total Tests" refers to the number of gene-level tests conducted across regions. "Significant Genes" indicates the number of genes reaching FDR < 0.01, with the percentage in parentheses denoting the proportion of these that are located in the MHC region. The final column shows the number of significant genes with positive versus negative logistic regression coefficients ($\beta > 0$; $\beta < 0$). For the "All Regions" row, counts reflect the union across regions, and the percentage appears lower due to gene overlap. Beta direction is not provided here because directionality may differ across regions for the same gene.

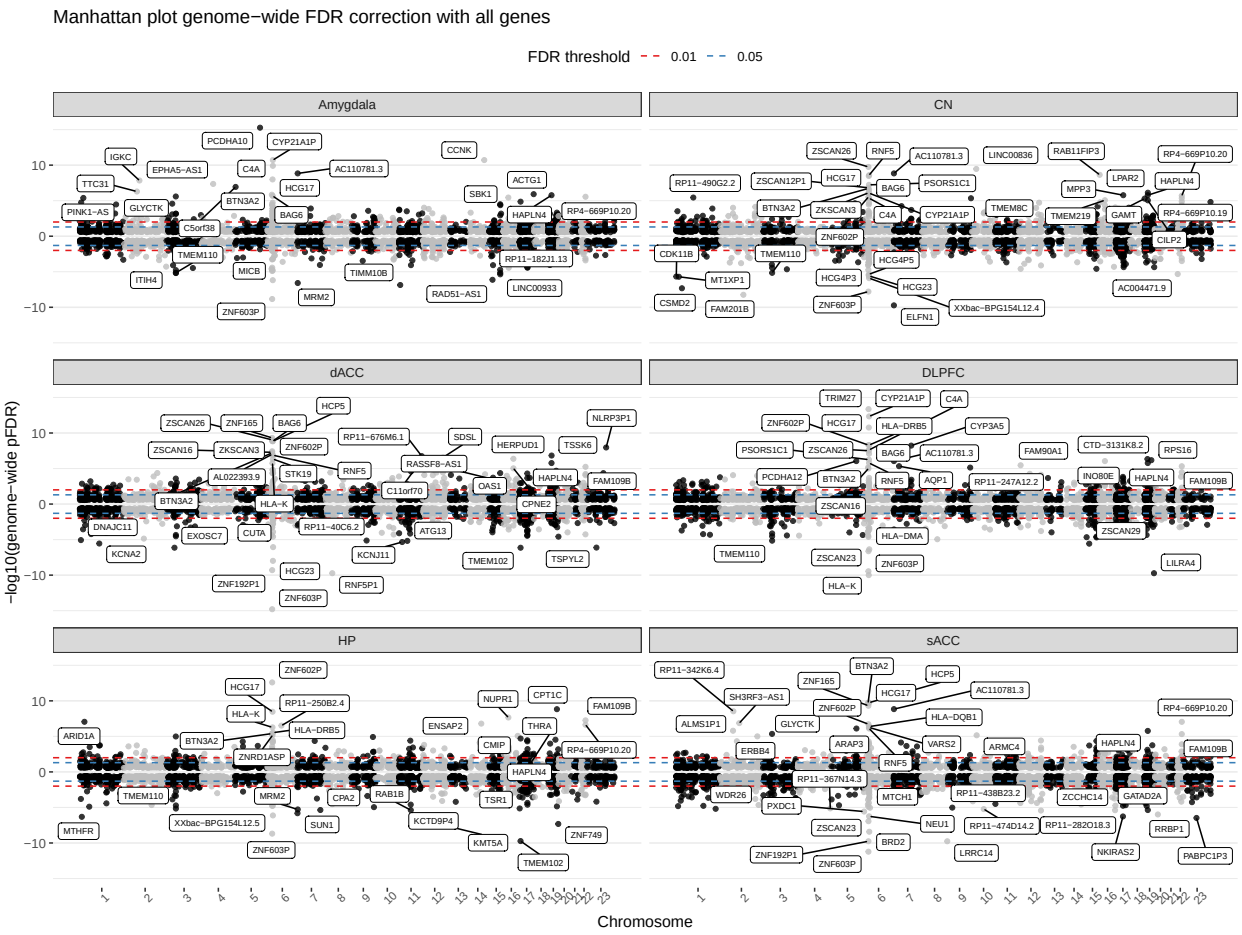

Figure 15: **coTWAS significant genes across brain tissues.**
Y-axis shows -$\log_{10}$(FDR-adjusted p-values), while the x-axis shows chromosomes. Red and blue lines mark FDR thresholds of 0.01 and 0.05, respectively. Positive direction (upper panel) represents up-regulated genes in SCZ; negative direction (lower panel) represents down-regulated genes. Due to clutter, a maximum of 50 gene labels are shown per region.

We performed Gene Ontology (GO) enrichment analyses to explore biological processes associated with coTWAS-significant genes. For downregulated genes ($\beta < 0$), we observed significant enrichment (qFDR < 0.05) in immune regulation, synaptic organization, cellular adhesion, and response to stimuli (Figure 16A). No significant GO terms were detected for upregulated genes.

Assessment of Statistical Inflation. To evaluate whether our coTWAS findings were affected by statistical inflation, we analyzed the distribution of p-values across 114,954

gene–region tests. The histogram of adjusted p-values (Fig. S10A) showed a clear enrichment of small values relative to the uniform null expectation. Quantile–quantile plots further illustrated deviation from the diagonal (Fig. S10B), while comparison of observed versus expected counts across p-value ranges confirmed an excess of significant associations (Fig. S10C).

We next quantified genomic inflation by computing the overall inflation factor ($\lambda = 0.391$, $\lambda_{1000} = 0.995$). Regional values ranged from 0.339 to 0.458 (Fig. S10D), indicating no evidence of systematic inflation. In fact, values below 1 suggest a slightly conservative test statistic distribution. Despite this, we observed enrichment of small p-values, with 2,232 associations (1.94%) reaching FDR < 0.01 and 5,302 (4.61%) reaching FDR < 0.05. Both Kolmogorov–Smirnov and chi-square goodness-of-fit tests ($p < 1 \times 10^{-16}$) confirmed deviation from the uniform null, consistent with the presence of true polygenic signal rather than random noise.

Together with the application of a conservative multiple testing threshold ($\alpha = 0.01$), these results suggest that our significant findings are unlikely to be driven by systematic false positives.

***cis* vs. *trans* model contributions to SCZ associations.** Of the coTWAS-significant genes, 430 were uniquely predicted by one model across regions: CIS (13.3%), EpiXcan (3.5%), INGENE (77%), and MODULE (6.2%). We focused on genes significant in at least two regions to highlight robustness. Of these, 35 (20.7%) were exclusively *cis*-predicted, 42 (24.9%) exclusively *trans*-predicted, and 92 (54.4%) included at least one *trans*-component in their prediction (Figure S11A). Across brain regions, *trans*-predictions accounted for the majority (49.9%–57.6%), followed by *cis-trans* (32.5%–40.7%), and *cis*-only predictions (Figure S11B).

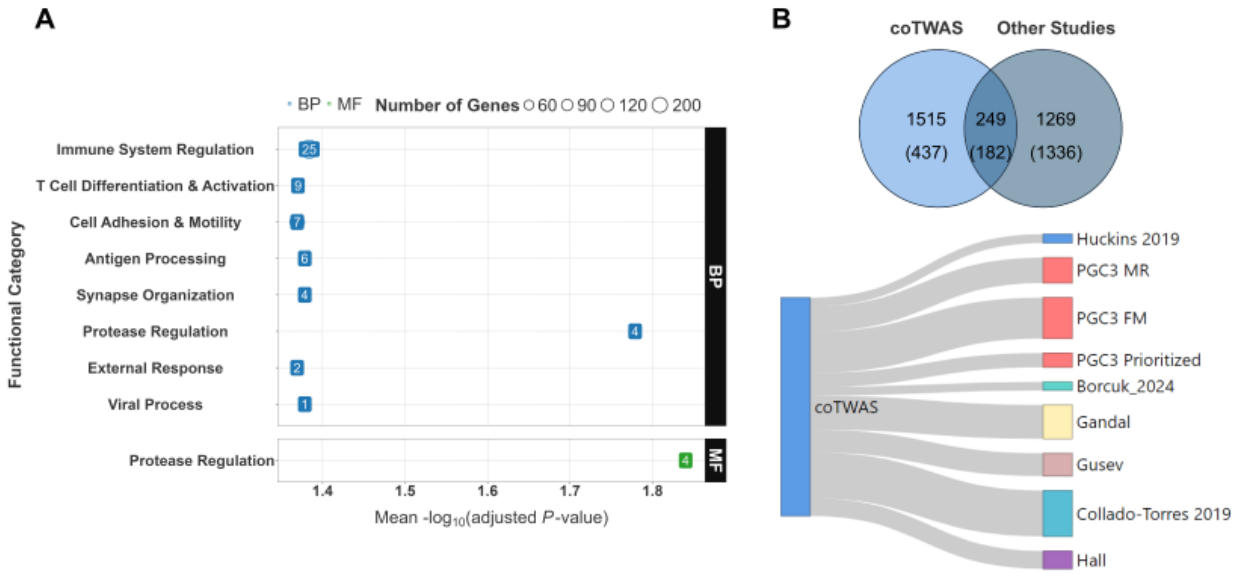

Figure 16: **Functional Enrichment and Cross-Study Overlap of coTWAS-Identified Genes Associated with SCZ. A)** GO enrichment of coTWAS-significant genes with $\beta < 0$. The x-axis indicates significance as -$\log_{10}$(FDR-adjusted p-value), while the y-axis lists functional categories grouped by GO domains. Point size corresponds to the number of genes per category; numbers in circles indicate the count of categories grouped by each functional domain. **B)** Sankey diagram (bottom) showing the overlap between coTWAS genes (blue bar, left) and SCZ-related gene sets from published studies. Gray streams represent the fraction of coTWAS hits present in each reference set. The Venn diagram (top) illustrates the total overlap between coTWAS genes (blue circle) and all SCZ gene sets (gray circle). Values in parentheses indicate the overlap obtained when restricting to genes predicted exclusively by *cis* models.

Functional Validation and Cellular Context of coTWAS Signals. To further elucidate the relationship between coTWAS signals and SCZ genetic risk, we examined the correspondence between coTWAS-derived predictions and proximity-based gene-level associations from MAGMA (de Leeuw et al., 2015). Specifically, we correlated MAGMA Z-scores with *cis*- and *trans*-based coTWAS predictions (Figure S12). Genes predicted via *cis*-models showed a statistically significant correlation (Pearson $r = 0.37$, $p < 2.2 \times 10^{-16}$), whereas predictions based exclusively on *trans*-regulatory information exhibited no significant association (Pearson $r = 0.03$, $p = 0.30$). These findings reinforce the notion that proximity-based approaches such as MAGMA are insufficient to capture the regulatory effects mediated by

trans-eQTLs. To highlight these novel gene associations with SCZ, Figure S13 highlights the top 50 genes with strong coTWAS support ($-\log_{10} p_{\text{adj}} > 5$) but weak MAGMA associations ($|Z| < 4$).

To explore the cellular context of these gene associations, we performed cell-type specificity enrichment using a human brain single-cell atlas (Habib et al., 2017) (Figure S14). In the amygdala, upregulated genes ($\beta > 0$) were enriched in excitatory DG-like neurons (exDG). In the dACC, two oligodendrocyte subtypes (ODC1 and ODC2) exhibited enrichment for upregulated genes, whereas exDG-like neurons showed downregulation ($\beta < 0$). In the HP, astrocyte subtypes (ASC1, ASC2) were enriched for upregulated genes, while exPFC neurons and microglia (MG) showed downregulation. Finally, in the sACC, both microglia and endothelial cells (END) displayed a pattern of downregulation. Together, these findings suggest that SCZ-associated transcriptional alterations exhibit both region- and cell-type specificity, highlighting dysfunctional neuron–glia interactions as a potential hallmark of disease pathophysiology.

Association strength between predicted gene expression and SCZ was evaluated by the meta-analytic p-values within brain regions (Figure 15). Among the consistently highly associated genes across regions *AC110781.3* (upregulated in Amygdala, CN, dACC, DLPFC, sACC) (Huckins et al., 2019; Hindley et al., 2022), *HAPLN4*, *FAM109B*, and *AS3MT* (Duarte et al., 2016; Li et al., 2016; Pardiñas et al., 2018; Roussos et al., 2014; Huo et al., 2019) have previously been linked to SCZ. Additionally, the coTWAS identified SCZ-related genes, including *SNX19* (Ma et al., 2020), *ZNF804A*, *GATAD2A*, *TCF4* (Teixeira et al., 2021), *CYP21A1P* (Cai et al., 2018), *C4A* (Sekar et al., 2016), and *CYP2D6* (Ma et al., 2021). While some of these genes were identified via *cis*-predictions, many—including *ZNF804A*, *AS3MT*, *FAM109B*, *C4A*, and *SNX19*—also exhibited a *trans*-component in their predictions.

Convergence with established SCZ gene sets. To further validate our findings, we cross-referenced our coTWAS significant genes with previously identified SCZ-associated genes. Specifically, we compared our results with SCHEMA (Singh et al., 2022), PGC3-prioritized genes (Trubetskoy et al., 2022), as well as those identified through Summary-based Mendelian Randomization (MR) and Fine-Mapping (FM) (Trubetskoy et al., 2022) (Figure 16B). Additionally, we included genes identified by Borcuk (Borcuk et al., 2024) as network neighbors of SCZ risk genes, and previous SCZ TWAS (Gusev et al., 2018; Gandal et al., 2018a; Collado-Torres et al., 2019; Huckins et al., 2019; Hall et al., 2020) (Table 8; Figure 16B).

Our analysis replicated many well-established SCZ associations, validating these predictions, while also identifying novel candidates (Figure 16B on the top). The coTWAS set included 249 genes previously implicated in other studies, alongside 1,515 uniquely identified genes, demonstrating the strength of *trans*-eQTL-driven modeling in capturing regulatory effects missed by traditional *cis*-only methods. To enable a fair comparison with prior TWAS analyses that primarily focused on *cis*-regulatory variation, we restricted the replication analysis to the 619 genes with a *cis*-predictor component across all brain regions. Within this subset, we replicated 182 previously reported SCZ-associated genes and identified 437 novel associations (Figure 16B, values shown in parentheses). Notably, the difference between the total and *cis*-restricted overlaps indicates that 67 previously implicated genes were captured exclusively through *trans*-based predictors in our framework. This observation underscores the ability of co-expression-informed *trans* models to recover established SCZ-associated genes that would have been missed by *cis*-based prediction alone, further demonstrating the complementary value of modeling distal regulatory architecture in complex traits like SCZ.

Among the 249 *cis* + *trans* overlapping genes, a subset was consistently detected in at least four independent studies (Figure 17A). Although this set of overlapping genes was not significantly enriched for any gene ontology, key risk loci emerged across independent studies, including *ZNF804A*, a well-replicated SCZ GWAS hit, as well as *TSNARE1* and *PCDHA*

family members (*PCDHA2*, *PCDHA7*, *PCDHA8*), implicated in synaptic connectivity and adhesion. Figure 17B highlights the robustness of coTWAS, as fold-enrichment analysis reveals a significant overrepresentation of recurrently identified genes across independent studies. Notably, coTWAS is the only method, apart from PGC approaches, to show a high overlap Szymkiewicz Simpson coefficient with MR genes (Trubetskoy et al., 2022) (Figure 17C), emphasizing its ability to bridge transcriptomic and genetic risk analyses.

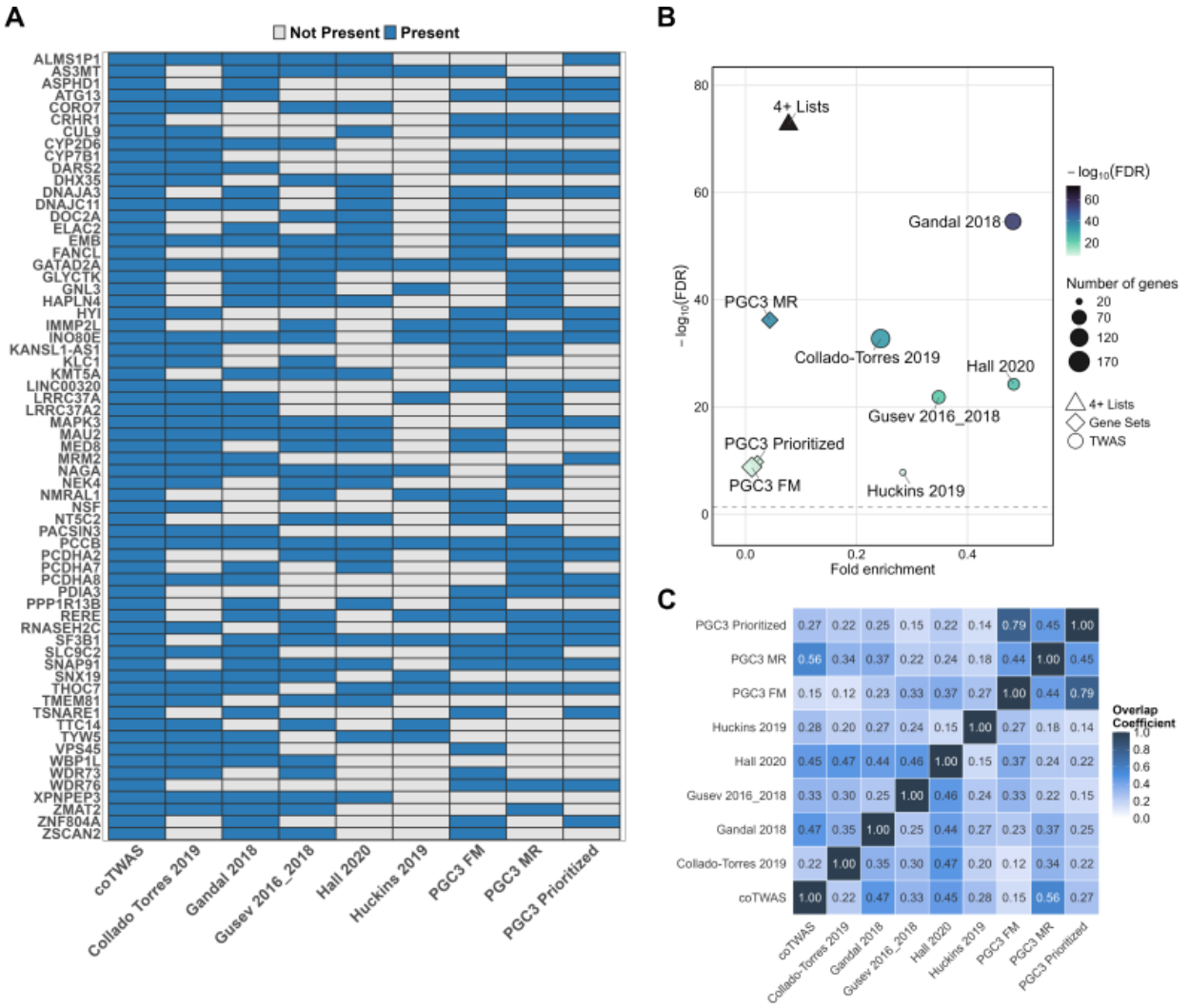

Figure 17: **Analysis of Gene Set Intersection and Enrichment Between coTWAS Hits and SCZ-Associated Gene Sets. A)** Presence/absence heatmap showing individual genes (y-axis) across studies (x-axis). **B)** Fold-enrichment (x-axis) versus $-\log_{10}(\text{FDR})$ (y-axis) for the overlap between coTWAS and each gene set. The dashed line marks an FDR significance threshold of 0.05. **C)** Matrix of Szymkiewicz–Simpson coefficients quantifying pairwise proportional overlap among SCZ-associated gene sets.

Summary. Our coTWAS analysis identified 1,764 significant gene associations with SCZ, including 1,515 novel TWAS hits. These results underscore the value of incorporating gene co-expression and *trans*-eQTL data into TWAS frameworks to uncover biologically meaningful mechanisms underlying psychiatric disorders.

4.5 Discussion

Beyond Proximity: A Network-Based Perspective on TWAS. Traditional TWAS approaches—whether at the gene (Gamazon et al., 2015; Zhang et al., 2019) or isoform level (Bhattacharya et al., 2023)—have enabled the prioritization of genes at GWAS loci that fall below genome-wide significance thresholds. These methods often implicate SNPs with borderline p-values ($5\times10^{-8} < p < 10^{-3}$), suggesting that enhanced power or complementary methods can recover functionally relevant associations. Our coTWAS framework extends this line of work by incorporating co-expression-informed models, developed in Study 1, that redefine variant-to-gene mapping based on statistical association—rather than physical proximity. This network-based strategy allows for the inclusion of both *cis*- and *trans*-eQTLs, capturing regulatory effects across spatially distributed yet functionally cohesive genes.

Large-Scale Application and Gene Discovery in SCZ. We applied *cis* and *trans* prediction models, developed in Study 1 (refer to Chapter 3 section 3.3.1) to 102,613 cases and controls from the PGC3 collection (Table ST5), constituting a large transcriptomic analysis of SCZ risk. Across six brain regions, we identified 1,764 significant associations (Table 4; Figure 15), with 1,515 of these representing novel associations with SCZ (Figure 17B) in respect to previous studies (Gusev et al., 2018; Gandal et al., 2018a; Pardiñas et al., 2018; Collado-Torres et al., 2019; Huckins et al., 2019; Hall et al., 2020; Singh et al., 2022; Trubetskoy et al., 2022; Borcuk et al., 2024). The significant overlap between genes identified through coTWAS analysis and Mendelian Randomization (Trubetskoy et al., 2022) (Figure 17B; Figure 17E) underscores the relevance of coTWAS signals to putative causal pathways

in SCZ. However, since both TWAS and MR share an instrumental variable framework and rely on genetic instruments to infer transcript-trait relationships Yuan et al. (2020), their overlap should not be viewed as an entirely independent validation. Rather, the convergence across these methods highlights consistent genetic evidence pointing toward the involvement of these genes in SCZ pathophysiology.

The Central Role of Trans-Regulation. A key aspect of our analysis involved the examination of genes predicted exclusively by *trans*-models. Notably, 49.9% to 57.6% of SCZ-associated genes were identified as exclusively *trans*-predicted across individual regions (Figure S11B), emphasizing the importance of *trans*-regulatory effects in SCZ. Additionally, 32.5% to 40.7% of SCZ-associated genes were predicted by both *cis* and *trans* models across regions (Figure S11B). The robust positive correlation between the coTWAS association strength and the MAGMA Z-score for *cis*-derived predictions indicates direct and localized control by *cis*-regulatory elements near the genes they regulate. Conversely, the lack of significant correlation for exclusive *trans*-predictions (Figure S12) highlights the additional insight available when using *trans*-eQTLs. In this scenario, integrating both *cis* and *trans* models provides a more comprehensive understanding of genetic influences on SCZ.

Functional Enrichment Highlights Immune and Synaptic Pathways. GO enrichment analysis on downregulated ($\beta < 0$) coTWAS significant genes (Figure 17A) highlights immune system regulation as the most significantly enriched functional category. The enrichment of T-helper cell differentiation, CD4-positive T cell activation, and antigen presentation via MHC class I and II suggests that adaptive immune processes are particularly susceptible to the molecular effects set in motion by genetic risk for SCZ. It is important to note, however, that this enrichment characteristic is based on relatively reduced expression of these immune components, suggesting that SCZ risk is not associated with immune activation. In this respect, GO terms related to synapse organization and postsynaptic assembly align with the crucial role of these processes in SCZ, this time not solely based on genetic proxim-

ity but also including *trans*-eQTL evidence. Additionally, enrichment in cell adhesion, cell motility, and junction assembly suggests disruptions in neuronal connectivity and immune cell trafficking, more in general implicating cell-to-cell communication in SCZ pathology.

Cell-Type and Region-Specific Signals. Single-cell enrichment analyses revealed distinct patterns across brain regions and cell types. Excitatory neurons were upregulated in the amygdala but downregulated in the dACC and hippocampus, consistent with altered cortical–limbic communication. Oligodendrocyte (ODC1, ODC2) and astrocyte (ASC1, ASC2) subtypes were consistently upregulated, suggesting myelination and glial support as potential contributors to SCZ pathology. Conversely, the downregulation of microglia (MG) in the hippocampus and sACC may indicate impaired neuroimmune surveillance or aberrant synaptic pruning functions.

These regional differences in cell-type signatures are further supported by the distribution of cross-region correlations of coTWAS effect directions (Fig. S9). While many risk-associated genes show consistent directionality across regions, a subset exhibit region-specific effects. This heterogeneity likely reflects differences in cell-type composition, local regulatory networks, and molecular pathways engaged within each brain area. Such region-dependent effects may help explain why psychiatric disorders disproportionately affect certain neural circuits, despite broad genetic overlap across conditions. They also suggest that genetic risk mechanisms may act through distinct cellular contexts, with some genes exerting pathogenic effects in one region while remaining neutral—or even compensatory—in another.

From a methodological perspective, these findings underscore the importance of interpreting TWAS results in their regional context, as pooled analyses may obscure biologically meaningful heterogeneity. Replication across multiple brain regions is therefore essential for assessing the generalizability of signals. Integrative approaches such as MODULE, which combine co-expression networks with region-specific information, provide a framework for distinguishing shared from region-limited genetic effects. This regional perspective is also

relevant for translational applications, where broadly shared pathways may support general therapeutic strategies, whereas region-specific alterations could inform more targeted interventions.

Alignment with Emerging Multi-Omic Findings. These functional enrichments are further confirmed by region-specific expression changes in neuronal and glial populations of significant coTWAS genes (Figure S14). Results implicate excitatory neuron dysregulation—particularly upregulation in the amygdala and downregulation in the dACC and HP—in SCZ risk, pointing to cortical-limbic integration. The consistent upregulation of oligodendrocyte and astrocyte subtypes (ODC1, ODC2, ASC1, ASC2) underscores a possible role for altered myelination and glial support mechanisms in SCZ, whereas downregulated microglia (MG) in the HP and sACC could signal diminished neuroimmune or synaptic-pruning functions.

Revaluating the Role of Microglia. Notably, recent studies support the notion that SCZ arises from both neuronal and non-neuronal perturbations across multiple brain regions. Our data again highlight that association varies by brain region, just as it has been shown to vary by cell type/state (Skene et al., 2018). Large-scale single-cell transcriptomic consortia have shown that excitatory neurons exhibit pronounced disease-associated expression changes (Gandal et al., 2018a), paralleling our results. Moreover, the upregulation of oligodendrocyte-related genes in SCZ revealed by various genomic and transcriptomic analyses (Tkachev et al., 2003; Ripke et al., 2014) emphasizes the importance of myelin integrity and white-matter function, mirroring our observation that oligodendrocyte subtypes are upregulated in the dACC. Similarly, the astrocyte dysregulation noted in our HP data aligns with evidence that compromised glial support for synapses and metabolic regulation contributes to SCZ risk (Bernstein et al., 2024).

Immune-focused genomic studies (Sekar et al., 2016) have long implicated microglia—central regulators of synaptic pruning—in SCZ etiopathology. In our findings, microglial downreg-

ulation in the HP and sACC does not entirely support the prevalent hypothesis of the role of activated microglia in SCZ, either in terms of immune activation or in microglia initiated synaptic pruning (Sekar et al., 2016; Notter and Meyer, 2017; Birnbaum and Weinberger, 2020). We observe both upregulated and downregulated immune cell types—consistent with prior studies (Birnbaum and Weinberger, 2020). A key limitation of our analysis is that we are investigating bulk tissue, which may obscure cell-type-specific expression changes. Although certain genes are preferentially expressed in certain cell types, the eQTL effects we measure may also depend on their expression in other cell types, hence entangling attempts to parse out cell specificity purely based on computational approaches.

Limitations

Despite the strengths of this study, several limitations must be acknowledged. First, the predictive models used for gene expression imputation were trained on a relatively small number of postmortem brain samples. Prior work has shown that larger training sets are crucial for improving model performance and stability in TWAS frameworks (Gamazon et al., 2015; Fryett et al., 2020). To mitigate this issue, we validated model generalizability using an independent testing framework described in Study 1.

Second, our analyses were based on bulk tissue transcriptomic data, which cannot resolve cell-type-specific regulatory effects. Although we performed cell-type enrichment using external single-nucleus datasets, future work incorporating larger-scale single-cell eQTL resources (Batiuk et al., 2020; Jaffe et al., 2020; Ruzicka et al., 2024) could significantly enhance the resolution of *trans*-regulatory mechanisms and improve interpretability.

Third, the study did not model sex-specific gene expression effects. Given known sex differences in gene regulation and disease vulnerability in SCZ (Brown, 2011; Hoffman et al., 2022), stratifying by sex could reveal important biological differences. However, doing so would have substantially reduced the sample size available for model training and analysis, thereby reducing statistical power.

Fourth, the genetic analyses were limited to individuals of European ancestry. This decision was driven by the ancestry composition of the PGC3 cohorts, which are predominantly European. While this ensures population homogeneity, it limits the generalizability of our findings across ancestries. Expanding this framework to diverse populations remains a critical future direction.

Finally, although we applied stringent imputation and quality control protocols, reliance on genotype imputation introduces uncertainty, particularly in regions of low coverage or rare variant density. This may impact both the detection of significant associations and downstream biological interpretation.

Future Perspectives: Reducing False Positives and Improving Replicability

An important next step will be to extend the coTWAS framework with additional procedures aimed at reducing false positives and improving the robustness of discoveries. One promising direction is the implementation of locus-based conditional analyses, such as those described by Huckins et al. (2019), in which the genome is partitioned into independent windows and associations within each locus are tested conditionally. This approach would help disentangle correlated signals, reduce redundancy, and yield a set of associations that more faithfully reflect independent biological effects.

In parallel, replication strategies across contributing cohorts can be developed to evaluate the stability of associations in the absence of equally powered external datasets. For example, leave-one-site-out analyses across the PGC3 cohorts could provide empirical measures of replicability by testing whether associations remain consistent in direction and effect size when subsets of cohorts are held out. While conservative, such strategies would allow us to distinguish highly stable signals from those more sensitive to sample composition.

Integrating these methodological extensions into the coTWAS pipeline will be an important direction for future research, ensuring that novel transcriptome-wide associations

represent not only statistically significant but also replicable and biologically meaningful signals.

Conclusion

Our results confirm that gene expression imputation based on biological network priors, rather than solely on local genotype, adds interpretive value to genetic association studies. This approach complements GWAS and PRS by prioritizing mechanistically informed, transcript-level associations. Moreover, our cell- and region-specific findings reinforce the idea that SCZ involves a broad interplay between synaptic signaling, glial support, immune modulation, and vascular integrity.

Taken together, this chapter advances the notion that co-expression structure and *trans*-heritability are key to interpreting the distributed genetic architecture of SCZ. These findings underscore the importance of network-informed models in complex disease genomics and lay the groundwork for their application in future multi-omic and precision psychiatry frameworks.

Chapter 5

Study 3: Prediction of Behavioral Traits from Genetically Regulated Brain Expression

5.1 Introduction

Antisocial behaviour refers to a broad range of actions that violate social norms and harm or infringe upon the rights of others (Tuvblad and Beaver, 2013). Antisocial behaviour represents a serious clinical and public health concern, given its links to criminality, psychiatric disorders, and significant societal costs (Romeo et al., 2006). Impulsivity, defined as the tendency to act on urges without forethought or consideration of consequences, is a central trait in antisocial behaviour Patton et al. (1995); Moeller et al. (2001); Bakhshani (2014) and will be a focal construct in this study. Individuals with antisocial tendencies often exhibit elevated impulsivity, which is thought to predispose them to aggressive or rule-breaking behaviours (Bezdjian et al., 2011). In fact, subtypes of antisocial behaviour that are more aggressive (and typically more impulsive) show especially high heritability (approximately 65% genetic influence) compared to less aggressive rule-breaking (approximately 48%) (Burt,

2009). Both genetic and environmental factors contribute to the risk for impulsivity. Twin and adoption studies indicate that roughly half of the variance in antisocial outcomes is attributable to genetic influences (Rhee and Waldman, 2002; Burt, 2009), and meta-analytic evidence confirms that impulsivity itself is moderately heritable, with approximately 45% of variation explained by genetics (Congdon and Canli, 2008). The remaining variation in impulsivity is largely due to non-shared environmental factors—experiences unique to the individual (Rhee and Waldman, 2002). Moreover, genes and the environment work in tandem: genetic predispositions can shape one's exposure to environmental risks (for instance, child genes can evoke parenting behaviours that influence the child's impulsivity and antisocial outcomes) (Kendler and Baker, 2007). This gene–environment interplay implies that the development of impulsive tendencies—and by extension antisocial behaviour—arises from a complex interaction between hereditary factors and life experiences (Moffitt, 2005; McAdams et al., 2013).

While both genetic and environmental factors shape impulsivity and antisocial behaviour, this Chapter focuses specifically on the genetic underpinnings of impulsivity. It examines whether genetically imputed brain expression profiles—developed using co-expression-informed modeling (Chapters 3)—can predict individual differences in impulsivity. Impulsivity is assessed using the Barratt Impulsiveness Scale (BIS-11) (Patton et al., 1995), which captures cognitive, motor, and non-planning dimensions, each showing modest but significant heritability (Congdon and Canli, 2008). These impulsivity traits have been consistently linked to antisocial behaviour, substance use, and a range of psychiatric disorders (Moeller et al., 2001), underscoring the relevance of investigating their genetic basis.

Study Overview and Rationale. Building on the coTWAS framework introduced in Chapter 4, this exploratory study applies machine learning (ML) pipelines to predict Barratt Impulsiveness Scale (BIS-11) scores in a high-risk forensic cohort of incarcerated adult males from correctional facilities in New Mexico (see the Data Overview section for cohort details).

The analysis leveraged postmortem brain-derived GReX models across six anatomically distinct regions developed in Chapter 3. Predictions were generated from genotype data using models that integrated both cis- and co-expression-informed trans-eQTL information.

The central aim was to test whether GReX—grounded in biologically meaningful, co-expression-informed predictors—could generalize to a highly complex, environmentally sensitive behavioral phenotype such as impulsivity.

Modeling Strategy and Analytical Framework. To assess predictive potential, the study employed a ML pipeline framed in a nested cross-validation strategy, incorporating:

- Multiple feature selection techniques,
- Hyperparameter optimization,
- Comparison of three distinct ML algorithms: random forest (RF), eXtreme Gradient Boosting (XGBoost), and support vector machines (SVM).

These models were selected to represent a spectrum of ML paradigms commonly applied to complex, high-dimensional datasets: RF as an ensemble of decision trees robust to overfitting and capable of modeling complex feature interactions; XGBoost as a highly efficient, regularized variant of gradient boosting known for capturing subtle, nonlinear patterns; and SVM as a margin-based algorithm well-suited for high-dimensional, small-sample problems. Together, these algorithms were chosen to capture both linear and nonlinear relationships between genetically imputed expression scores and impulsivity traits.

Preview of Results and Interpretation. This chapter examines the feasibility of predicting individual differences in impulsivity using GReX. The predictive performance of the models was limited, consistent with the modest heritability of impulsivity—approximately half that of more genetically influenced traits such as SCZ (see Chapter 2, Section 2.2.1)—and the behavioural complexity of the phenotype.

The findings of this study illustrate several well-established statistical and biological challenges, including the curse of dimensionality inherent to transcriptome-based prediction, limited statistical power due to the small sample size (approximately two orders of magnitude smaller than large-scale psychiatric GWAS), and the likely influence of unmeasured environmental and developmental factors. It remains unclear whether the limited signal reflects a fundamental constraint of GReX in modelling behaviourally complex traits, insufficient genetic variation in the present sample, or limitations in the operationalisation of impulsivity.

Rather than demonstrating successful prediction, this Chapter provides a boundary analysis: it outlines the current limitations of our GReX-derived scores, given the present data scale, phenotype definition, and modeling approach, in capturing behavioural phenotypes, and identifies key design considerations for future research. By systematically mapping where and why predictive failure occurs, this study offers insights for the design of future efforts seeking to extend genomically anchored prediction frameworks into the realm of behavioural and psychiatric phenotypes.

5.2 Data Overview

This study uses data obtained through an institutional research collaboration between the University of Bari (UNIBA), the University of Pisa (UNIPI), and the Mind Research Network (MRN), a U.S.-based neuroscience consortium affiliated with the University of New Mexico (https://www.mrn.org/). The partnership enabled access to a large-scale forensic dataset comprising behavioral and genomic data from ≈985 white adult male inmates (age: 35 ± 10 years), along with those of 168 white adolescent males in reeducation facilities (age: 17.0 ± 1.1 years). Each participant contributed a DNA sample and completed standardized psychometric assessments.

5.2.1 Cohort Construction

Initial Dataset and Study Population. Genome-wide genotyping and initial quality control of DNA samples were conducted by UNIPI (see Appendix D.1 for details).

Subsequent genotype processing and quality control were performed at UNIBA following the same protocol used in Study 1 (Chapter 3) and Study 2 (Chapter 4), ensuring methodological consistency across studies. Additional details are provided in Appendix D.1.

The final quality-controlled cohorts consisted of:

- 605 adult males
- 98 adolescent males

After all filtering steps, the resulting dataset comprised 4,894,072 high-quality genotyped variants.

Inclusion and Exclusion Criteria. To ensure data integrity, biological plausibility, and adequate statistical power, we further applied a series of filters based on demographic and sample characteristics—including age, sex, genetic ancestry, and phenotypic completeness—to construct the final analytic sample. These criteria were selected based on both theoretical relevance and empirical considerations:

- **Adults**: The study aimed to investigate impulsivity and antisocial traits in an adult population, where such behaviors are more stable and diagnostically relevant. Excluding individuals under 18 reduced developmental heterogeneity, as impulsivity and personality traits are subject to maturational changes during adolescence.
- **Biologically male**: All analyses were restricted to participants assigned male at birth to reduce sex-related variability in gene expression regulation and behavioural phenotype expression. Prior research has documented significant sex differences in impulsivity, neurodevelopment, and antisocial behaviour (Weinstein and Dannon, 2015), which could introduce confounding effects if not accounted for.

- **Availability of complete BIS-11 total scores**: The BIS-11 scale served as the primary outcome variable for the study. Individuals with missing total scores were excluded to prevent bias in model training and evaluation, and to ensure consistency in phenotype measurement across the sample.

After applying these criteria, the final analysis sample comprised $N = 468$ adult male individuals.

Variable	**Mean (SD)**	**Range**
Age (years)	35.1 (9.7)	19–65
IQ	97.4 (13.8)	66–131
BIS-11 Total Score	70.0 (11.1)	44–102
Attentional	17.7 (4.0)	8–30
Motor	25.4 (4.6)	14–40
Non-planning	28.7 (5.0)	12–43
PCL-R Total Score	20.2 (6.6)	3.2–37.9
Factor 1 (Affective)	5.6 (3.3)	0–15
Factor 2 (Deviance)	12.4 (3.8)	0–20
BSMSS Education Score	11.6 (2.7)	5.0–18.5
BSMSS Occupation Score	20.8 (6.8)	5.0–40.0
Ethnicity	Hispanic (non-Hispanic)	52% (48%)
Sex at Birth	Male	100%

Table 5: **Demographic and Psychometric Characteristics of Final Analysis Sample (N = 468)**.

5.2.2 Phenotypic Measures

Primary Behavioral Phenotype Impulsivity was assessed using the BIS-11, one of the most widely used and validated instruments for measuring trait impulsivity in both clinical

and forensic populations. The BIS-11 is a 30-item self-report questionnaire that yields a total score and three subscale scores reflecting distinct facets of impulsivity (Figure 18 shows scores distribution):

- **Attentional impulsivity - BIS factor 1** – Difficulty in focusing and cognitive instability (e.g., "*I get easily bored when solving thought problems*").
- **Motor impulsivity - BIS factor 2** – Acting without thinking or the inability to inhibit behavioral responses (e.g., "*I act on the spur of the moment*").
- **Non-planning impulsivity - BIS factor 3** – A lack of future orientation or forethought (e.g., "*I am more interested in the present than the future*").

Each item on the BIS-11 is rated on a 4-point Likert scale, which captures the respondent's level of agreement with impulsivity-related statements. The scale ranges from "Rarely/Never" to "Almost Always/Always," providing an ordinal measure of trait impulsivity without offering a neutral midpoint. Scores range from 30-120 with higher scores reflecting greater impulsivity. In the present study, the **BIS-11 total score**-the sum of each subscale-was used as the primary continuous outcome variable in all predictive models due to its reliability, psychometric validity, and direct relevance to antisocial behaviour.

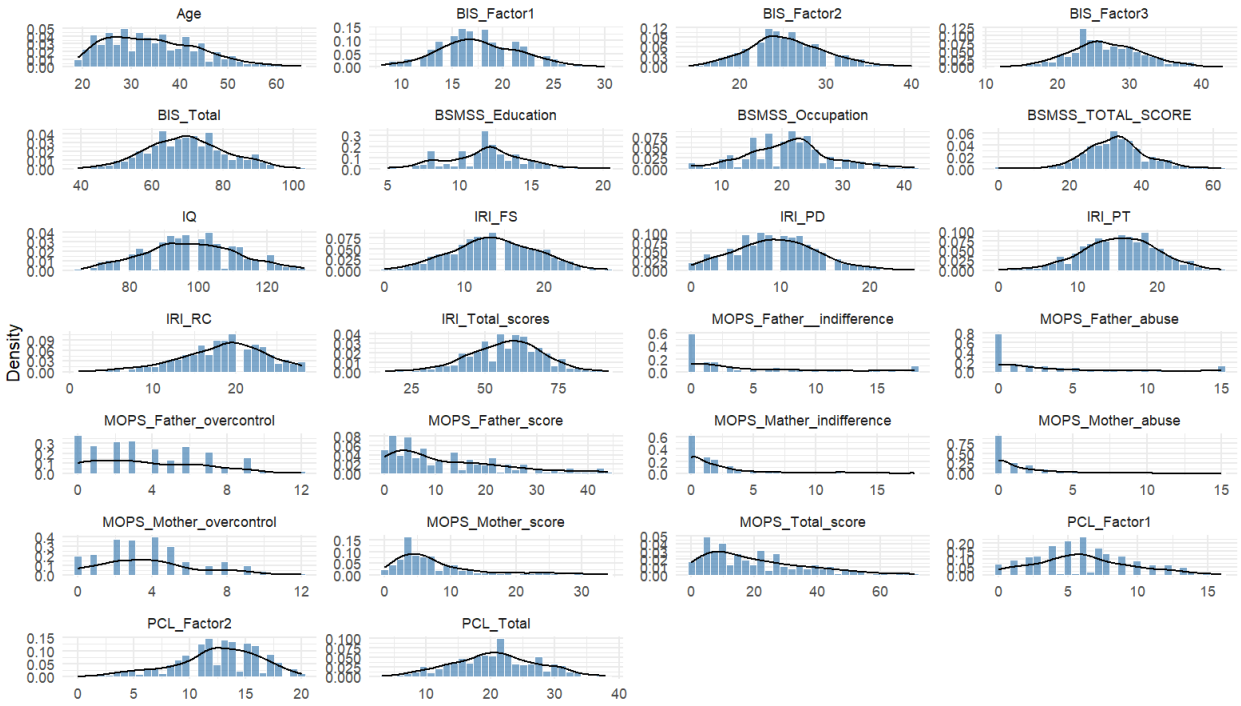

Figure 18: **Density plots of key psychometric and demographic variables in the adult inmate sample (N = 468).** Variables include impulsivity scores (BIS-11 total and subscales), socioeconomic status (BSMSS), cognitive ability (IQ), empathy traits (IRI), retrospective parenting measures (MOPS), and psychopathy dimensions (PCL-R). All scores were assessed using validated instruments. Distributions highlight individual differences in behavioral phenotypes and their potential variance structure for modeling.

Additional Psychometric Instruments Several additional behavioral and cognitive instruments were available. Although not directly used in the modeling pipeline due to missing data (Table 6), these measures provide important contextual information about participants' developmental and psychological profiles:

- **Psychopathy Checklist–Revised (PCL-R)** – A clinician-administered 20-item rating scale assessing core psychopathic traits across two factors: interpersonal/affective features (Factor 1) and antisocial lifestyle/deviance (Factor 2). The PCL-R is a gold-standard measure of psychopathy, frequently used in forensic risk assessments.

- **Interpersonal Reactivity Index (IRI)** – A 28-item self-report scale designed to assess components of empathy, including both cognitive and affective dimensions. The

IRI includes subscales such as Perspective Taking, Empathic Concern, Personal Distress, and Fantasy.

- **Measure of Parental Style (MOPS)** – A retrospective self-report measure that captures recalled parenting behaviour and childhood emotional experiences. It assesses perceptions of parental neglect, overprotection, and abuse, offering insight into early environmental adversity.

- **Barratt Simplified Measure of Social Status (BSMSS)** – Provides indices of socioeconomic status based on self-reported education and occupation. It includes separate subscales for the participant's education level and that of their parents, offering a broad index of developmental social background.

- **Estimated IQ** – Intelligence was assessed using an abbreviated version of the Wechsler scales or a comparable validated screener, providing a general estimate of cognitive functioning for descriptive characterization.

Table 6: **Missingness Summary for All Variables**

Variable	**Missing Count**	**Missing (%)**
BSMSS Education Score	238	51.0
BSMSS Occupation Score	238	51.0
PCL-R Total Score	43	9.1
IRI Total Score	82	0.18
MOPS Total Score	295	63.0
IQ	3	0.6
Age, Ethnicity, Sex	0	0.0

To preserve sample size, only **age** and **ethnicity** were retained as covariates. Ethnicity was encoded as a binary variable (Hispanic vs. non-Hispanic).

5.2.3 Predictors: Genetically Regulated Expression

Gene-level predictors for this study were derived from GReX, imputed directly from individuals' genotype data using predictive models developed and benchmarked in Chapter 3. These models estimate individual-level expression profiles by leveraging statistical associations between genetic variation and transcriptomic activity observed in postmortem brain tissues.

Specifically, we applied four complementary gene expression prediction frameworks to the high-quality genotype data from the adult inmate sample: CIS, EpiXcan, INGENE, and MODULE (see Chapter 3, Section 3.3.1 for methodological details, and Appendix D.2 for gene-expression imputation protocol). Briefly, CIS and EpiXcan models capture local (*cis*) regulatory effects, whereas INGENE and MODULE incorporate network-based approaches to model distal (*trans*) regulation through co-expression-informed priors. All models were originally trained on transcriptomic data from six neuroanatomically distinct brain regions, enabling both region-specific and integrative prediction strategies.

5.3 Methods

5.3.1 Regional and Combined Modeling Strategies

Two complementary modeling configurations were implemented:

- **Region-specific models**: Used GReX features from a single brain region to predict BIS-11 total scores, allowing localized evaluation of transcriptomic signals.
- **Combined-region models**: Concatenated residualized features from all six brain regions to create a high-dimensional integrative predictor matrix.

This dual approach enabled the assessment of whether integrating transcriptomic information across regions enhances predictive accuracy via potential synergistic effects. The

number of predictors for each configuration is reported in Table 7.

Table 7: **Number of Genes Imputed per Brain Region**

Brain Region	**Number of Genes**
All Regions	113,655
Amygdala	19,123
CN	19,440
dACC	17,997
DLPFC	19,847
HP	18,897
sACC	18,363

5.3.2 Machine Learning Pipeline

To model inter-individual variation in impulsivity, we implemented a robust and interpretable ML pipeline. The framework incorporated nested cross-validation, stratified data partitioning of the outcome variable, multi-step feature selection, and algorithm-specific hyperparameter optimization across three regressors: RF, XGBoost, and SVM. Modeling was conducted independently for each brain region and for the combined multi-region feature set.

Nested Cross-Validation Strategy

A 4-fold **nested cross-validation** scheme was employed. In each outer fold, 80% of the data were used for model training and 20% were held out for evaluation. Internal folds were used for feature selection and hyperparameter tuning. Stratification based on BIS-11 score quantiles ensured consistent outcome distributions across folds.

Feature preprocessing involved first removing genes in the bottom 5% of variance across individuals to eliminate low-information features. The remaining gene expression values were then centered and scaled to have zero mean and unit variance. The outcome variable was

similarly z-scored within each training fold to ensure that both predictors and outcome were standardized to a comparable scale.

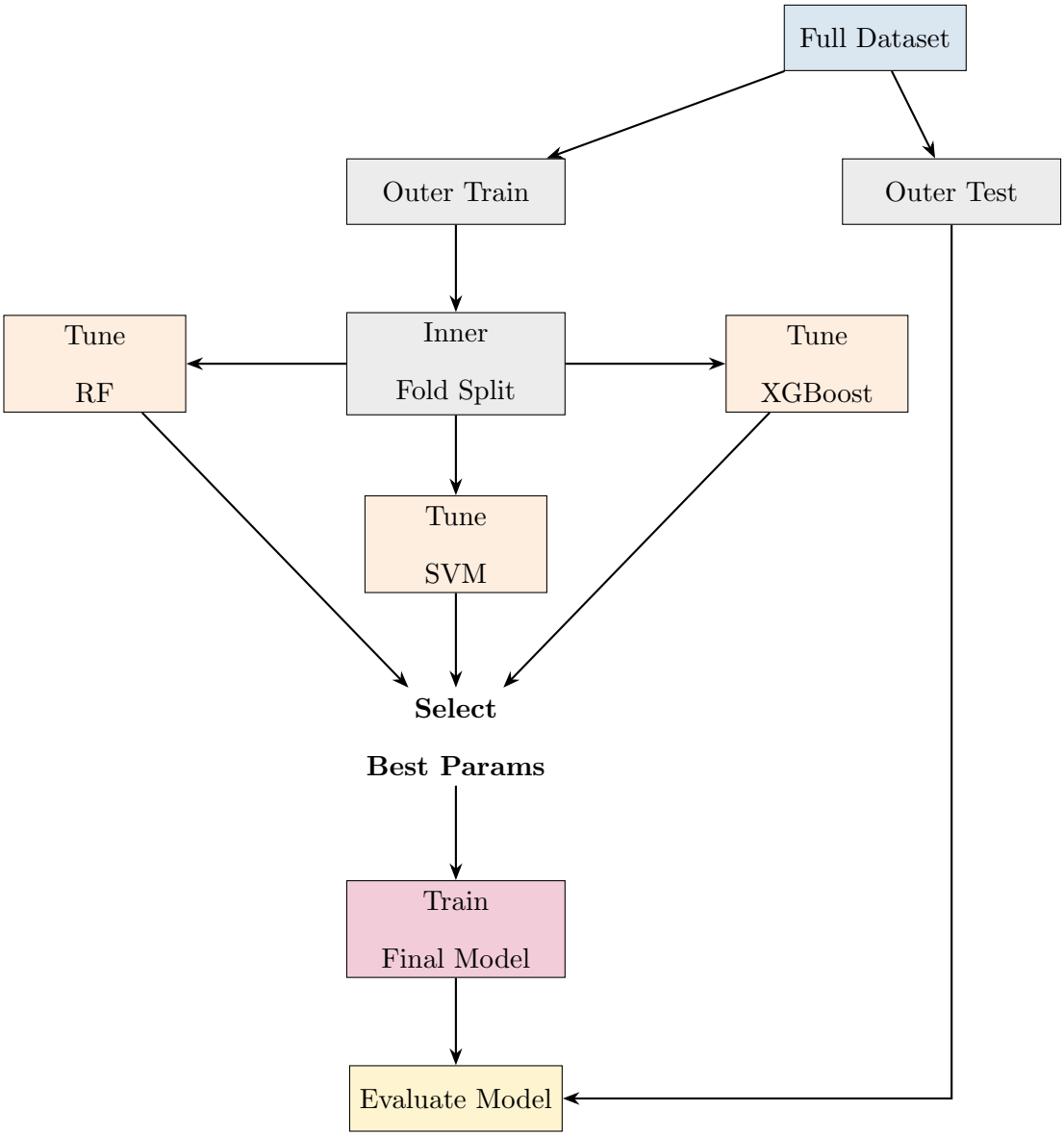

Figure 19: **Simplified ML pipeline overview**. Inner folds are used for algorithm tuning (RF, XGBoost, and SVM), followed by final model training on the outer training set and evaluation on the held-out outer test set.

Feature Selection Benchmarking

Feature selection was embedded within the training phase of each outer fold using a dedicated benchmarking routine. Four selection strategies were tested:

- **Correlation Filtering:** Features were selected based on statistically significant correlations ($p < 0.05$) with the outcome.
- **Boruta:** A wrapper-based method relying on random forest importance to iteratively filter features (Kursa and Rudnicki, 2010).
- **Combined Pipeline:** Features were first filtered by correlation, then further refined using the Boruta algorithm, followed by redundancy pruning (removing highly correlated pairs, $r > 0.8$); within each correlated pair, the feature with lower variance across individuals was discarded.

Each method was evaluated across three feature subset sizes (1%, 2%, and 5% of all genes) to assess whether progressive integration of predictors improved model accuracy. Performance at each subset size was assessed using an internal 80/20 train/test split within the training fold. For each configuration, we trained a RF model on the selected features and calculated the coefficient of determination R^2 on the 20% split. The features selected by the method and percentage with the highest predictive performance were retained and used in the main modeling phase.

This systematic benchmarking allowed us to identify the most informative and least redundant subset of features *per fold*, enabling fold-specific adaptation to the complexity and structure of the data.

Algorithm Tuning

Hyperparameter grids were selected to balance model complexity, generalizability, and computational feasibility given the high-dimensional, low-sample-size structure ($p \gg n$) of the predictive data. Because the number of gene expression predictors far exceeded the number of observations, tuning parameters were designed to promote regularization, prevent overfitting, and encourage robust learning:

- For RF, smaller `mtry` values (e.g., $\sqrt{p}$, $p/3$) and constrained `min.node.size` settings (1, 5, 10) were used to limit model complexity and reduce variance.
- For XGBoost, low learning rates (`eta` $\in \{0.01, 0.05, 0.1\}$) and shallow trees (`max.depth` $\in \{2, 3\}$) encouraged conservative, incremental learning suited to high-dimensional spaces.
- For SVM, tuning the `C` and `epsilon` parameters controlled the trade-off between margin maximization and model flexibility, promoting generalization in a sparse feature space.

Overall, the hyperparameter choices reflected the need to manage model capacity and avoid overfitting in a predictive setting dominated by far more features than available samples (see Appendix D.2 for a detailed description of tuning parameters and their rationale).

5.3.3 Performance Evaluation

Model predictions were generated for held-out outer folds and aggregated to evaluate overall performance. For each outer fold, the following metrics were computed separately on the training and test sets:

- **Root Mean Squared Error (RMSE)**: quantifies the average magnitude of prediction errors, giving greater weight to large deviations.
- **Mean Absolute Error (MAE)**: measures the average absolute difference between predicted and observed scores, offering a robust alternative to RMSE.
- **Coefficient of Determination (R^2)**: indicates the proportion of variance in the outcome explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- **Adjusted** R^2: provides a bias-corrected measure of model fit by accounting for the number of predictors. In this study, adjusted R^2 was extracted directly from the output of a linear model (`lm()` function in R) regressing observed scores on predicted scores within each outer test fold. This approach ensures that the penalty for model complexity is based on the effective degrees of freedom in the held-out data.

All metrics were averaged across the outer folds to obtain global performance estimates.

5.4 Results

5.4.1 Feature Selection Benchmarking: Maximizing Signal in The Presence of Uncertainty

Given the high dimensionality of the predictor space and the expected weak associations with the behavioral outcome, the first analysis step benchmarked different feature selection strategies. Pseudo-validation performance was estimated for each feature selection method using a RF trained on an internal 80/20 split, and the best-performing subset was retained.

Although this internal split introduces optimistic bias—since feature selection was performed on the full fold prior to splitting—it provided a controlled framework for comparing feature selection strategies. Importantly, these results informed method selection only; final model performance was evaluated separately on the outer test fold, which remained fully independent of feature selection and tuning.

To further illustrate how each method filtered the data, Figure 20 shows the reduction in feature count across selection steps for one representative cross-validation fold (Fold 1), using features from all six brain regions combined. This setup tested the maximum dimensionality context in our framework, providing a stringent benchmark for feature selection strategies. Starting with 1,425 features from correlation filtering, Boruta independently selected 42 features, while the combined correlation-Boruta approach retained 64 predictors.

Overall, the figure highlights the challenge of stable feature selection in high-dimensional gene expression data and the benefits of layered filtering strategies for managing noise and sparsity.

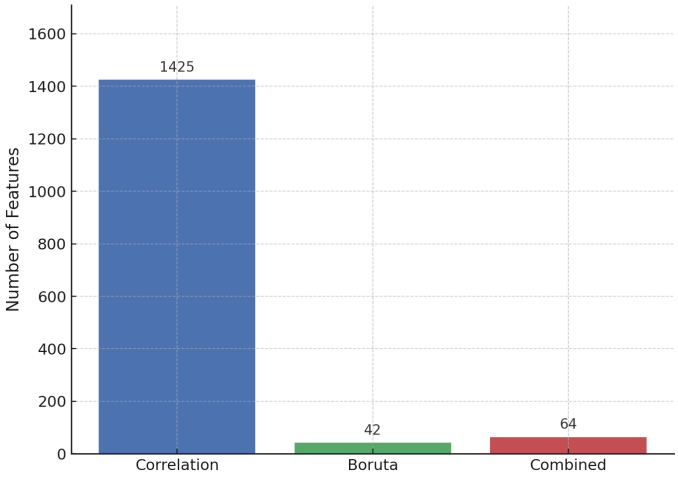

Figure 20: **Feature Count Across Selection Steps (Fold 1).** Feature selection was performed on the combined predictor space from all six brain regions. Correlation filtering (blue bar) initially retained 1,425 features, Boruta alone (green bar) selected 42 features, and the combined correlation-Boruta approach (red bar) retained 64 features. This figure, based on Fold 1, is representative of the general patterns observed across cross-validation folds. It illustrates the divergence between feature selection methods in high-dimensional transcriptomic data and the stringency achieved through multi-step filtering strategies.

Following feature selection, model performance was evaluated in two stages to assess predictive utility and generalization. Figure 21A shows the pseudo-validation R^2 obtained from the inner 80/20 split within each outer training fold, used to benchmark different feature selection strategies during internal model tuning. After identifying the best-performing feature subset, models were retrained on the full outer training fold and evaluated on the corresponding held-out test set, as shown in Figure 21B. For completeness, performance on the full outer training fold is also reported, allowing direct comparison between training and testing behaviour and assessment of potential overfitting.

The results, summarized in the diagnostic plots for each fold (Figure 21), reveal several critical patterns:

- **Limited predictive power even within training folds.** The variance explained by

the selected features—even when evaluated within the training fold—was generally low. Pseudo-validation R^2 values rarely exceeded 0.25 (Figure 21A), indicating that, even under favorable conditions, retained features offered limited predictive utility. This highlights the inherent difficulty of modeling impulsivity from transcriptomic data, where true signal is likely weak, sparsely distributed across genes, and potentially sensitive to sample characteristics such as environment, developmental stage, or tissue specificity.

- **Poor generalization to unseen data.** Final performance on the outer test folds—the truly held-out data—was uniformly poor across all folds and models. As shown in the predicted-vs-actual plots (Figure 21B), test set predictions (yellow lines) were nearly flat, indicating a failure to capture individual-level variation and a tendency to predict values close to the sample mean. This pattern indicates that models, despite fitting the training data to varying degrees, failed to generalize and produced non-informative predictions on unseen data—a sign of overfitting.

- **Algorithm- and feature selection–specific behaviors.** Distinct patterns emerged depending on the feature selection strategy. In folds where correlation filtering was selected (Figure 21A), training R^2 values often approached 1 (Figure 21B, blue lines), suggesting near-perfect in-sample fit driven by spurious noise structures. Correlation-based filtering, while efficient, likely captured many spurious linear associations, inflating in-sample performance. In contrast, when more stringent methods such as Boruta or the combined correlation-Boruta approach were used, training performance dropped considerably. Under these conditions, algorithm-specific variability became more evident: particularly, SVM and XGBoost showed greater fluctuations in training performance, likely due to their higher sensitivity to sparse feature spaces. While these algorithms are capable of modeling complex nonlinear relationships, this flexibility can lead to overfitting and instability when the underlying signal is weak or poorly defined.

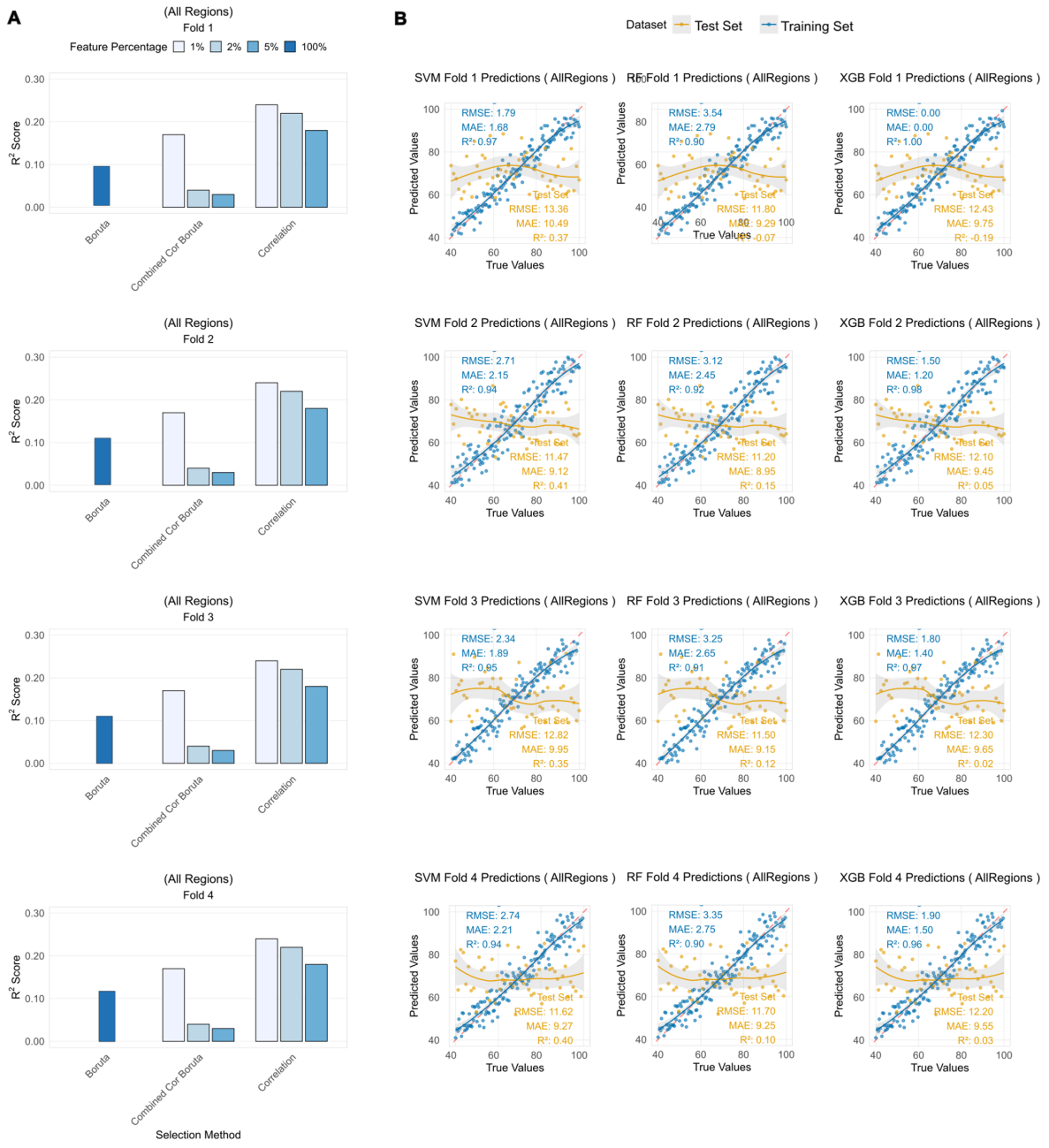

Figure 21: **Feature Selection and Model Performance Across Cross-Validation Folds.** Each row represents one outer CV fold (1–4), with plots showing (left to right): (i) R^2 scores on the pseudo-validation (inner 80/20 split) for each feature selection method, (ii) SVM performance on training and outer test folds, (iii) Random Forest (RF) predictions, and (iv) XGBoost predictions. Bars in the selection plots reflect the highest R^2 achieved for each method and feature subset size. The predicted vs. true scatter plots show model performance on both training and test sets, with blue lines for training and orange for test. Note the near-horizontal alignment of test predictions in many cases, indicating the model's tendency to regress to the mean—a hallmark of overfitting. SVM and XGBoost exhibit more fluctuation in training fit compared to RF, especially when conservative feature selection methods are used.

Taken together, these findings illustrate the core challenge of this modeling task: striking a balance between selecting informative features and preventing overfitting in a regime dominated by noise and weak signal. Correlation-based filtering frequently resulted in high in-sample performance, but this came at the cost of poor generalizability, indicating clear overfitting. More conservative approaches reduced overfitting to some extent but also led to substantial drops in apparent accuracy, highlighting the difficulty of identifying predictive signal under these conditions. Overall, none of the strategies fully overcame the trade-off between model fit and generalization.

Given the challenges outlined above, modest predictive performance across folds was anticipated; nevertheless, we systematically report model outcomes to capture the structure of model behaviour and failure modes.

5.4.2 Model Performance Across Brain Regions and Algorithms

All-Region Model Performance As shown in Figure 22, models trained on features concatenated across all brain regions consistently achieved near-perfect performance on the training folds, particularly for XGBoost ($R^2 \approx 1$) and SVM ($R^2 > 0.9$), albeit with high MAE and RMSE values. This highlights a clear overfitting issue, where models captured structure in the training data but failed to generalize.

Figure 22: **Model Performance Using All Brain Regions.** Barplots show the training (left) and testing (right) performance metrics across: RF, SVM, and XGB. Metrics include adjusted R^2, r^2, R^2, MAE, and RMSE. Training scores for SVM and XGBoost reach near-perfect levels, indicative of overfitting, while testing performance is uniformly poor across all metrics and models.

On the outer test folds—the true measure of generalizability—all three models displayed extremely poor predictive performance. Adjusted R^2 values hovered near zero or dipped into the negative range for all algorithms, with the strongest performer (SVM) achieving only a marginal adjusted $R^2 \approx 0.005$. Similarly, the R^2 values for the test set were negative across all algorithms, indicating that predictions were worse than simply using the mean outcome as a baseline. While SVM showed a slightly stronger correlation structure ($r^2 > 0.01$), the practical effect was negligible, and performance remained insufficient for any substantive behavioral prediction.

We further explored whether restricting predictors to individual brain regions might enhance performance; however, no substantial gains were observed (see Appendix, Figure S15).

Permutation Testing. To assess whether the modestly positive (though near-zero) test set metrics could reflect any real predictive signal—or were simply artifacts of model complexity and overfitting—we conducted permutation testing. For each algorithm, the target outcome was randomly shuffled 100 times, and the models were retrained and re-evaluated using the same pipeline. The distribution of test set performance under these null conditions was then compared to the observed results.

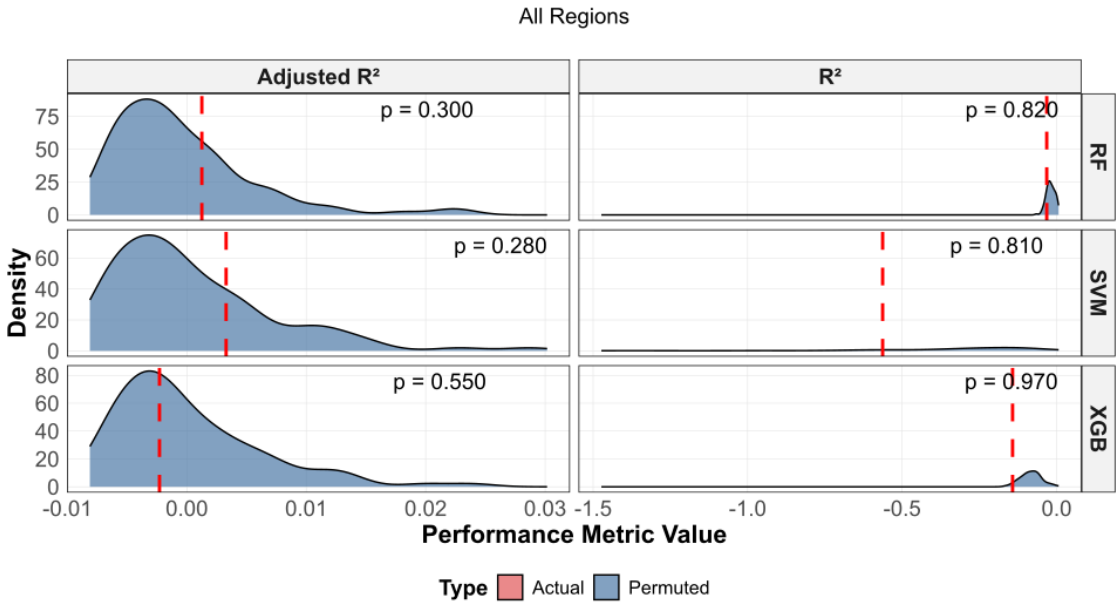

Figure 23: **Permutation Testing for All-Region Models.** Distribution of adjusted R^2 and R^2 values across 100 permutations of the BIS outcome. Vertical red lines indicate the actual (unpermuted) performance. P-values reflect the proportion of permuted values exceeding the performance observed.

As shown in Figure 23, the observed R^2 and adjusted R^2 scores for all models fell well within the distribution of permuted outcomes. The corresponding permutation p-values—0.30 for RF, 0.28 for SVM, and 0.55 for XGB on adjusted R^2—confirm that model performance was not significantly better than chance. This further corroborates the interpretation that this pipeline could retrieve no predictive signal from the transcriptomic features for impulsivity in this dataset.

Summary. Despite employing standard cross-validation procedures and extensive algorithm tuning, none of the models achieved meaningful out-of-sample predictive performance. These results suggest that the current modeling setup failed to prevent overfitting, likely due to both the high dimensionality and noise inherent in the transcriptomic data and the limited sample size. Moreover, the cross-validation design was fixed and not systematically varied; alternative folding schemes—such as double-nested leave-one-out or two-fold strategies—were not explored. As a result, the analysis does not capture how different levels of overfitting susceptibility might influence generalization performance.

Taken together, these findings reinforce the difficulty of predicting impulsivity, as measured by the BIS-11 questionnaire in this high-risk forensic cohort, from static gene expression proxies. The observed limitations highlight not only the sparsity and weak signal typical of behavioural phenotypes, but also the need for more robust validation strategies to quantify and mitigate overfitting in small-sample, high-dimensional settings.

5.5 Discussion

The results of this study underscore the substantial challenges in predicting antisocial traits, proxied here by BIS-11 impulsivity scores, from genetically inferred brain transcriptomic data. Despite a careful modeling pipeline incorporating nested cross-validation, feature selection, and multiple machine learning algorithms, out-of-sample predictive performance remained consistently poor. These findings point to several methodological and conceptual limitations that help contextualize the negative results and provide direction for future research.

Predictive Signal and Dimensionality Challenges. First and foremost, the weak performance across models is attributable in part to the high dimensionality of the feature space relative to the sample size. The predictive models indeed operated on hundreds to thousands of predictors in a dataset with fewer than 500 individuals. In such high-dimensional

settings, even small statistical noise or technical variation—whether in feature selection or in the prediction of individual-level scores—can be amplified, resulting in models that overfit the training data while failing to generalize (Huynh et al., 2020).

Furthermore, the pseudo-validation step showed that even under optimistic conditions, where features were selected using the full training fold, the variance explained by these features rarely exceeded 20–25% (Figure 22A). The difference in feature counts between Boruta alone and the combined approach offers important methodological insights. Correlation filtering likely captures a broad range of linear associations with the outcome, including many spurious signals, which explains the initially high number of selected genes. In contrast, Boruta—designed to identify features with robust importance relative to random noise—may struggle to confidently classify predictors as important in a sparse and noisy transcriptomic setting, where signal-to-noise ratios are low. As a result, Boruta alone yields a more conservative, smaller feature set. Combining correlation filtering with Boruta reduces the predictor space before applying significance testing, likely improving Boruta's ability to distinguish relevant features and yielding a slightly larger but still stringently filtered subset.

Furthermore, when applied to unseen data, predictive performance dropped to near-zero (Figure 22B), with predicted values clustering tightly around the sample mean. This outcome may reflect the instability of feature selection under high noise, the sparsity of truly informative features for the behavioural phenotype in question, or a combination of both (Gupta and Gupta, 2019).

Limitations of GReX in Behavioral Prediction. This work also highlights the limited translational value of GReX when applied to complex behavioral traits like impulsivity. GReX models, trained in postmortem datasets using transcriptomic-genomic associations, capture only the genetically driven portion of expression (Gamazon et al., 2015). While this framework has shown utility in linking genes to disease risk (Gusev et al., 2018; Gandal et al., 2018a), it may be insufficiently nuanced to capture the dynamic, environmentally

modulated, and context-sensitive mechanisms underlying behaviour at the single individual level.

Antisocial behaviour and impulsivity are particularly multifactorial, influenced not only by genetic predispositions but also by environmental, developmental, and psychosocial factors (Raine, 2002; McAdams et al., 2013). The BIS-11 questionnaire, while validated and widely used (Patton et al., 1995), captures self-reported impulsivity across cognitive, motor, and non-planning dimensions—all heavily shaped by sociocultural context (Tuvblad and Beaver, 2013).

Low Heritability of Trait-Level Antisocial Behaviour. A central limitation of this predictive framework stems from the nature of the outcome trait itself. Impulsivity—assessed in this study using scores from the self-report BIS-11 questionnaire—is not strongly heritable. Twin studies estimate that genetic influences on impulsivity-related traits are modest, typically ranging from 30–40% (Rhee and Waldman, 2002; Burt, 2009; Congdon and Canli, 2008), with even lower estimates reported for self-reported measures like the BIS-11. Notably, these heritability estimates reflect both additive and non-additive genetic effects. In contrast, predictive models based on common SNPs and eQTL-derived features are restricted to capturing additive genetic variance, which represents only a subset of the total heritable signal captured by twin designs. This mismatch between broad-sense heritability and the additive-only nature of SNP-based models inherently limits the upper bound of predictive accuracy. Consequently, the overall predictive ceiling of any GReX-based approach for impulsivity—as operationalized in this study—is constrained by the modest genetic architecture of the trait.

Unlike traits such as SCZ, which benefit from relatively well-characterized polygenic architectures and large-scale genomic studies (Sullivan et al., 2003; Trubetskoy et al., 2022), the genetic underpinnings of antisocial behaviour—and related traits like impulsivity—remain less well understood. It is possible that their genetic architecture involves more complex

gene–environment interplay, context-specific regulatory mechanisms, or influences beyond the scope of GReX models, such as post-transcriptional regulation or epigenetic effects (Guloksuz et al., 2019; Schoeler et al., 2019). However, the relative contributions of rare variants or non-genetic factors to these traits remain to be clarified and cannot be inferred from the present data.

Cohort and Phenotyping Constraints. Several cohort-specific limitations may have contributed to the limited predictive performance observed. The sample—adult males incarcerated in North America—represents a unique and highly heterogeneous population in terms of environmental exposure, cultural background, and genetic diversity. Although genomic eigenvectors were included to control for ancestry, subtle population stratification may still have influenced results (Gurdasani et al., 2019).

Furthermore, impulsivity was assessed using the BIS-11, a self-report measure that is subject to response biases—particularly in forensic contexts where strategic underreporting may be more pronounced (Tuvblad and Beaver, 2013). Despite this, the mean BIS-11 total score in this sample (70.0 ± 11.1) was substantially higher than normative values in general population samples (typically 62–64) (Stanford et al., 2009; Vasconcelos et al., 2015), supporting the interpretation of this cohort as an elevated-risk group and consistent with the study's focus on extreme phenotypic expression.

It is important to note, however, that such response biases may not weaken the genetic signal, but rather affect the interpretation of what is being predicted. If self-perceived impulsivity—as captured by the BIS-11—is itself heritable, GReX models may still detect genetic associations. In this case, what is being predicted may reflect genetically influenced self-assessment of impulsivity rather than objective behavioural impulsivity per se.

Conclusion and Future Directions

This study evaluated whether antisocial traits, operationalized via BIS-11 impulsivity scores, could be predicted from genetically imputed brain gene expression using ML pipelines. Despite extensive modeling efforts, no meaningful out-of-sample predictive performance was achieved. This outcome may reflect a convergence of challenges: the weak relationship between genotype and behaviour, the modest heritability of impulsivity, and the limitations of using static transcriptomic proxies to model dynamic behavioural traits.

Nevertheless, the study lays important groundwork for methodological and conceptual refinement. Strengthening the modeling pipeline, reassessing phenotype definition, and establishing internal validation benchmarks all represent key directions for future work. These steps are particularly relevant in light of the complex, high-dimensional nature of behavioural genomic data, where signal is often subtle and difficult to disentangle from noise.

A major methodological concern involves uncontrolled overfitting. The current analysis relied on a fixed cross-validation scheme, which limited insight into how different validation strategies affect model stability. Exploring a range of nested cross-validation designs—such as leave-one-out or two-fold schemes—could help parametrise overfitting risk and clarify how performance varies with partitioning structure and sample size.

The use of positive control tasks may also offer insight into the model's capacity to detect signal under ideal conditions. For example, prediction of genetically structured outcomes—such as Hispanic vs. non-Hispanic group membership, without correcting for ancestry—could serve as a benchmark for signal recovery. These types of validation tasks are especially valuable when working with traits of modest heritability and noisy measurement.

Beyond methodological adjustments, expanding the range of phenotypes under consideration could provide a more comprehensive understanding of the relationship between genotype and antisocial behaviour. While this study focused on BIS-11 scores, the dataset includes additional measures (see section 5.2.2 for details) such as the PCL-R, the IRI, and develop-

mental and environmental indicators like the MOPS, the BSMSS, and estimated IQ. These variables were not formally tested here due to limited sample size, but may represent more stable or informative targets. Clinician-rated instruments like the PCL-R could mitigate some of the noise associated with self-report tools, while traits like empathy or early-life adversity might reflect distinct but genetically relevant aspects of antisocial behaviour.

If sample size increases or comparable datasets become available, future analyses could evaluate whether these alternative phenotypic representations yield stronger transcriptomic signals, especially when modeled alongside environmental moderators or in stratified subgroups. Such extensions may help clarify the boundary conditions under which GReX-based approaches become informative.

Taken together, these next steps aim to move beyond the limitations of the current study by interrogating both modeling strategy and phenotype design. By incorporating more flexible validation procedures, refining the choice and structure of behavioural measures, and integrating contextual environmental information, future work may contribute to a more reliable and interpretable framework for behavioural genomic prediction—one that remains responsive to the complexity of psychological traits and the constraints of available data.

Chapter 6

General Discussion & Conclusion

The General Introduction outlined the challenges inherent in studying the genetic basis of brain-related traits characterized by complex heritability. This thesis explored the hypothesis that the regulatory impact of genetic variation is best captured through network-informed models of gene expression that integrate both local and distal regulatory influences. Rather than focusing exclusively on proximal genes or *cis*-eQTLs, it posits that genetic risk is mediated through structured, modular systems of transcriptional regulation. These expression programs reflect the combined influence of local (*cis*) and long-range (*trans*) regulatory variants, whose complex architecture can be effectively modeled using gene co-expression networks (Pergola et al., 2016, 2017, 2019a, 2023c).

To disentangle such system-level regulatory effects, this thesis proposed leveraging gene co-expression networks derived from human brain transcriptomic data as biological priors for modeling gene regulation at the level of individual genes. These networks capture patterns of transcriptional coordination and serve as scaffolds for identifying coherent regulatory modules. By conducting *trans*-eQTL analyses within these modules, it becomes possible to detect biologically meaningful long-range regulatory effects with greater interpretability and statistical power, thereby addressing key challenges such as effect size dilution and the multiple-testing burden inherent to genome-wide scans (Battle et al., 2014; Võsa et al., 2021;

Liu et al., 2022).

The three studies presented in this thesis operationalize the central hypothesis at progressively broader analytic scales (Figure 24):

Study 1. The first study establishes the core modeling framework by introducing INGENE and MODULE—novel methods for predicting gene expression using co-expression-informed *trans*-eQTL signals. Trained on high-resolution postmortem brain data and rigorously validated across independent datasets, these methods achieve substantially increased transcriptomic coverage and predictive accuracy compared to conventional *cis*-based approaches (Gamazon et al., 2015; Zhang et al., 2019; Huckins et al., 2019).

Study 2. The second study applies these network-informed expression models to the largest SCZ genetic dataset to date (PGC3), in order to investigate associations between genetically predicted expression and diagnostic status. This analysis is grounded in the hypothesis that SCZ heritability is distributed across distinct gene networks, each representing a different biological risk architecture across individuals. The goal was to elucidate how dispersed genetic risk converges on specific biological pathways through which susceptibility genes may mediate their effects.

Study 3. The final study investigates whether the GReX models can generalize to predict individual differences in antisocial behaviour, proxied through the BIS-11 questionnaire for impulsivity categorization, within a challenging real-world forensic population. The individual-level predictive performance in this study was negligible. This null result highlights critical limitations in translating GReX models from population-level association frameworks to individual-level behavioral prediction, particularly in the context of an environmentally shaped and limited sample sizes population.

Summary. Together, the three studies developed in this thesis advocate for a conceptual shift: viewing gene expression as a dynamic bridge between genotype and phenotype—one that is shaped by the complexity and long-range structure of gene regulation. The next section discusses the scientific findings and their significance to the field. The final sections will address the strengths and limitations of the methods developed and propose future research directions emerging from these results.

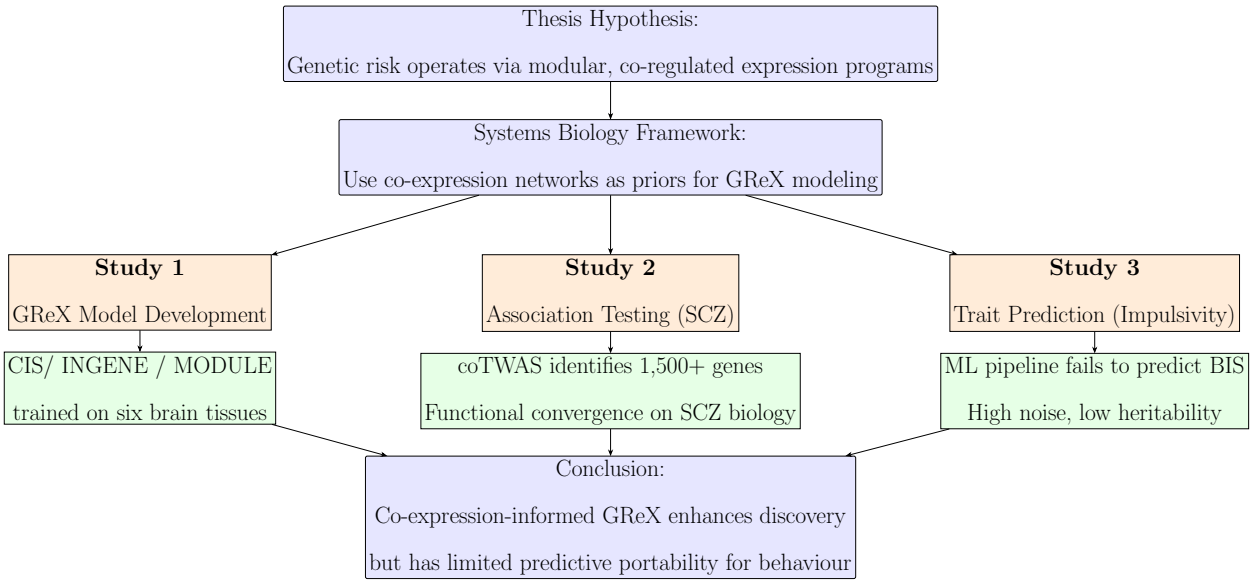

Figure 24: **Conceptual Flow of the Thesis.** From modeling co-expression-informed GReX (Study 1), to disease association testing (Study 2), and behavioral prediction (Study 3), this thesis investigates how modular regulation informs psychiatric biology and its predictive boundaries.

6.1 *Trans*-eQTL Models: Why Might Network-Based Models Recover Additional Genetic Information?

A central methodological innovation of this thesis is the development of two network-aware frameworks—INGENE and MODULE—that extend GReX modeling into the *trans*-regulatory domain. Unlike traditional approaches constrained to *cis*-acting variation, these models are grounded in the hypothesis that polygenic risk manifests not through isolated gene effects

but through perturbation of structured transcriptional programs. By embedding gene-level expression prediction within co-expression networks, INGENE and MODULE exploit regulatory interdependencies that conventional models overlook, offering a tractable framework for capturing distal effects with biological and statistical coherence.

Modeling Assumptions and Conceptual Rationale

At the core of these models is a system-genetic perspective: genetic variants influence gene expression not as isolated units but as elements embedded in complex, modular networks. INGENE leverages genetically imputed *cis*-regulated partner genes as intermediate predictors. Rather than modeling each distal variant–gene pair independently, it exploits the stability of co-expression networks to infer regulatory influence through shared transcriptional partners. MODULE complements this strategy by summarizing gene modules using eigengenes—principal components that capture dominant variance across functionally related gene sets—thereby reducing the dimensionality of the inference space and focusing analysis on coherent regulatory units.

These designs rest on two key assumptions. First, that co-expression reflects biologically meaningful regulatory architectures—whether via TFs, chromatin conformation, or shared signaling pathways. Empirical studies support this assumption: co-expressed genes frequently share upstream regulators (Kustatscher and et al., 2022), exhibit spatial proximity driven by chromatin architecture (Zhang and et al., 2019), and tend to participate in common signaling pathways relevant to diseases (Gandal et al., 2018a).

Second, embedding models within this architecture serves both statistical and biological purposes: it constrains the search space, reduces the dimensionality of expression prediction, and improves signal-to-noise ratios by aggregating weak signals across co-regulated genes (Pergola et al., 2017, 2023b). As demonstrated in Study 1, this network-guided design substantially increases the number of genes with reliable expression prediction and enhances model performance across independent datasets—validating both its statistical utility and

translational relevance.

Toward a Systems-Level Understanding of Transcriptional Regulation

Methodologically, the INGENE and MODULE frameworks reflect a broader shift toward network-constrained learning in transcriptomics, where biological structure is explicitly encoded into predictive models. Rather than treating genes as statistically independent features, these methods leverage co-expression as a functional prior—encoding inter-gene dependencies that shape both model architecture and biological interpretability. This design choice resonates with emerging theories of complex trait architecture, including the omnigenic model (Boyle et al., 2017), which argues that core regulatory processes are not isolated, but are diffusely influenced by peripheral genes acting through regulatory networks (Borcuk et al., 2024).

Importantly, the utility of these models need not rely on strong assumptions about biological modularity being perfectly captured by co-expression structure. Instead, their effectiveness may stem from more pragmatic considerations. For example, in MODULE, a SNP's association with a module eigengene aggregates evidence across multiple target genes. This consolidation may increase statistical power and robustness, even if the underlying regulatory effects are subtle or heterogeneous. In particular, such SNPs may fail to reach significance in single-gene eQTL analyses due to context-specificity—e.g., effects confined to specific cell types, developmental stages, or brain regions. Thus, part of what makes MODULE effective may be its ability to compensate for the limitations of bulk transcriptomic assays, rather than an intrinsic alignment with regulatory modularity.

In contrast, INGENE may be less sensitive to context-specific biases introduced by bulk transcriptomic data, such as the averaging of distinct cell types or developmental stages. Because it leverages genetically imputed *cis*-expression (GReX) from a gene's co-expressed

partners—rather than relying solely on observed expression—it offers a degree of insulation from assay-specific noise. This strategy requires fewer assumptions about the consistency of co-regulation across biological conditions, and may therefore provide a more stable basis for inferring distal regulatory relationships.

This potential for latent specificity opens compelling future directions. With the continued maturation of single-cell and spatial transcriptomics technologies (Stuart and et al., 2019; Maynard and et al., 2021), these frameworks could be adapted to incorporate high-resolution context, enabling the identification of trans-regulatory effects that are cell-type–specific or spatially localized. Additionally, integrating other regulatory modalities—such as chromatin accessibility (e.g., ATAC-seq), enhancer–promoter interactions, or protein–DNA binding—may help resolve causal chains linking non-coding genetic variation to coordinated gene expression.

Ultimately, these models represent a step toward a systems-level synthesis of gene regulation, where prediction and interpretation are co-informed by biological structure. By leveraging modularity as both a statistical tool and a biological principle, INGENE and MODULE offer a scalable framework for decoding the distributed regulatory logic that underlies complex brain traits.

6.2 From Discovery to Prediction: Why Predictive Power Diverges

A central question emerging from this thesis is why the GReX framework, which produced robust group-level associations with SCZ diagnosis in Study 2, showed limited predictive power for individual differences in impulsivity in Study 3. This divergence highlights a critical distinction between explaining population-level mechanisms and predicting individual-level outcomes. Explanatory models aim to identify statistically reliable relationships that reveal causal or mechanistic structure, while predictive models are evaluated by their ability to

generalize to new individuals, often in the presence of noise, complexity, and environmental variability (Shmueli, 2011; Yarkoni and Westfall, 2017).

This distinction is particularly salient in psychiatric genetics, where traits are shaped by complex polygenic architectures and embedded within diverse biological and environmental contexts. In Study 2, the association analysis focused on a well-characterized diagnostic phenotype with established transcriptional correlates, enabling GReX models to capture meaningful variation at the group level. In contrast, Study 3 focused on impulsivity—a multifaceted behavioral construct assessed in a smaller, non-clinical sample—where predictive performance may have been constrained by lower statistical power, higher measurement variability, and greater contextual sensitivity in the phenotype.

Differences in outcome definition and measurement further contributed to this divergence. SCZ diagnosis is a clinically validated, binary phenotype that aggregates heterogeneous symptoms into a standardized diagnostic category and has been extensively studied in genomic and transcriptomic research. Impulsivity, by contrast, is a continuous behavioral trait typically assessed through self-report or task-based measures, both of which are more sensitive to situational variation and measurement error. This contrast—clinical vs. behavioral, binary vs. continuous—can markedly affect the signal-to-noise ratio available to predictive models.

Moreover, the smaller sample size and cross-sectional design of Study 3 limited statistical power to detect modest associations, even if biologically meaningful variation was present. Rather than undermining the utility of transcriptomic imputation, these findings highlight the importance of aligning model design with phenotype characteristics—balancing the power of group-level signal aggregation against the complexity of individual-level behavioral prediction.

Predictive performance also varies substantially across individuals, even within ancestrally matched populations. Variability in local ancestry, LD, gene–environment interactions, and phenotype measurement all contribute to inconsistency in model generaliza-

tion (Mostafavi and et al., 2020). In this thesis, we observed a marked difference in ancestry composition between studies: Study 2 was conducted entirely in individuals of European ancestry (Trubetskoy et al., 2022), whereas PCA revealed broader genomic diversity in Study 3, comprising individuals of European, Central American, and mixed ancestry (see Appendix D.1). This population heterogeneity reduces the portability of genetic predictors, particularly when models—such as those used in Study 1—are trained on European reference panels (see Section 3.2.1 for sample inclusion details). These ancestry differences can impair predictive performance (Mostafavi and et al., 2020; Lupi et al., 2024), not because of flaws in model design, but due to mismatches between the ancestry of the training data and that of the individuals being predicted.

Altogether, the divergence between Study 2 and Study 3 underscores the need to approach discovery and prediction as distinct but complementary objectives. Group-level association studies, such as those used to investigate SCZ, can reveal interpretable links between genetic variation and biological processes. However, translating these findings into reliable individual-level predictions—particularly for behavioral traits—requires more than statistical replication. As demonstrated in Study 3, limitations in sample size, ancestry match, and phenotype measurement precision can significantly constrain predictive performance, even when biologically informed models are used.

While increasing model complexity or incorporating environmental covariates is often proposed as a solution, our findings suggest that predictive gains may also depend on scaling up sample sizes, improving the quality of expression predictors, and tailoring model architectures to the trait of interest. Rather than assuming any single explanation, these results call for systematic comparisons across modeling strategies, trait definitions, and population compositions. Bridging the gap between statistical discovery and individual-level prediction will depend not on embracing complexity as a barrier, but on identifying which factors most directly limit generalizability—and optimizing accordingly.

The following paragraphs briefly revisit the key strengths and limitations identified in

the analyses presented in Study 2 and Study 3.

Study 2: Conditions Guiding Accurate Association. The coTWAS findings presented in **Study 2** highlight a fundamental yet frequently overlooked statistical principle: *association analyses at the population level are strengthened by aggregating numerous weak genetic effects*. In this context, three essential aspects contribute uniquely to the power of the co-expression-informed association framework used here:

1. **Systems-level nature of SCZ.** SCZ is known to exhibit a strong polygenic signal with coherent neurobiological underpinnings and it is increasingly conceptualized not as the result of isolated gene dysfunction, but as a disorder of large-scale transcriptional network perturbation emerging across neurodevelopment (Birnbaum and Weinberger, 2017; Gandal et al., 2018a; Fromer et al., 2016). During brain maturation, gene co-expression patterns shift dynamically (Pergola et al., 2023a), reflecting changing cellular composition, synaptic remodeling, and region-specific regulatory programs. In this context, genetic risk may not act directly on single genes, but instead influence the trajectories of co-regulated gene groups (Borcuk et al., 2024), particularly through central "hub" genes that orchestrate broader transcriptional programs. The co-expression network approach adopted in this thesis captures these interdependencies by embedding genes within structured modules that reflect shared regulatory history. This biologically informed dimensionality reduction compresses millions of potential SNP–gene relationships into interpretable regulatory units, mitigating the multiple-testing burden and enhancing the detection of subtle, developmentally mediated *trans* effects (Pergola et al., 2023b).

2. **Large effective sample size.** By meta-analyzing 62 independent cohorts from the PGC3 (total $N = 102,613$), we translated subtle, dispersed genetic influences into robust, genome-wide significant associations. Because sampling error decreases proportionally to $1/\sqrt{N}$, expanding the sample size through additional cohorts yields

substantially greater gains in statistical power than could be achieved through any incremental optimization of individual analytical parameters (Visscher et al., 2017).

3. **Outcome definition: categorical diagnosis vs. dimensional behaviour.** Another key factor contributing to the stronger performance of Study 2 is the nature of the outcome variable. SCZ is a clinically defined categorical diagnosis with operationalized diagnostic criteria established through decades of psychiatric nosology (Rajiv Tandon, 2013). While the condition remains biologically heterogeneous, its categorical boundaries help constrain phenotypic variability and reduce noise in association analyses. This structure improves the signal-to-noise ratio and allows statistical models to detect group-level effects more effectively, especially when supported by large sample sizes (Trubetskoy et al., 2022). In contrast, Study 3 focused on impulsivity—a dimensional, self-reported behavioral construct assessed using the BIS-11. Unlike categorical diagnoses, such traits are often shaped by transient states, environmental context, and respondent interpretation, which collectively introduce significant measurement error (Tuvblad and Beaver, 2013).

Study 3: Learning from Predictive Limitations In contrast to the robust group-level findings observed in Study 2, **Study 3** offered a valuable opportunity to test the boundaries of GReX-based prediction in a behavioral context. Attempts to predict BIS-11 impulsivity scores in a smaller, heterogeneous sample ($N = 468$) did not yield significant results. However, the analysis highlighted several biological and statistical constraints that elucidated the factors limiting predictive performance:

Low trait heritability. Twin studies consistently estimate modest heritability for impulsivity-related traits (Congdon and Canli, 2008), with genetic factors accounting for up to 30% of phenotypic variance. However, due to the *missing heritability* phenomenon, GReX models built from common SNPs typically capture only a fraction (20–30%) of this heritability (Dudbridge, 2013). Even under ideal conditions, the maximum R^2 achievable

for individual-level prediction remains below 10%—a ceiling insufficient to overcome the substantial noise inherent in self-reported behavioral assessments.

Mismatch between biological signal and behavioral complexity. GReX captures stable, genetically determined transcriptional regulation (Gamazon et al., 2015), whereas impulsivity is a dynamic, context-sensitive phenotype influenced by trauma, substance use, incarceration, and broader environmental exposures (Burt, 2009; Caspi et al., 2002; Kendler and Baker, 2007). Critically, growing evidence suggests that environmental risk factors induce persistent biological changes through epigenetic mechanisms. For example, maternal behaviour has been shown to produce stable alterations in histone acetylation and DNA methylation patterns that impact offspring behaviour (Weaver et al., 2004). Similarly, early life adversity—including sexual abuse, physical maltreatment, and bullying—has been linked to lasting epigenetic modifications that reshape gene expression profiles and behavioral outcomes (Burns et al., 2018). This evidence highlights a fundamental limitation of static GReX models: by focusing solely on inherited genomic variation, they are unlikely to capture the epigenetically mediated impact of environmental experience on behaviour.

Curse of dimensionality. Following quality control, approximately 20,000 gene expression predictors per brain region were available for each individual, totaling over 113,000 predictors across all regions. This extremely high predictor-to-sample ratio poses a major challenge for machine learning models (Domingos, 2012; Wainberg et al., 2018). Even with regularization and feature selection, the risk of overfitting remains substantial, increasing the likelihood that models will capture noise rather than true biological signal (Hawkins, 2004).

Population and environmental divergence. GReX models were trained on postmortem brain tissue from individuals of genetically confirmed European ancestry. In contrast, although PCA indicated majority of European-like structure, the forensic cohort self-

identified as Latino/not-Latino—a group with extensive admixture involving European, Native American, and African ancestries (Moreno-Estrada et al., 2013). Subtle differences in local ancestry, LD structure, and allele frequency distributions can substantially reduce the portability of GReX models across admixed populations (Martin et al., 2019). Moreover, environmental exposures unique to the forensic cohort (e.g., trauma, incarceration-related stress, substance use) may have induced additional epigenetic and transcriptional divergence that expression models trained in different population contexts are unlikely to capture. Together, these genetic and environmental mismatches compounded the transportability problem, further limiting predictive accuracy.

These limitations do not undermine the value of the modeling framework, but rather highlight the need for greater specificity in how predictive models are constructed and applied. For instance, genetic effects on impulsivity may be more detectable in specific subgroups, such as individuals with histories of maltreatment or high environmental stress exposure. Research has shown that childhood adversity can moderate the behavioural effects of MAOA variants (Nilsson et al., 2018), suggesting that gene–environment interactions (G×E) may amplify predictive signals in stratified contexts. More broadly, G×E effects often surface only under specific environmental conditions such as familial instability or prenatal stress (McGue and Carey, 2017; Ruisch, 2020).

This suggests a shift toward integrative and stratified modeling approaches. Future efforts may benefit from combining GReX with environmental covariates or developing dynamic models that account for life history, social context, and epigenetic change. Stratification by trauma exposure or comorbidity has already improved genetic signal detection in studies of ADHD and related traits (Franke and Buitelaar, 2018; van Hogezand, 2016), and emerging work on behavioral plasticity in adolescence suggests that developmental stage and social environment play a critical role in shaping genetic expression (Richards, 2015; Gidziela, 2024).

Ultimately, Study 3 highlights the need to align predictive tools with the complexity of their target phenotypes. While GReX-based models show promise for capturing population-level effects, individual-level prediction—particularly for context-dependent behavioral traits—may require models that are both biologically informed and contextually responsive.

Summary. Collectively, this thesis delivers a two-fold message. **First**, network-informed GReX models are highly effective tools for biological discovery: in large-scale association studies, they aggregate weak and distributed genetic signals into coherent, mechanistically interpretable insights that would otherwise remain undetected. **Second**, when the goal shifts from group-level inference to individual-level prediction, integrative approaches may be necessary. Progress in this domain will likely require multimodal frameworks that incorporate genetic, epigenetic, and environmental data to capture the full multidimensionality of human phenotypic variation.

6.3 Limitations

While the network-based methodologies developed in this thesis demonstrate significant potential for uncovering *trans*-regulatory mechanisms, several important limitations must be considered.

Limited Sample Size and Tissue Representation A fundamental limitation of the predictive models developed in this thesis lies in their dependence on post-mortem brain transcriptomic data, which, while biologically relevant, remain constrained in both scale and anatomical coverage. Due to ethical, logistical, and technical challenges, brain tissue resources are far more limited than those available for blood or other peripheral tissues (GTEx, 2020). This restricts the effective sample size for model training, particularly in the case of *trans*-eQTL mapping, where effect sizes are typically small and highly context-dependent.

While the integration of co-expression structure helps mitigate some of this sparsity by enabling statistical aggregation, the detection of long-range regulatory effects is nonetheless power-limited when data are scarce.

Moreover, uneven tissue representation across brain regions introduces another critical constraint. Gene regulatory dynamics vary substantially across spatial and functional domains of the brain (Hawrylycz et al., 2012; Maynard and et al., 2021), and the absence of uniformly sampled regions limits the model's capacity to capture region-specific transcriptional programs. This is particularly consequential for psychiatric traits like SCZ, which are known to involve distinct circuits and cell populations distributed across tissues (Skene et al., 2018). Models trained on aggregated or partially sampled tissue may thus dilute or obscure biologically specific regulatory relationships that are essential for fine-grained mechanistic interpretation.

In addition, the post-mortem nature of the reference data imposes inherent temporal limitations. Regulatory processes that are dynamic across development, aging, or disease progression cannot be fully captured in static adult tissue samples (Pergola et al., 2023a), potentially biasing models toward stable regulatory architectures and away from those that mediate trait-relevant plasticity (Birnbaum and Weinberger, 2017; Li et al., 2018b). Together, these factors underscore the need for future efforts to incorporate larger, more diverse transcriptomic datasets—ideally incorporating longitudinal, single-cell, and spatial resolution—to fully realize the promise of GReX models in complex brain traits.

Dependence on Bulk RNA-seq Data A key limitation of the current modeling framework lies in its reliance on bulk RNA-seq data to identify co-expression *trans*-eQTLs. While bulk transcriptomic data provide high-throughput, regionally resolved snapshots of gene activity, they conflate signals from multiple cell types—thereby integrating both genuine co-regulation and variation in cellular composition (Avila Cobos et al., 2020). As a result, the modules derived in INGENE and MODULE may capture not only transcriptional regula-

tion but also shifts in the abundance of cell populations, particularly across heterogeneous brain regions.

Although this admixture does not undermine the utility of the models for discovering broad transcriptional patterns, it limits interpretability at the level of specific regulatory mechanisms. For example, it is difficult to determine whether a *trans*-regulatory effect inferred by these models reflects upstream transcriptional control, cell-type–specific activation, or simply co-variation in cell population proportions. This ambiguity is particularly relevant in the brain, where gene expression is tightly coupled to cell identity and where risk for psychiatric disorders like SCZ is known to converge on discrete neuronal subtypes (Skene et al., 2018; Gandal et al., 2018a).

Uncertainty Associated with Gene Expression Imputation A central dependency of both INGENE and MODULE is the use of GReX imputed from genotype data—a process that introduces multiple layers of uncertainty. These include potential errors from genotype imputation, limitations in the reference transcriptome panels, and assumptions embedded within the underlying predictive models (Gamazon et al., 2015; Barbeira et al., 2018). Additionally, expression imputation accuracy varies substantially across genes and tissues, with particularly low reliability observed in transcripts exhibiting low expression heritability, high inter-individual variability, or strong environmental sensitivity (Wainberg et al., 2019; Mostafavi and et al., 2020).

Such uncertainty poses significant challenges for downstream interpretation, particularly in *trans*-eQTL contexts where effect sizes are small and model signal may be further attenuated by biological heterogeneity. While model regularization and validation across independent cohorts partially mitigate this issue, the imputed values remain probabilistic estimates rather than direct molecular readouts.

Restricted Ancestral Generalizability A notable limitation of the current study is its exclusive focus on individuals of European ancestry, a constraint driven by the demographic

composition of available reference panels and validation cohorts. While this choice enhances internal consistency, it substantially limits the external validity and equitable applicability (Popejoy and Fullerton, 2016) of the INGENE and MODULE frameworks. Genetic architectures—including allele frequencies, LD patterns, and eQTLs—are known to vary significantly across ancestral populations (Martin et al., 2019; Benjamin et al., 2023). As a result, models trained in one population may exhibit reduced accuracy or even systematic bias when applied to others.

Reliance on Fixed Network Priors and Absence of Causal Inference The current implementation of INGENE and MODULE relies on fixed co-expression modules sourced from previously published bulk RNA-seq datasets (refer to Table ST3 for details about the study chosen). While this decision facilitates the generalization of previous findings and computational tractability, it also imposes structural constraints that may propagate biases inherent to the original data—such as limited cell-type resolution, cohort-specific noise, or developmental stage effects (Oldham and et al., 2008; Langfelder and Horvath, 2008). These fixed priors may obscure regulatory heterogeneity and prevent the discovery of novel or context-dependent gene–gene interactions.

Beyond the structural assumptions inherent to fixed network priors, it is also essential to acknowledge a conceptual limitation shared by both INGENE and MODULE: these frameworks infer statistical associations rather than causal regulatory relationships. While the models are designed to prioritize biologically interpretable signals by leveraging co-expression structure, the directionality and functional impact of the identified *trans*-eQTLs remain unresolved without experimental validation. In some cases, the predicted regulatory nodes may act as proxies for latent confounding factors, or they may reflect correlation structures within gene modules that do not correspond to direct regulatory mechanisms.

Bridging this interpretive gap will require the integration of causal inference methodologies and experimental perturbation strategies. Functional follow-up studies—such as

CRISPR-based regulatory element screens or perturb-seq platforms—offer scalable means of testing whether candidate variants or regulators exert causal effects on downstream gene expression (Gasperini et al., 2019).

Summary. Collectively, these limitations do not undermine the potential of the developed methodologies but rather delineate their current boundaries. Recognizing and systematically addressing these challenges will be crucial for transitioning from associative findings to robust mechanistic insights and eventually informing therapeutic strategies. The following section outlines opportunities for future research to address and mitigate them.

6.4 Future Directions

Building on both the limitations outlined in the previous section and the systems-level framework established in this thesis, three priority areas emerge for advancing co-expression–informed GReX modeling: (i) enhancing biological specificity, (ii) improving ancestry generalizability, (iii) and expanding clinical applicability. These directions reflect not only the methodological constraints encountered, but also the conceptual and translational opportunities uncovered throughout this work.

Cross-Ancestry Portability and Global Applicability

Extending predictive genomic models beyond European ancestry is both a scientific imperative and an ethical necessity. The long-standing Eurocentric bias in GWAS, TWAS, and eQTL resources has restricted discovery, limited clinical relevance for under-represented populations, and exacerbated disparities in biomedical research (Popejoy and Fullerton, 2016; Gurdasani et al., 2019; Kachuri et al., 2024). These imbalances not only reduce the equity of genetic applications but also undermine the robustness and replicability of findings across human populations.

Traditional GWAS frameworks are highly sensitive to ancestry-specific patterns of LD, which constrains the transferability of association signals. In contrast, GReX-based approaches that model genetically regulated expression offer a mechanistically informed alternative. Rather than relying solely on variant proximity, these models leverage intermediate molecular phenotypes—such as transcriptional programs and co-regulated gene modules—that may be more evolutionarily conserved across populations (Bhattacharya et al., 2022; Zeng et al., 2022).

Recent evidence suggests that the regulatory architecture governing gene expression is more portable across ancestries than raw GWAS signals, particularly when intermediate phenotypes are modeled (Kachuri et al., 2024; Benjamin et al., 2024). Coexpression-informed models, by distilling complex regulatory relationships into stable transcriptional modules, may transcend population-specific LD patterns and enhance the cross-ancestry portability of transcriptome-based predictions.

Speculatively, one may conceptualize co-expression networks as encoding "biological priors" that reflect constraints imposed by the functional organization of the genome, rather than the statistical structure of allele frequencies or LD patterns that dominate population-specific GWAS signals. Whereas traditional association studies rely on ancestry-dependent correlations between variants and traits, co-expression modules capture patterns of gene regulation that arise from conserved cellular programs—such as neurodevelopment, synaptic signaling, or immune response—that are largely shared across human populations (Russell et al., 2023).

These transcriptional modules are highly conserved across human populations and even across species, suggesting they are shaped by evolutionary constraints rather than population-specific genetic architecture (Crow et al., 2022; Oldham et al., 2006; Stuart et al., 2003; Russell et al., 2023). Importantly, patterns of coordinated gene expression are preserved even in the absence of homologous tissues or identical regulatory inputs, underscoring the robustness of modular gene regulation (Stuart et al., 2003). In this sense, coexpression networks can

be conceptualized as encoding "biological priors": regulatory frameworks imposed by functional necessity rather than demographic history. By modeling these conserved regulatory programs, coexpression-informed GReX approaches are well-positioned to generalize across ancestries, potentially outperforming variant-centric methods that rely on population-specific linkage disequilibrium.

In this sense, the transcriptional architecture inferred from co-expression networks may represent a form of "universal regulatory logic", grounded in biological function rather than genomic background. Realizing this potential will require coordinated investment in diverse transcriptomic and eQTL resources. Specifically, future work should aim to: (i) construct multi-ancestry reference panels; (ii) incorporate local ancestry-aware modeling into *trans*-eQTL frameworks; and (iii) empirically benchmark the portability of network-informed predictions across globally representative cohorts. Such efforts will be essential to ensure that the interpretability and predictive power of system-level models extend equitably across the full spectrum of human genetic diversity.

Toward Cell-Type and Isoform-Specific Regulation

Advancing the biological resolution of gene expression prediction is a critical frontier for improving both the interpretability and translational utility of GReX-based models. While *cis*-eQTLs typically exert relatively stable, gene-proximal effects, *trans*-eQTLs are markedly more context-dependent—varying across cell types, developmental stages, and environmental exposures (Yao et al., 2017; Võsa et al., 2021; Ouwens et al., 2020). Accurately modeling these distal effects requires a regulatory framework that accounts for the cellular and spatial environment in which gene expression is regulated. Without this resolution, genetically predicted expression may fail to capture the relevant biology linking genotype to complex traits.

Most TWAS to date have relied on bulk RNA-sequencing data, which average gene expression across heterogeneous cellular populations. This averaging masks critical variation in

cell-type-specific gene regulation, particularly from rare, transient, or functionally specialized cell types that may be disproportionately involved in disease pathogenesis (Lee, 2022; Skene et al., 2018). In the context of brain disorders, where specific neuronal and glial subtypes underlie distinct functional and pathological circuits, the lack of cellular resolution presents a major barrier to mechanistic interpretation (Maynard and et al., 2021; Bryois et al., 2022).

Recent advances in single-cell RNA-sequencing (scRNA-seq), spatial transcriptomics, and integrative deconvolution methods offer the opportunity to model gene expression at single-cell resolution and within anatomical context (Stuart and et al., 2019; Maynard and et al., 2021). Embedding co-expression network construction within this framework could yield regulatory modules that reflect true cell-type and spatial specificity, thereby enabling GReX models to capture the relevant transcriptional programs active in specific neural subpopulations. Such refinement is particularly promising for psychiatric genetics, where clinical heterogeneity and poor diagnostic boundaries obscure molecular stratification.

Equally fundamental is the regulation of gene expression at the isoform level. Alternative splicing dramatically expands the proteomic and regulatory complexity of the genome, and a substantial proportion of genetic regulation operates not at the gene level, but through transcript-specific mechanisms (Glinos and et al., 2022; Takata and et al., 2017). Aggregating expression across all isoforms of a gene risks obscuring biologically relevant regulatory variation. Indeed, isoform-resolved studies have shown that transcript-level eQTLs often reveal novel signals and trait associations that are undetectable at the gene level (Bhattacharya et al., 2023). As long-read sequencing technologies mature and provide more accurate isoform quantification, the construction of isoform-specific expression panels will become feasible, enhancing the granularity and accuracy of GReX prediction.

Speculatively, integrating co-expression networks with cell-type and isoform-level resolution could enable the development of multi-dimensional regulatory models that more faithfully reflect the complex transcriptional architecture of the brain. Such models would allow for the detection of regulatory programs that are temporally restricted, cell-type–specific,

and isoform-selective—attributes essential for understanding psychiatric phenotypes, which often arise from the convergence of polygenic risk onto distinct cellular systems during key developmental periods (Birnbaum and Weinberger, 2017; Akbarian et al., 2015).

An emerging hypothesis from this work is that *trans*-eQTLs may be particularly enriched for isoform-level regulatory effects. Unlike *cis*-eQTLs, which often act through proximal regulatory elements to influence overall gene expression, *trans*-eQTLs are more likely to involve distal intermediates—such as splicing factors, RNA-binding proteins, or chromatin remodelers—that modulate transcript-specific outcomes (Takata and et al., 2017; Wainberg et al., 2019). These regulators exert their influence not by changing total gene abundance, but by altering isoform usage, exon inclusion, 3′ UTR length, or transcript stability.

This distinction is especially relevant in the brain, a tissue characterized by high transcriptional complexity and extensive alternative splicing (Glinos and et al., 2022). Neuronal function depends on tightly controlled isoform programs that are spatially, developmentally, and functionally specialized. Thus, genetic variants that subtly alter isoform ratios—without impacting total gene-level expression—may still have profound phenotypic consequences. The brain therefore provides a particularly promising landscape for *trans*-eQTL discovery at the isoform level.

Co-expression–informed frameworks such as INGENE and MODULE, which leverage the modular organization of gene expression, may be uniquely sensitive to these transcript-specific effects. When gene-level expression is aggregated across multiple isoforms, subtle regulatory relationships can be diluted or entirely lost. By contrast, isoform-level resolution preserves this regulatory granularity, offering clearer mechanistic links between *trans*-acting variants and their downstream targets—especially when those effects are confined to a single transcript within a broader gene locus.

Altogether, these insights suggest that *trans*-eQTLs offer an interesting entry point into understanding isoform-level regulation, particularly in brain contexts where splicing complexity and functional specialization intersect. Future work incorporating isoform-resolved quan-

tification into network-based *trans*-eQTL models has the potential to reveal novel dimensions of transcriptional control and improve the biological specificity of GReX predictions. In doing so, such models could move beyond gene-centric abstraction to more accurately capture the fine-grained molecular disruptions that drive complex brain disorders.

Clinical Translation and Stratification

A central objective of GReX modeling is to move beyond statistical discovery toward clinically actionable insight. While GWAS have significantly advanced psychiatric genetics by identifying common risk loci, their reliance on case–control allele frequency contrasts provides only an indirect and often biologically opaque view of disease mechanisms (Visscher et al., 2017; Tam et al., 2019). One of the most persistent limitations of GWAS is its dependence on broad phenotypic aggregation: in heterogeneous conditions like SCZ, collapsing diverse symptom trajectories into a binary case label obscures meaningful subtypes and dilutes the signal from rare or context-specific effects (Boyle et al., 2017).

In contrast, transcriptome-wide models that estimate GReX offer a mechanistically grounded alternative. By modeling the cumulative effects of polygenic variation on gene expression, GReX provides biologically interpretable molecular phenotypes that are sensitive to context, tissue specificity, and regulatory structure (Barbeira et al., 2018; Zhang et al., 2019; Huckins et al., 2019). These features make GReX particularly valuable for tasks where GWAS often underperforms—such as patient stratification, early risk detection, and treatment response prediction.

By aggregating the effects of thousands of small-effect variants into coherent gene-level signals, GReX enables the identification of transcriptional programs that underlie shared disease liability. For instance, individuals at clinical high risk for psychosis might be stratified according to GReX-derived transcriptional profiles: upregulation of neuroinflammatory modules may signal elevated conversion risk, whereas preserved synaptic integrity might suggest relative resilience. In this way, GReX supports a shift from syndromic classification

to biologically informed subtyping.

Embedding GReX within co-expression and *trans*-eQTL frameworks—such as those developed in INGENE and MODULE—further amplifies its clinical potential. Co-expression networks capture modules of coordinately regulated genes that reflect shared biological functions and regulatory control across tissues, cell types, and developmental stages (Fromer et al., 2016; Hartl et al., 2021; Pergola et al., 2023a). In psychiatric disorders, which are increasingly understood as syndromic entities involving multiple biological subtypes, this systems-level perspective provides a principled approach to molecular stratification. Co-expression modules enriched for dopaminergic signaling, glutamatergic dysfunction, immune dysregulation, or disrupted neurodevelopment can define distinct patient subgroups that may share surface symptoms but diverge at the molecular level (Sportelli et al., 2024).

This perspective has particular relevance for understanding treatment response. Antipsychotics targeting D_2 dopamine receptors remain the first-line intervention for SCZ (Miyamoto et al., 2012), yet roughly one-third of patients respond poorly, and relapse rates remain high even among initial responders (Buckley and Miller, 2017; Leucht et al., 2022). This variability strongly suggests underlying biological heterogeneity. Attempts to define genetically informed subtypes using PRS have provided some value (Chen et al., 2020; Lu et al., 2023), but PRS models are agnostic to regulatory context and aggregate genome-wide variation indiscriminately, potentially obscuring pathway-specific mechanisms (Boyle et al., 2017).

In contrast, GReX models guided by co-expression networks can localize genetic risk to interpretable regulatory modules. For example, Sportelli et al. (2024) identified a co-expression module enriched for dopaminergic signaling in the striatum that was both genetically associated with SCZ and predictive of dopamine synthesis capacity and striatal activation.

Such mechanistically anchored signatures may help predict differential treatment response. Patients with elevated dopamine-related GReX activity may benefit more from D_2 antagonists, while those with immune- or glutamate-related transcriptional profiles may

respond better to adjunctive or alternative interventions.

In summary, co-expression networks may serve as a functional scaffold for precision psychiatry. Unlike PRS, which summarizes risk across thousands of loci without regard for biological coherence, co-expression-informed GReX embeds polygenic signal into structured modules that mirror the brain regulatory architecture. These modules provide interpretable units of molecular stratification, reflect stable biological programs across individuals, and may be leveraged as dynamic biomarkers that evolve with disease progression or treatment exposure. Their modularity facilitates hypothesis-driven exploration of drug targets, and their compatibility with other omics layers opens pathways for integrative multi-modal modeling.

6.5 Conclusion

Together, these future directions emphasize a central theme: gene expression prediction is not merely a statistical tool for association, but a lens through which to interpret the biological mechanisms linking genotype to complex traits. Coexpression-informed *trans*-eQTL modeling, as developed in this thesis, offers a robust and adaptable foundation for addressing emerging challenges in genomics—from global population diversity and cellular specificity to translational utility in health care. Efforts in refining these models, expanding their scope, and integrating new data modalities will be essential to unlocking their full potential in both discovery and clinical contexts.

This thesis presents a novel functionally enriched framework for modeling the transcriptional consequences of genetic variation in the brain. By leveraging co-expression networks and *trans*-eQTL integration, the work pushes the boundaries of TWAS methodologies and offers new tools for decoding the molecular underpinnings of psychiatric and behavioral phenotypes.

As multi-omic datasets continue to grow and become more representative, the strategies

developed here lay a strong foundation for next-generation efforts in precision psychiatry, systems neuroscience, and translational behavioral genomics.

Appendix A

Extra Tables

A.1 SCZ Transcriptome-Wide Association Studies

Table 8: **Summary of Major TWAS Studies in Schizophrenia**

Study	**Tissue / Dataset**	**Key Results**	**Pathways / Innovations**
Gusev et al., 2018	GWAS summary statistics from 79,845 individuals (PGC)	Applied TWAS across SCZ, ASD, and BD; Identified 157 TWAS-significant genes, including 35 outside GWAS loci	42 genes linked to chromatin features.
Gandal et al., 2018	PsychENCODE (brain transcriptomics)	193 SCZ-associated genes	Synaptic signaling and neuroimmune pathways
Collado-Torres et al., 2019	BrainSpan (DLPFC and HP)	Integrated GWAS with four expression feature types; found 1,656 features across 624 genes associated in both regions	Highlighted novel SCZ risk loci; neurodevelopmental processes
Huckins et al., 2019	CommonMind (DLPFC)+ 12 GTEx tissues	TWAS across 12 tissues; 413 gene–tissue associations, 67 independent signals, 19 novel genes	Tissue-specific gene expression associations; highlighted the importance of integrating multiple brain regions in TWAS analyses
Hall et al., 2020	UK Brain Banks (DLPFC)	Identified 89 significant genes, 20 novel; prioritized synaptic signaling pathways	Neurotransmission, synaptic vesicle cycling, synaptic integration
Bhattacharya et al., 2023	PsychENCODE + AMP-AD (adult and developmental prefrontal cortex	Developed isoTWAS, an isoform-level TWAS; identified dozens of isoform-specific associations in SCZ and other traits; prioritized specific isoforms of *AKT3*, *CUL3*, and *HSPD1* for SCZ	Highlighted distinct regulatory mechanisms at the isoform level; emphasized alternative splicing and isoform diversity in SCZ genetic risk

Appendix B

Supplementary Information (Study 1)

B.1 RNA-Seq Data Processing

Gene Expression Quantification Gene-level mRNA expression quantification was performed using the `recount3` R package (Wilks et al., 2021). Raw counts were converted to RPKM and subsequently to transcripts per kilobase million (TPM) to ensure consistency across datasets. To reduce the impact of low-expression and zero-inflated genes, we retained only genes with a median RPKM ≥ 0.1 and with fewer than 20% zero values across samples. Expression values were $\log_2$-transformed with an offset of 1. Samples were removed as outliers if their inter-array distance exceeded 3SD from the mean. Mitochondrial genes were filtered out prior to downstream analysis.

Confounder Adjustment and Residualization Expression values were residualized using linear regression models. For the LIBD dataset, covariates included diagnosis, sex, age, RNA Integrity Number (RIN), mitochondrial mapping rate, ribosomal RNA (rRNA) rate, gene mapping rate, five ancestry principal components (PCs), and the top three expression PCs.

In the GTEx dataset, log-transformed RPKM values were residualized using a linear model that included sex, mean age, RNA integrity number (RIN), rRNA content, postmortem interval (PMI), the first five genetic ancestry components, and the top three expression PCs.

For the CMC dataset, residualization of log-transformed RPKM values included diagnosis, sex, age, PMI, RIN, rRNA rate, the ratio of exon-mapped reads to total reads, intronic and intergenic mapping rates, the first five ancestry components, and the top three expression PCs.

For each dataset, we evaluated the correlation between estimated neuronal proportions (`neu`) and the top three principal components (PCs) of gene expression. In the LIBD dataset, Pearson's correlation coefficients were high across all regions: amygdala ($r = 0.85$), caudate

nucleus ($r = 0.64$), dorsal anterior cingulate cortex (dACC; $r = 0.76$), dorsolateral prefrontal cortex (DLPFC; $r = 0.70$), subgenual ACC (sACC; $r = 0.90$), and hippocampus (HP; $r = 0.83$), all with $p < 0.05$. Similar results were observed in the GTEx dataset: amygdala ($r = 0.87$), ACC ($r = 0.92$), caudate nucleus (CN; $r = 0.84$), DLPFC ($r = 0.91$), and HP ($r = 0.73$). In the CMC dataset, strong correlations were also found in ACC ($r = 0.75$) and DLPFC ($r = 0.62$). Given this redundancy, `neu` was excluded from all residualization models.

After covariate adjustment, residuals were normalized using the Blom transformation (Pergola et al., 2017), which approximates a normal distribution by transforming each residual value x_i based on its rank within the sample:

$$x_i^* = \Phi^{-1}\left(\frac{r_i - 3/8}{n + 1/4}\right)$$

where r_i is the rank of x_i, and Φ^{-1} is the inverse cumulative distribution function of the standard normal distribution.

B.2 Genotyping and Imputation Procedures

Genotype data for the LIBD cohort were generated using Illumina BeadChips and subsequently imputed to the Trans-Omics for Precision Medicine (TOPMed) reference panel (Taliun et al., 2021) and the Haplotype Reference Consortium (HRC) panel McCarthy et al. (2016).

For the GTEx dataset, genotypes were processed following the GTEx pilot analysis protocol (GTEx, 2015), with imputation performed using IMPUTE2 (Howie et al., 2009) and the 1000 Genomes Project Phase 1 reference panel (Auton et al., 2015b).

CMC genotype data were obtained from DNA extracted from the dorsolateral prefrontal cortex (DLPFC), genotyped using the Illumina Infinium HumanOmniExpressExome platform, and imputed to the HRC panel, as previously described (Fromer et al., 2016).

B.3 Elastic Net Model Training

The relationship between genotype and gene expression was modeled using *elastic net* regularization (Li et al., 2018c), a penalized linear regression technique that balances variable selection (via the L1 penalty of LASSO) with coefficient shrinkage (via the L2 penalty of ridge regression). This approach is particularly well-suited to high-dimensional genomic data where the number of predictors far exceeds the number of observations, and where correlated features (e.g., SNPs in linkage disequilibrium) are expected.

For each gene g, the model can be expressed as:

$$\hat{y}^{(g)} = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

where $\hat{y}^{(g)}$ is the predicted (normalized) expression of gene g, X_j are the genotype dosages for the selected SNPs, and β_j are the fitted coefficients. The elastic net objective function is given by:

$$\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + \frac{1-\alpha}{2} \sum_{j=1}^{p} \beta_j^2 \right) \right\}$$

Here, λ controls the overall strength of the penalty, while $\alpha \in [0, 1]$ determines the balance between L1 and L2 penalties. In this study, elastic net was implemented using the `glmnet` R package, with λ hyperparameter optimized via nested cross-validation and $\alpha = 0.5$ consistent with prior studies (Gamazon et al., 2015; Huckins et al., 2019; Zhang et al., 2019).

B.4 Lambda Tuning for MODULE Training

To fine-tune the elastic net regularization parameter λ in MODULE, we used the same fold indices employed during the co-eQTL discovery step. Specifically, we generated a sequence of 100 candidate λ values using the `lambdaseq` function from the `method` R package, setting

parameters to $\alpha = 0.5$, `lambdaRatio` $= 10^{-2}$, and `nLambda` $= 100$.

In each outer fold, we recomputed the module eigengene (PC1) on the training set and projected its loadings onto the testing set to generate a consistent PC1 for evaluation. To ensure reproducibility, we set a fixed random seed and created an inner 4-fold cross-validation loop within the training data to select the optimal λ for each outer fold.

The model was fitted on the training genotype using each candidate λ value, and performance on the projected PC1 in the test fold was assessed via multiple metrics: mean squared error (MSE), R^2, adjusted R^2, and Pearson correlation between the predicted and observed gene expression values. The optimal λ was chosen as the one minimizing the average MSE across inner folds.

Before training the final model using the full LIBD genotype data and the global PC1, we checked for sign consistency between the gene's expression and the PC1. Since the sign of a principal component is arbitrary, we flipped the PC1 if its Pearson correlation with the target gene's expression was negative. We also computed the Pearson correlation between observed gene-level expression and predictions from the cross-validated model fitted on the training genotype, as an additional measure of consistency and directionality.

B.5 Supplementary Figures

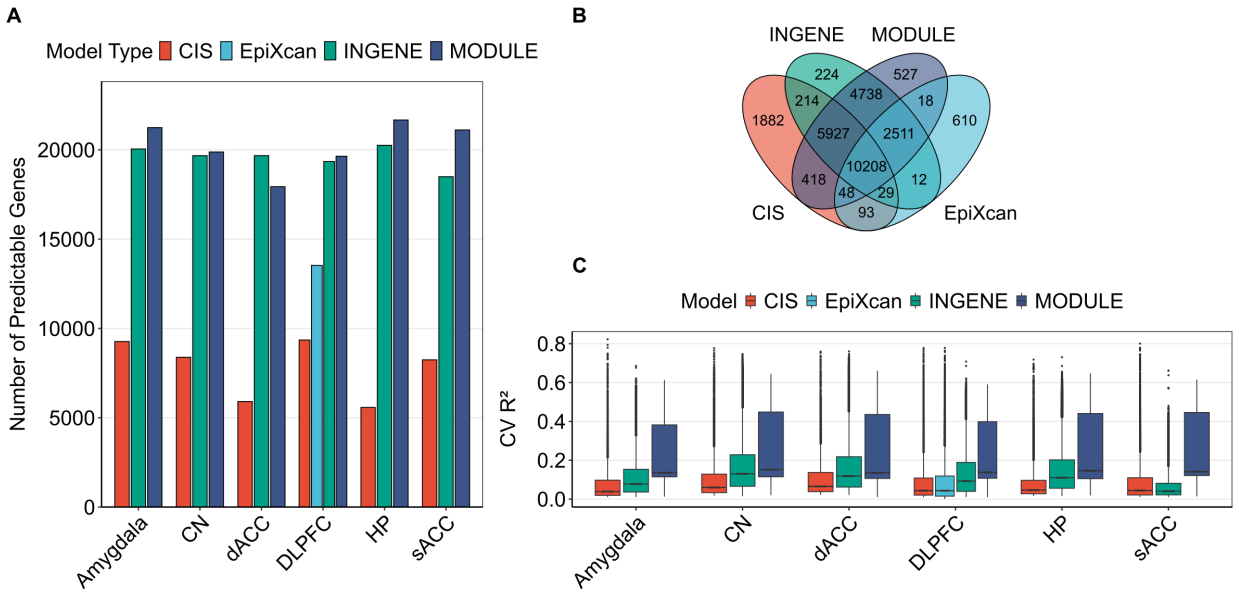

Figure S1: **Comparison of CIS, EpiXcan, INGENE, and MODULE model training performance**. (A) Barplot illustrates the number of genes meeting the threshold (cross-validated adjusted $R^2 \geq 0.01$) for CIS (red), EpiXcan (light blue), INGENE (green), and MODULE (blue) across brain regions. (B) Venn diagram showing the overlap of total predicted genes among models. (C) Distribution of cross-validated R^2 values for CIS, EpiXcan, INGENE, and MODULE across brain regions. Boxplots show the median (central line), interquartile range (IQR, box), and whiskers extending to $1.5 \times$ IQR; outliers are plotted as individual points.

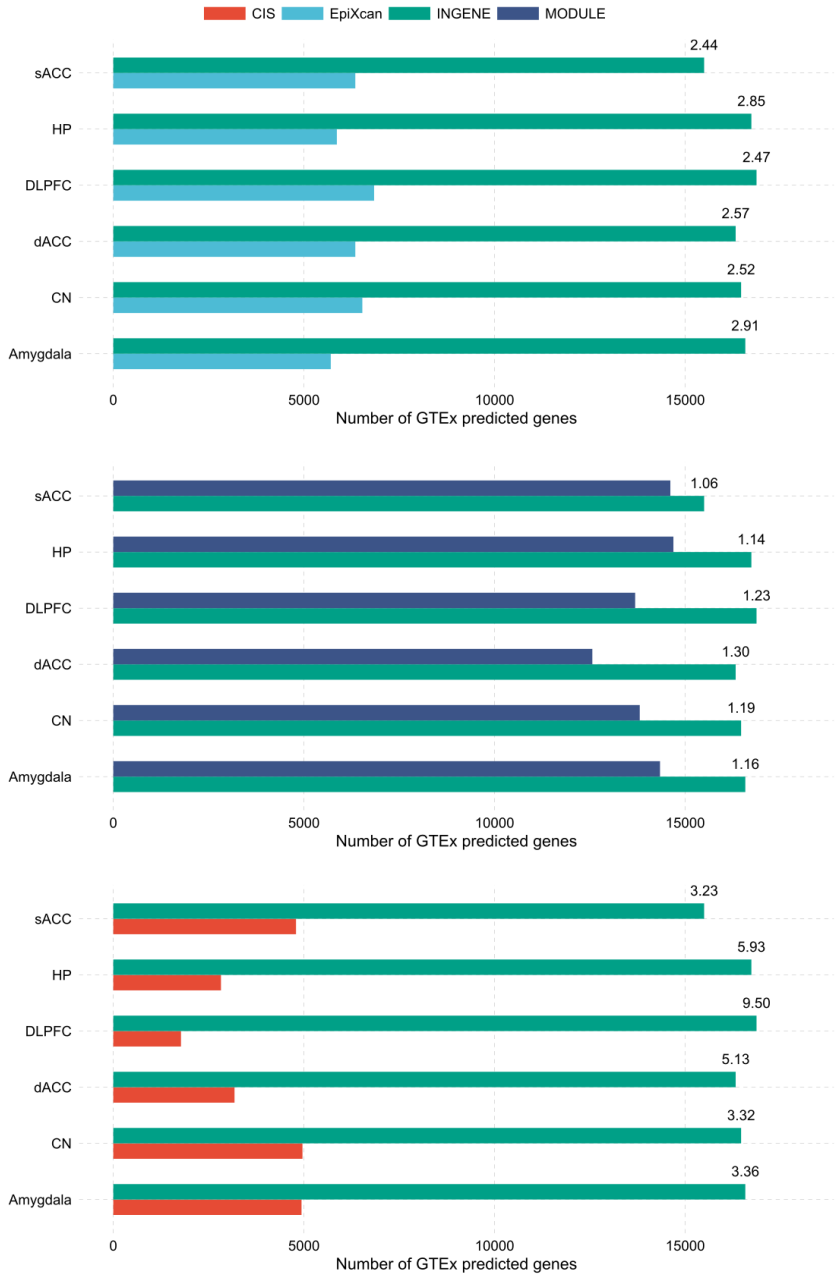

Figure S2: **Predictive models replicate across brain regions in GTEx external dataset and predict differ-ent genes at different performance**. Barplots show the number of predicted genes (x axis) in the GTEx dataset by CIS (red), EpiXcan (light blue), INGENE (green) and MODULE (blue) models. The number on the right indicates the ratio of INGENE gene counts divided by EpiXcan counts (top), MODULE counts (middle) and CIS counts (bottom).

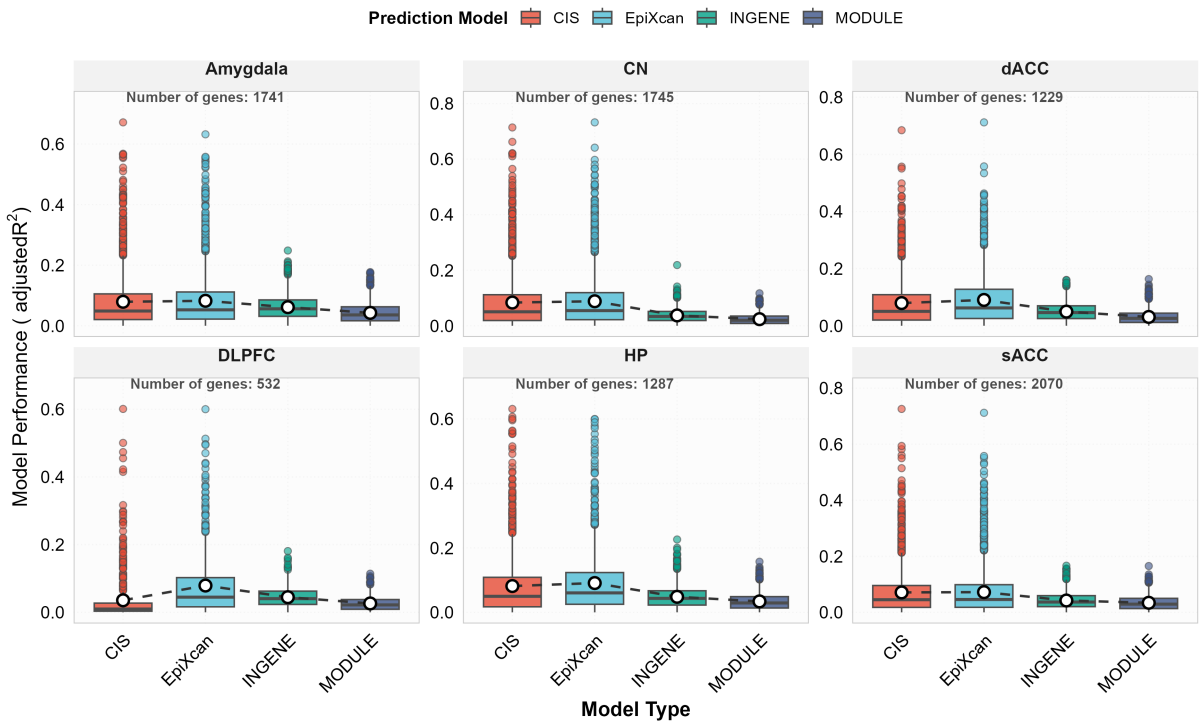

Figure S3: **Models predict common genes at different performance across brain regions in GTEx external dataset**. Box plots of adjusted R^2 values (y-axis) in predicting gene-level expression in GTEx using CIS (red), EpiXcan (light blue), INGENE (green) and MODULE (blue) for commonly "n" predicted genes within brain regions. The median is represented by the central line, with the interquartile range (IQR) as the box. Whiskers extend to 1.5×IQR, and outliers are plotted as individual points.

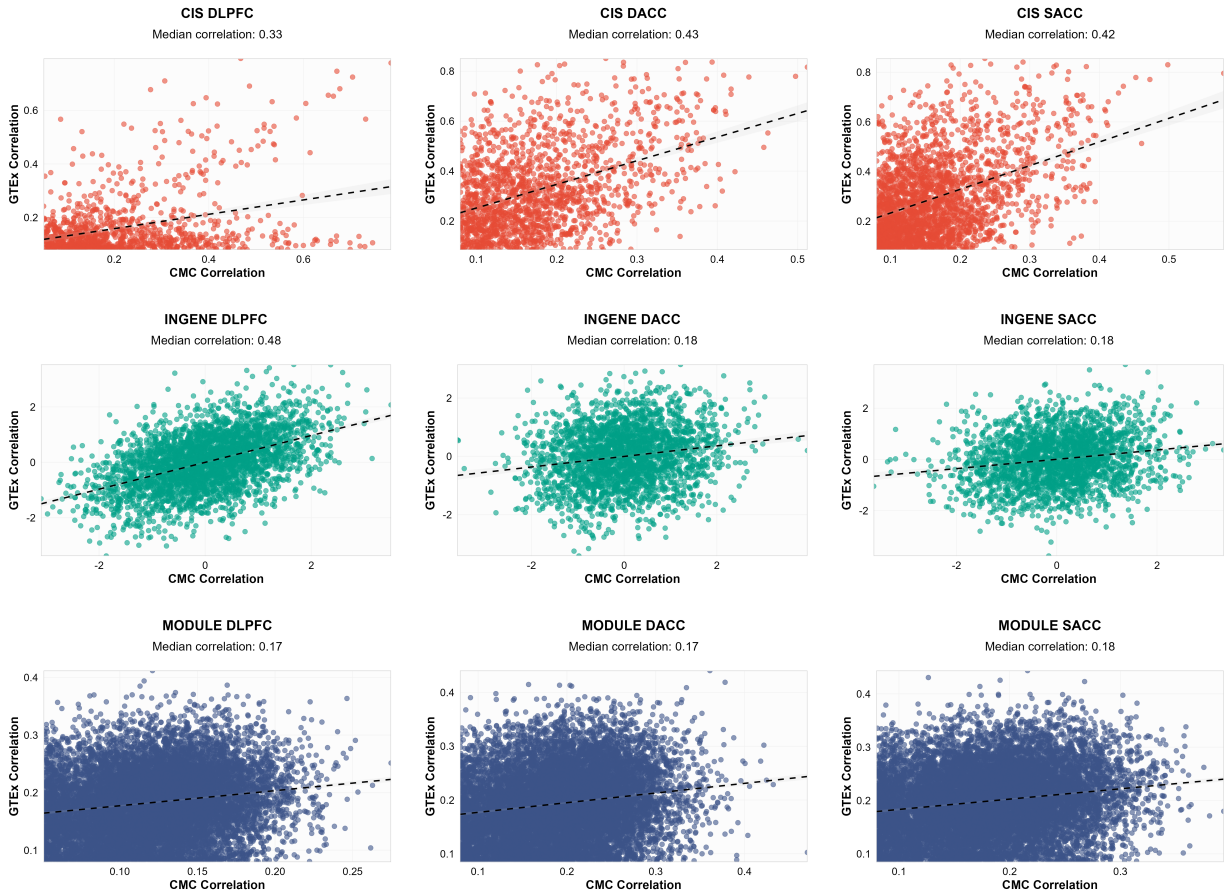

Figure S4: **The correlation of CIS (red), INGENE (green) and MODULE (blue) predictions be-tween CMC and GTEx in DLPFC, dACC and sACC**. The x-axis shows correlation coefficients between observed and predicted expressions in the CMC testing dataset, while the y-axis represents correlation coefficients between observed and predicted expressions in GTEx.

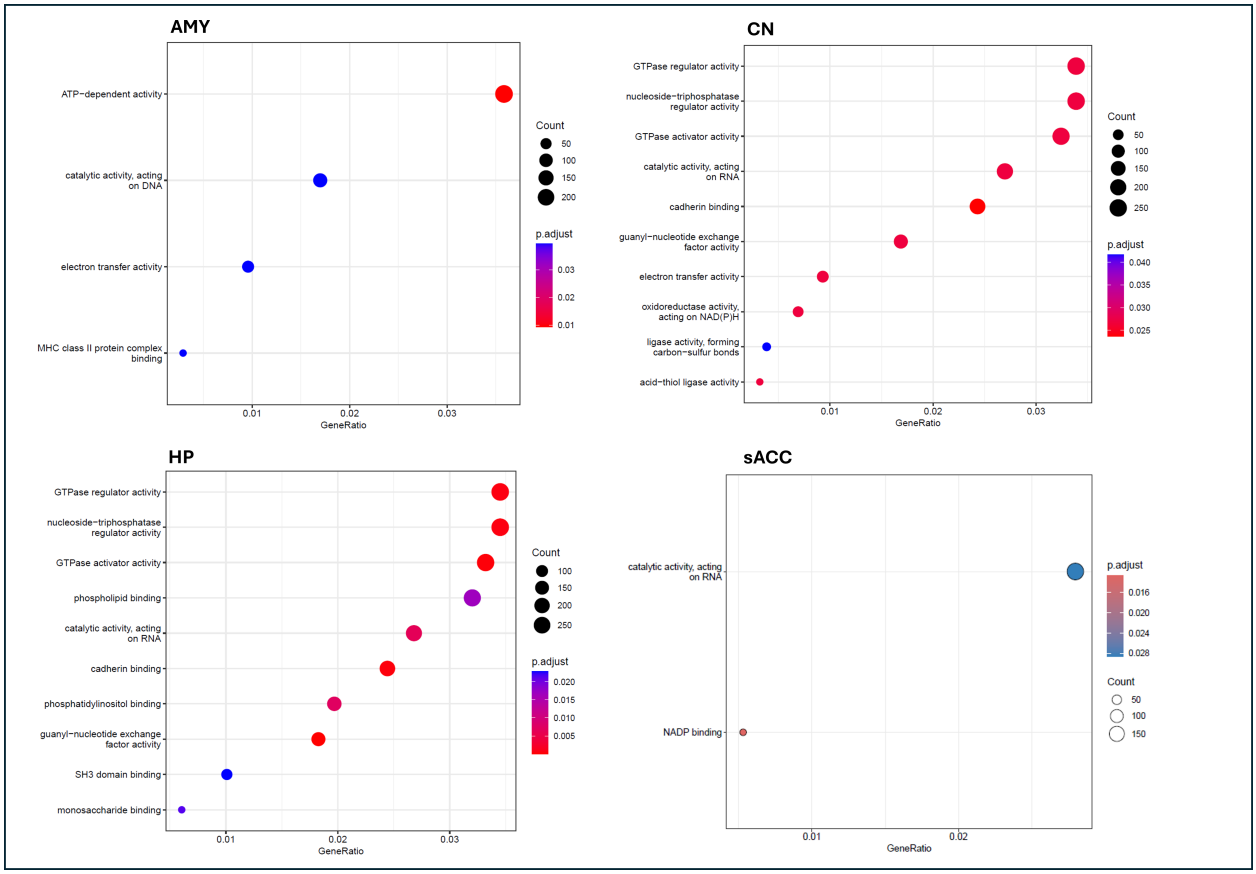

Figure S5: **GO Enrichment Analysis on GTEx eGenes.** The x-axis shows the gene ratio for each molecular function category (y-axis). P-adjusted val-ues refer to BH correction. Abbreviations: AMY: amygdala; CN: caudate nucleus bulk tissue data; HP: hippocampus bulk tissue data; sACC: subgenual anterior cingulate cortex bulk tissue data.

B.6 Supplementary Tables

Table ST1: **Percentage of MODULE-predicted genes regulated by cis-eQTLs of co-expression partners across brain regions.**

Region	% of Genes Regulated by Partner cis-SNPs
Amygdala	20.0%
CN	18.4%
dACC	16.3%
DLPFC	24.0%
HP	15.0%
sACC	23.0%

Note: Percentages indicate the proportion of trans-predicted genes for which at least one co-expression partner's cis-eQTL overlaps as a trans-eQTL. Abbreviations: CN = caudate nucleus; dACC = dorsal anterior cingulate cortex; DLPFC = dorsolateral prefrontal cortex; HP = hippocampus; sACC = subgenual anterior cingulate cortex.

Table ST2: **Performance of PrediXcan-family models trained on LIBD DLPFC samples.**

Model	Mean Adj. R^2	SD Adj. R^2	Number of Genes
EpiXcan	0.043	0.081	8,137
CMC	0.051	0.090	5,832
MASHR	0.034	0.080	4,158
PrediXcan	0.072	0.100	2,828

Note: Values correspond to adjusted R^2 for genes with non-zero predictive performance. Abbreviations: SD = standard deviation.

Table ST3: **Summary of co-expression networks, their sources, and methods of network construction**

Network Name	**Study**	**Method for Network Construction**
Hartl2021	Hartl et al. (2021)	RNA-seq across 12 brain regions; modules defined via WGCNA and categorized as region-specific or conserved.
Pergola2019	Pergola et al. (2019a)	WGCNA on DLPFC postmortem RNA-seq; replicated in multiple datasets; module validated for treatment response relevance.
Pergola2023	Pergola et al. (2023b)	WGCNA-derived consensus networks across brain regions and ages; focused on modules enriched for SCZ risk genes.
Radulescu2020	Radulescu et al. (2020)	WGCNA applied to DLPFC RNA-seq data (SCZ vs. control); prioritized modules enriched for PRS and GWAS loci.
Gandal2018	Gandal et al. (2018a)	RNA-seq across ASD, SCZ, and BPD; co-expression modules stratified by cell type and disorder; WGCNA across combined datasets.
Fromer2016_control	Fromer et al. (2016)	RNA-seq from DLPFC samples (SCZ and controls); case-control analysis with co-expression patterns derived from CommonMind data.
Werling2020	Werling et al. (2020)	WGCNA applied to RNA-seq from 176 DLPFC samples across development; identified 19 consensus modules analyzed for developmental trajectories, cell-type specificity, and GWAS enrichment.
Li2018	Li et al. (2018b)	WGCNA performed on RNA-seq from neurodevelopmental brain tissue; modules used for spatiotemporal expression modeling.
Walker2019	Walker et al. (2019)	WGCNA applied to fetal brain RNA-seq; modules integrated with eQTL and splicing QTL data to inform GWAS loci interpretation.

Appendix C

Supplementary Information (Study 2)

C.1 Supplementary Figures

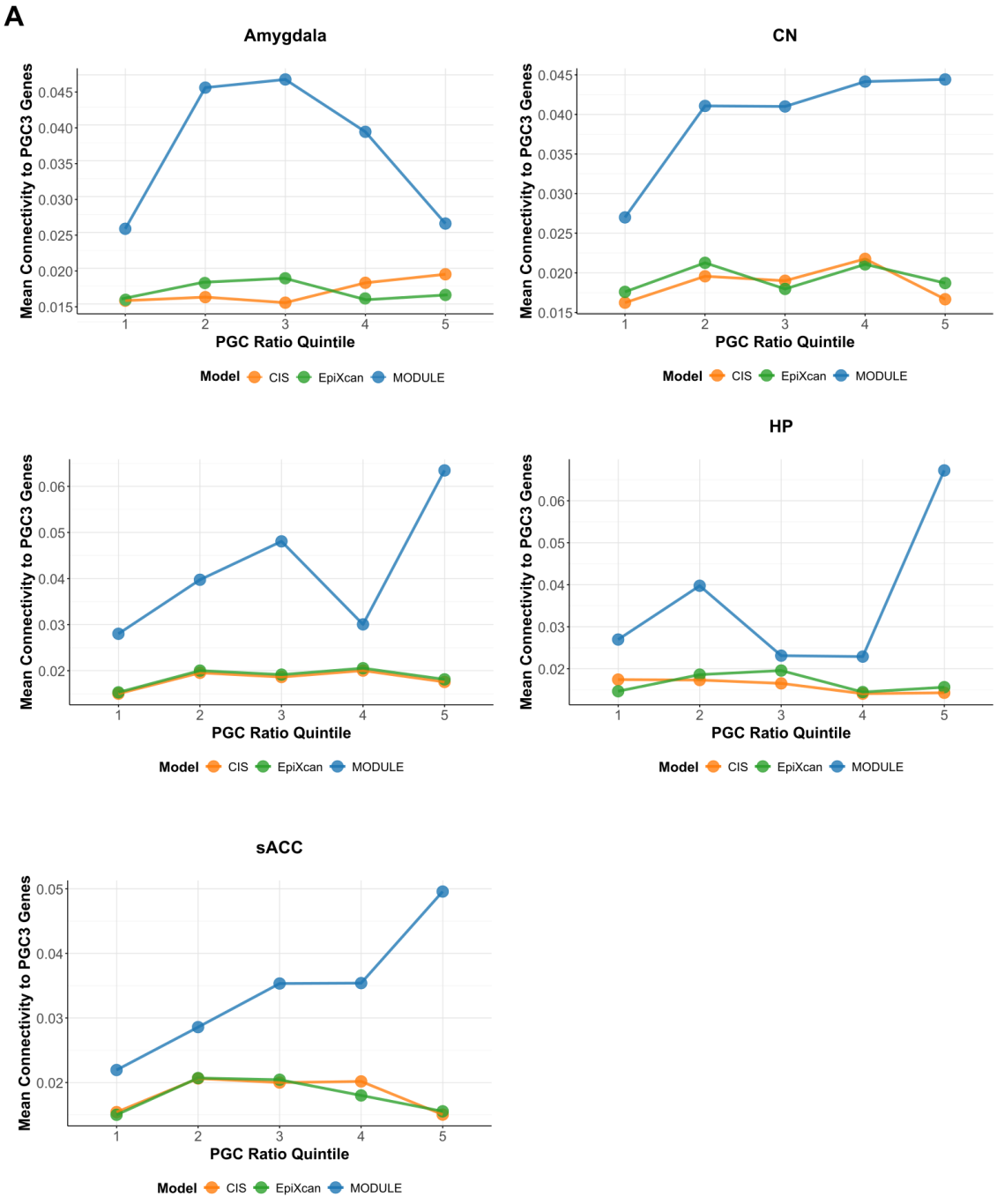

Figure S6: **Connectivity between predicted gene sets and PGC3 risk genes across PGC-weight quintiles.** MODULE models (blue) show robust increases in connectivity across several regions, while CIS (orange) and EpiXcan (green) show little or no trend.

Model	Region	Linear p	Spearman p	Perm Slope p	Perm Spearman p
MODULE	dlpfc	0.2182	0.2333	0.000	0.231
MODULE	amygdala	0.9059	0.9500	0.184	0.949
MODULE	caudate	0.0794	0.0833	0.000	0.083
MODULE	hippo	0.3513	0.9500	0.000	0.941
MODULE	sACC	0.0107	0.0167	0.000	0.018
CIS	dlpfc	0.4563	0.6833	0.226	0.685
CIS	amygdala	0.0637	0.2333	0.037	0.210
CIS	caudate	0.7293	0.6833	0.598	0.654
CIS	hippo	0.0255	0.0833	0.050	0.077
CIS	sACC	0.9112	0.6833	0.779	0.700
EpiXcan	dlpfc	0.4245	0.6833	0.169	0.680
EpiXcan	caudate	0.7664	0.6833	0.738	0.708
EpiXcan	hippo	0.8102	0.9500	0.674	0.942
EpiXcan	sACC	0.8771	1.0000	0.675	1.000

Figure S7: **Summary table of nominal and permutation test statistics highlights that enrichment of SCZ risk gene connectivity is specific to co-expression–based MODULE models.**

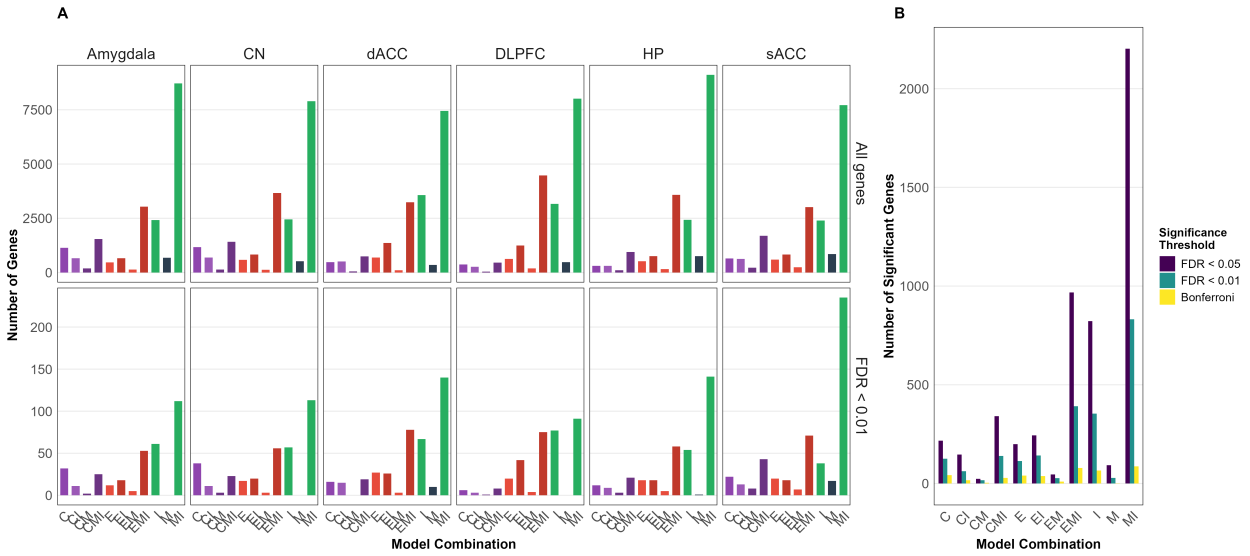

Figure S8: **Distribution of predicted genes by brain region and model across PGC3 cohorts. A)** Barplot showing the number of predicted genes, pooling predictions from all models and PGC3 cohorts. **B)** Number of genes surviving different thresholds of significance (FDR 0.05, FDR 0.01, and Bonferroni 0.05) across models. Abbreviations: C = CIS; E = EpiXcan; M = MODULE; I = INGENE; CN = caudate nucleus; dACC = dorsal anterior cingulate cortex; DLPFC = dorsolateral prefrontal cortex; HP = hippocampus; sACC = subgenual anterior cingulate cortex.

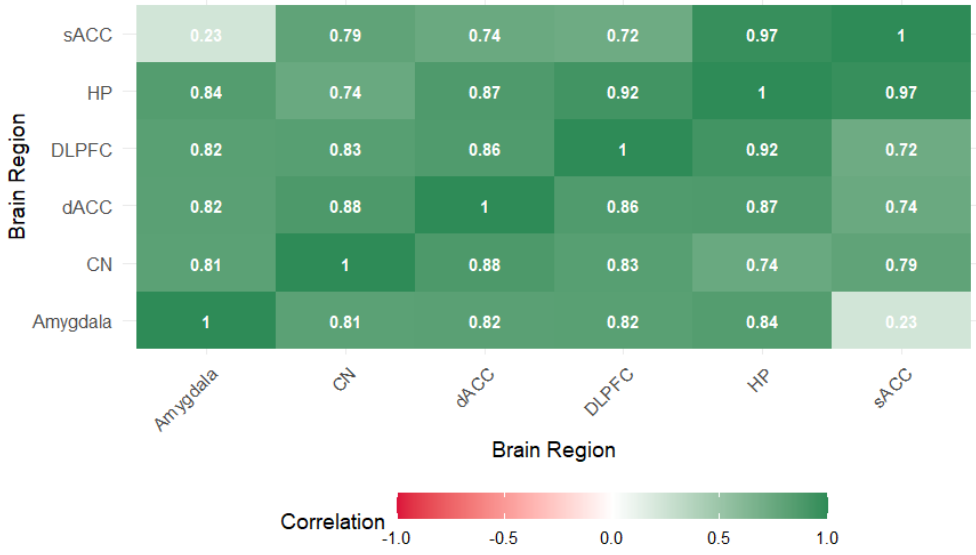

Figure S9: **Cross-region directionality of TWAS effects.** Pairwise correlations of gene-level β across brain regions for the multi-region significant genes (FDR < 0.01). Most pairs show high concordance (dark green), with pockets of lower concordance (lighter green), indicating region-specific effects.

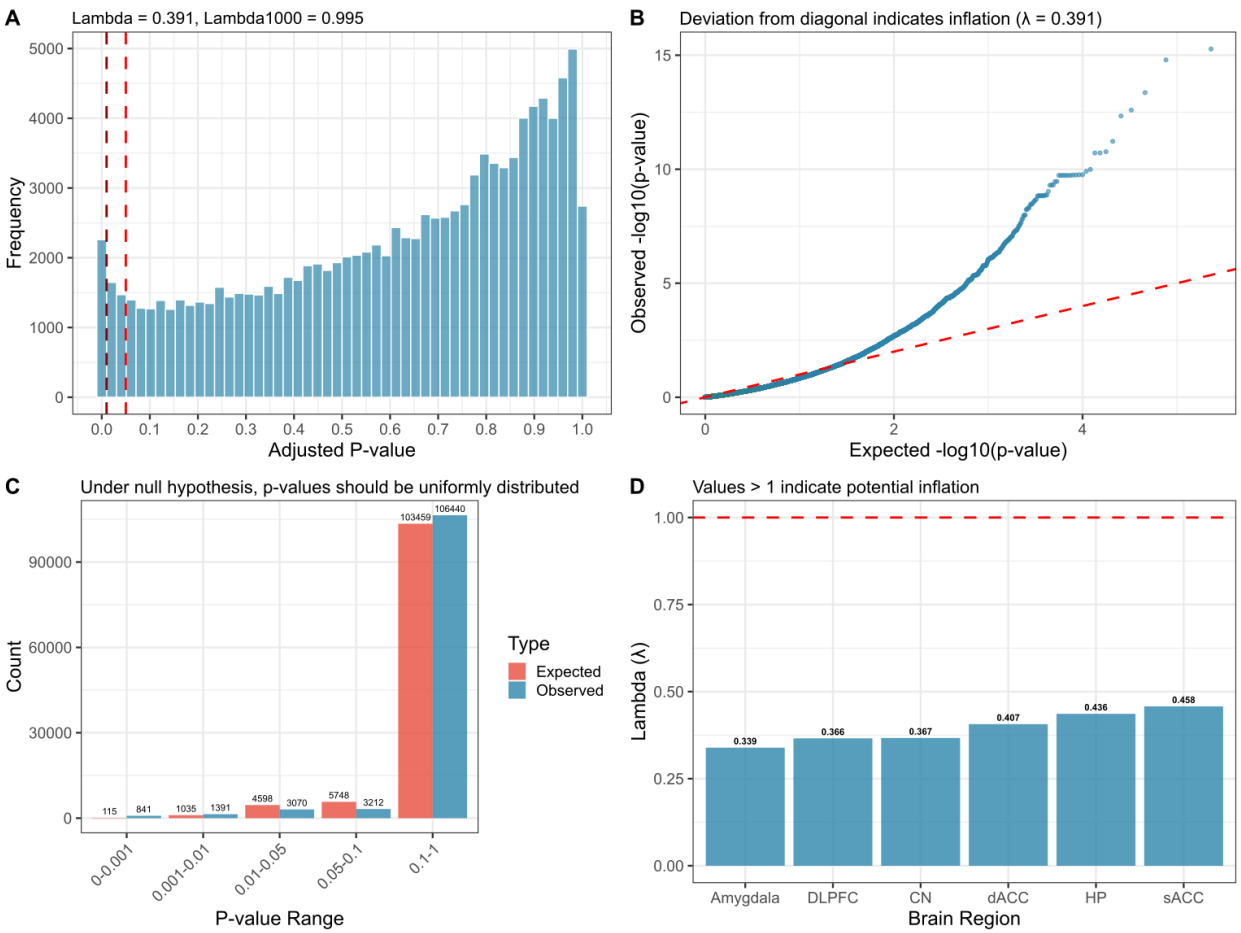

Figure S10: **Assessment of statistical inflation in coTWAS results. A)** Histogram of adjusted p-values across all gene–region tests, showing enrichment of small values relative to the null. **B)** Quantile–quantile (Q–Q) plot illustrating deviation from the expected diagonal under the uniform null distribution. **C)** Observed versus expected counts across p-value ranges, confirming an excess of significant associations. **D)** Genomic inflation factor (λ) estimated for each brain region, with values <1 indicating no evidence of systematic inflation.

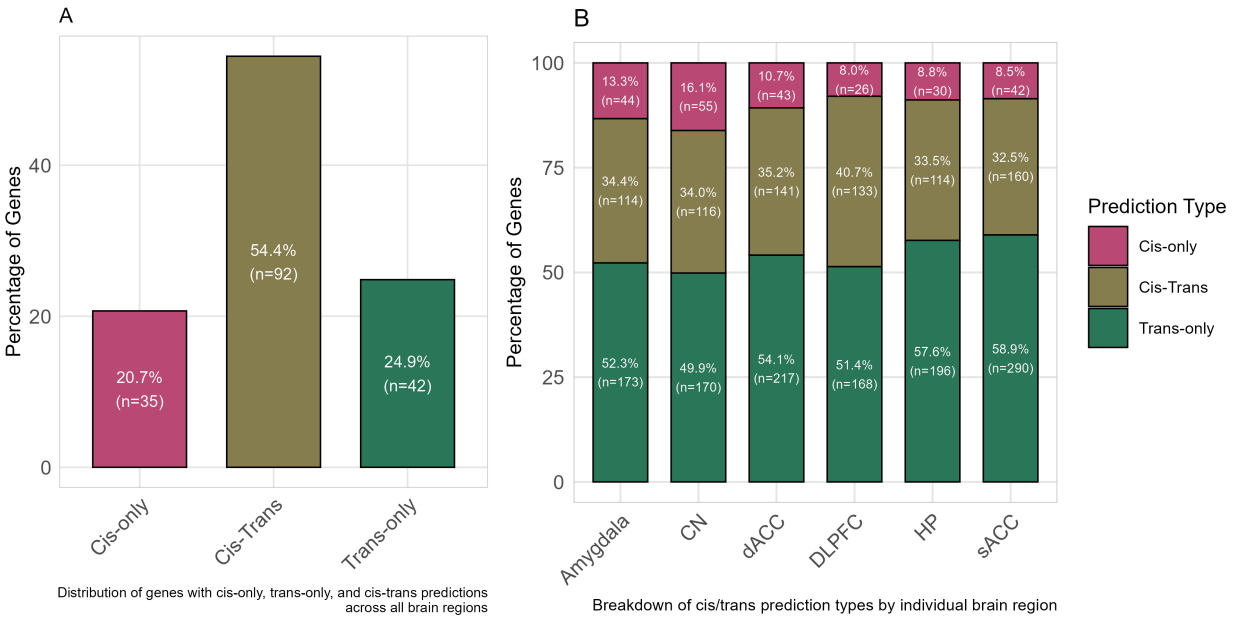

Figure S11: **Distribution of FDR 0.01 significant genes with *cis*-only, *trans*-only, and *cis-trans* predictions. A)** Barplot shows the percentage of predicted genes with a consistent prediction type in more than 2 regions. **B)** Percentage of genes within regions and across prediction types. Abbreviations: CN: caudate nucleus data; dACC: dorsal anterior cingulate cortex; DLPFC: dorsolateral prefrontal cortex; HP: hippocampus; sACC: subgenual anterior cingulate cortex.

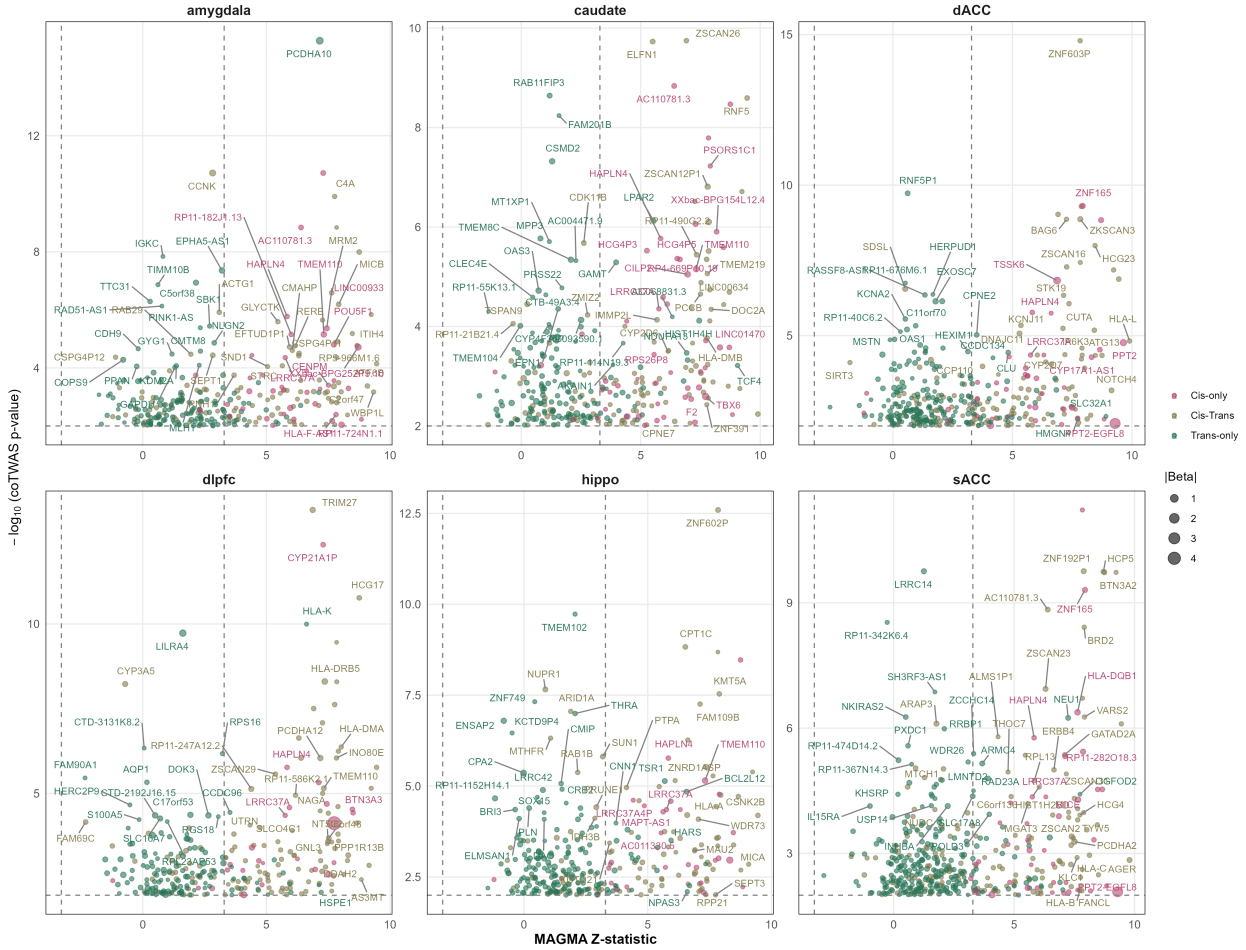

Figure S12: **Integration of MAGMA and coTWAS results.** Scatterplots showing gene-level association statistics across brain regions. The x-axis represents MAGMA Z-statistics, and the y-axis shows $-\log_{10}(p\text{-value})$ from coTWAS results. Each point corresponds to a gene, colored by prediction class: red for cis-only, green for trans-only, and gold for both cis and trans. Point size is proportional to the absolute value of the coTWAS logistic regression β coefficient, reflecting effect size magnitude.

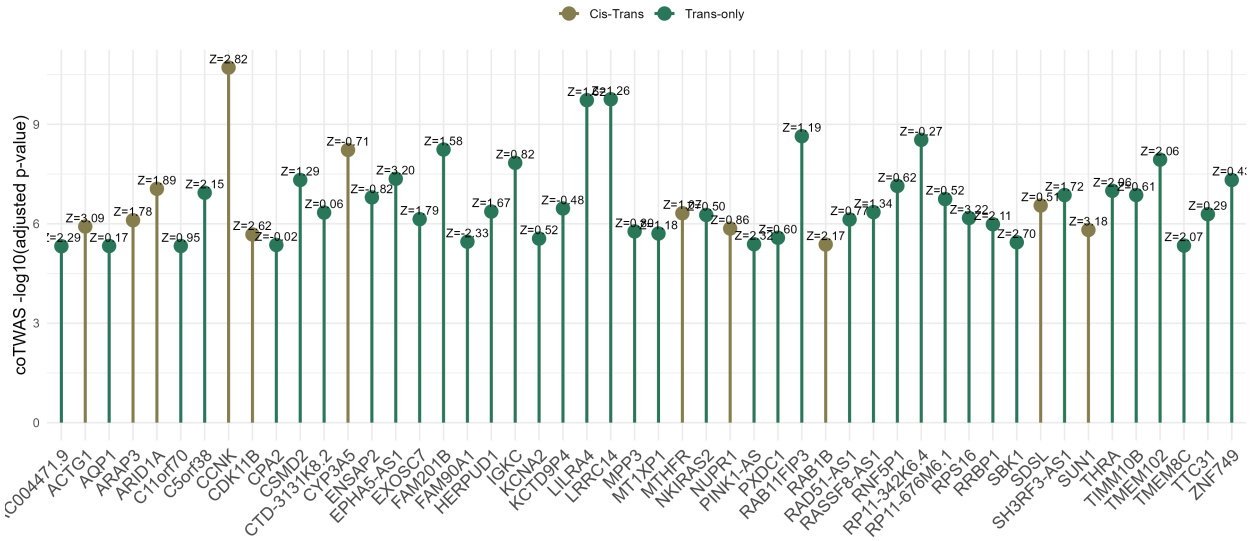

Figure S13: **Top 50 genes with strong coTWAS evidence but weak MAGMA association.** Genes are ranked by coTWAS significance, with each point representing a gene that exhibits a strong coTWAS association ($\log_{10}$ adjusted p-value > 5) but a relatively weak MAGMA Z-score ($|Z| < 4$). MAGMA Z-statistics are displayed above each point.

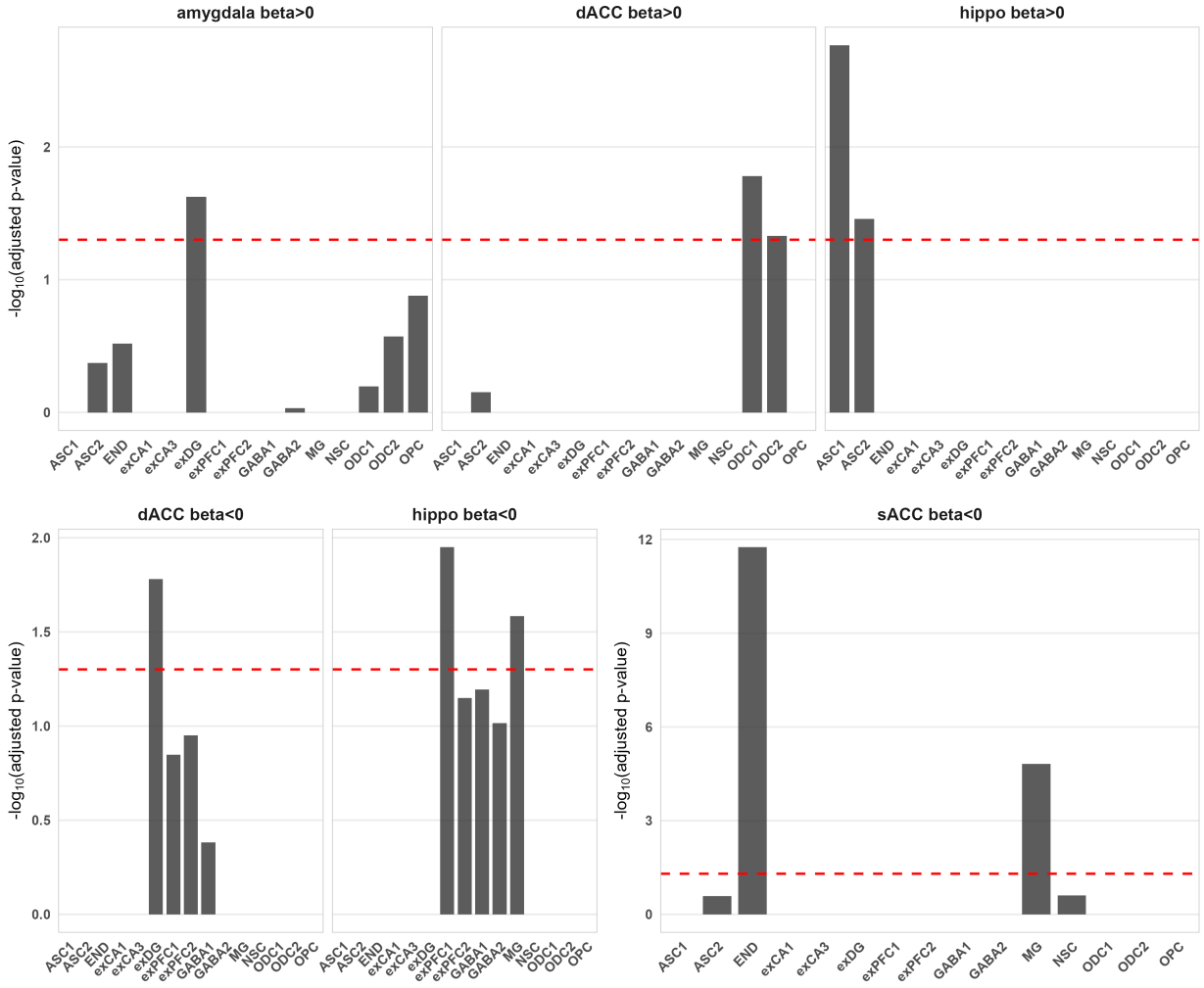

Figure S14: **Cell-type specificity of coTWAS-significant genes based on human single-cell transcriptomic data.** Enrichment p-values were obtained using the mean-rank Gene Set Test from the `limma` R package. The y-axis displays FDR-adjusted p-values, corrected for multiple comparisons across genes and cell types. Red dashed lines indicate the FDR significance threshold ($\alpha = 0.05$). The top panel distinguishes upregulated ($\beta > 0$) and downregulated ($\beta < 0$) genes in SCZ patients based on coTWAS logistic regression results.

C.2 Supplementary Tables

Table ST4: **Number of CIS, EpiXcan, and MODULE SNPs overlapping the 900,090 SNPs from the PGC3 SCZ summary statistics ($p < 0.05$) across brain regions for GTEx-validated genes**. Abbreviations: CN = caudate nucleus; dACC = dorsal anterior cingulate cortex; DLPFC = dorsolateral prefrontal cortex; HP = hippocampus; sACC = subgenual anterior cingulate cortex; OR = odds ratios; abs = absolute value.

Region	**Model**	**N° SNPs in Model**	**Overlapping SNPs**	**Pearson's Coeff.** log(OR), |weights|
Amygdala	CIS	62,664	5,919	0.14
Amygdala	EpiXcan	100,282	12,824	0.14
Amygdala	MODULE	75,455	805	0.40
CN	CIS	64,316	5,606	0.15
CN	EpiXcan	111,265	14,043	0.15
CN	MODULE	75,781	8,091	0.34
dACC	CIS	48,285	3,258	0.15
dACC	EpiXcan	132,312	16,413	0.18
dACC	MODULE	71,526	6,681	0.28
DLPFC	CIS	23,898	2,065	0.15
DLPFC	EpiXcan	119,388	15,006	0.15
DLPFC	MODULE	47,262	5,297	0.41
HP	CIS	3,708	2,814	0.14
HP	EpiXcan	10,039	12,667	0.15
HP	MODULE	72,991	7,588	0.33
sACC	CIS	66,533	6,118	0.15
sACC	EpiXcan	139,932	16,413	0.14
sACC	MODULE	57,414	6,556	0.42

Table ST5: **Distribution of PGC3 cohorts by site, sex at birth, and diagnosis.** PGC Site refers to the unique identifier for each participating cohort. *N* indicates the total number of participants at each site. *Sex at birth (Males/Females)* reports the number of individuals assigned male or female at birth. *Cases/Controls* represents the number of individuals diagnosed with schizophrenia and healthy controls, respectively. All individuals included in this table are of European ancestry.

PGC Site	N	Sex at Birth (M/F)	Cases/Controls
bep1b	849	352/497	293/556
braz2	433	238/195	102/331
butr	1298	656/642	649/649

Continued on next page

Table ST5 – continued from previous page

PGC Site	N	Sex at Birth (M/F)	Cases/Controls
celso	3499	2139/1360	1993/1506
clz2a	11970	6532/5434	5063/6907
cogs1	870	517/353	401/469
du2aa	555	333/222	321/234
enric	1274	735/539	700/574
eu5me	790	402/387	613/177
eusp2	792	512/280	308/484
eutu2	1083	570/513	393/690
gap1a	281	166/115	120/161
geba1	1037	491/546	358/679
gpc2a	3979	2264/1715	1923/2056
gro2a	376	238/138	210/166
grtr	290	238/52	145/145
lemu	1032	690/342	516/516
mcqul	2506	1302/1204	1222/1284
mosc2	841	442/399	408/433
price	1568	887/681	841/727
rive1	1521	836/685	319/1202
rouin	389	220/169	204/185
sb2aa	469	296/173	236/233
serri	454	241/213	216/238
to10c	5961	3030/2931	961/5000
uktr	140	110/30	70/70

Continued on next page

Table ST5 – continued from previous page

PGC Site	N	Sex at Birth (M/F)	Cases/Controls
xaber	1325	916/409	642/683
xajsz	2487	1744/743	893/1594
xasrb	755	445/310	463/292
xboco	3936	1923/2013	1778/2158
xbuls	801	381/420	193/608
xcati	614	466/148	397/217
xcaws	668	391/277	387/281
xcgs3	3580	1838/1742	474/3106
xcims	126	95/31	65/61
xclm2	7425	4476/2949	3345/4080
xclo3	4023	2476/1547	2053/1970
xcou3	1200	654/546	521/679
xdenm	930	542/388	474/456
xdubl	1100	423/677	260/840
xedin	598	370/228	319/279
xegcu	1387	372/1015	236/1151
xersw	591	369/222	271/320
xgras	2234	1437/797	1065/1169
xirwt	2231	1374/857	1235/996
xlacw	394	344/50	153/241
xlie2	401	219/182	132/269
xlie5	878	550/328	489/389
xmsaf	458	280/178	321/137

Continued on next page

Table ST5 – continued from previous page

PGC Site	N	Sex at Birth (M/F)	Cases/Controls
xmunc	731	411/320	421/310
xpewb	2158	1167/991	456/1702
xpews	381	222/159	148/233
xport	560	293/267	345/215
xs234	4232	2325/1907	1967/2265
xswe1	421	219/202	213/208
xswe5	4348	2410/1938	1770/2578
xswe6	2132	1148/984	978/1154
xtop8	746	401/345	369/377
xucla	1117	693/424	580/537
xuclo	977	549/428	492/485
xume2	2032	974/1056	544/1488
xzhh1	379	219/160	190/189

Appendix D

Supplementary Information (Study 3)

D.1 Genotype Data Preprocessing

DNA samples were genotyped at UNIPI using the Illumina Infinium Omni2.5Exome-8 v1.5 arrays and processed on a 550 NextSeq Illumina platform.

Post-imputation quality control was performed at UNIBA following the same protocol described in Study 1 (Chapter 3) and Study 2 (Chapter 4). Briefly, post-imputation quality control was conducted uniformly with PLINK toolkit version 1.07 (Purcell et al., 2007). SNPs were excluded based on minor allele frequency (MAF) ≤ 0.01, Hardy-Weinberg equilibrium p-value $< 10^{-6}$, or $> 5\%$ missingness. Individuals were excluded for $> 2\%$ missing genotypes or relatedness $\hat{\pi} > 0.125$. Population structure was inferred using principal components aligned to HapMap3 (Altshuler et al., 2010). Among the 707 genotyped individuals that passed QC steps, the ancestry composition included:

- 346 of European ancestry
- 269 of Central American ancestry
- 92 of mixed ancestry

The final number of genotypes after processing was 4,894,072.

D.2 Gene Expression Prediction

To generate predictors, we applied CIS, EpiXcan, INGENE and MODULE to the genotype data of each participant, yielding imputed expression levels for approximately 20,000 genes per region (Table 7). As described in Chapter 3 (section 3.4.2), model validity was assessed using cross-cohort replication in the GTEx brain dataset. Only genes that met minimum predictive performance criteria in GTEx—namely, adjusted $R^2 > 0$ and Pearson correlation $r > 0$—were retained for downstream use.

For each gene with multiple available predictors across models or regions, we computed a combined expression estimate using a linear weighting strategy introduced in Study 1 (refer to section 3.3.5). This integration step enhanced expression accuracy while preserving biological specificity.

To control for demographic confounding, gene expression predictions were residualized using linear models that included age and ethnicity (Hispanic vs. non-Hispanic) as covariates. This residualization was performed gene-wise prior to any modeling.

SVM, RF, XgBoost Model Tuning Parameters and Their Rationale

To optimize predictive performance while minimizing overfitting in a high-dimensional, low-sample setting, model hyperparameters were carefully selected for each algorithm. The choices reflect a bias–variance trade-off appropriate for exploratory behavioural prediction using transcriptomic data.

Random Forest

- `mtry`: The number of features randomly selected at each split. Smaller values (e.g., $\sqrt{p}$, $p/3$) were used to reduce model complexity and encourage diversity among trees,

thereby lowering variance and the risk of overfitting.

- `min.node.size`: The minimum number of observations allowed in a terminal node. Lower values (1, 5, 10) permit deeper trees and finer splits, which can increase model flexibility but also raise the risk of overfitting. These settings were included to explore the trade-off between bias and variance in small-sample contexts.

XGBoost

- `eta` (learning rate): Controls the contribution of each tree to the final model. Small values ({0.01, 0.05, 0.1}) ensure conservative, incremental updates, which are particularly beneficial in high-dimensional settings where overfitting is a concern.

- `max.depth`: Limits the maximum depth of individual trees. Shallow trees ({2, 3}) were used to prevent overly complex decision boundaries and to promote generalizability.

Support Vector Machine (SVM)

- `C`: Regularization parameter that controls the trade-off between achieving a low training error and maintaining a large margin. Smaller values promote simpler models that generalize better, especially in sparse feature spaces.

- `epsilon`: Defines the margin of tolerance around the predicted function in regression tasks (e.g., SVR). Tuning this parameter helps control model sensitivity and avoids overfitting to noise in behavioural data.

These parameter ranges were selected based on prior research applying machine learning to high-dimensional biomedical data and adapted to the specific characteristics of the present study's sample and feature space.

D.3 Supplementary Figures

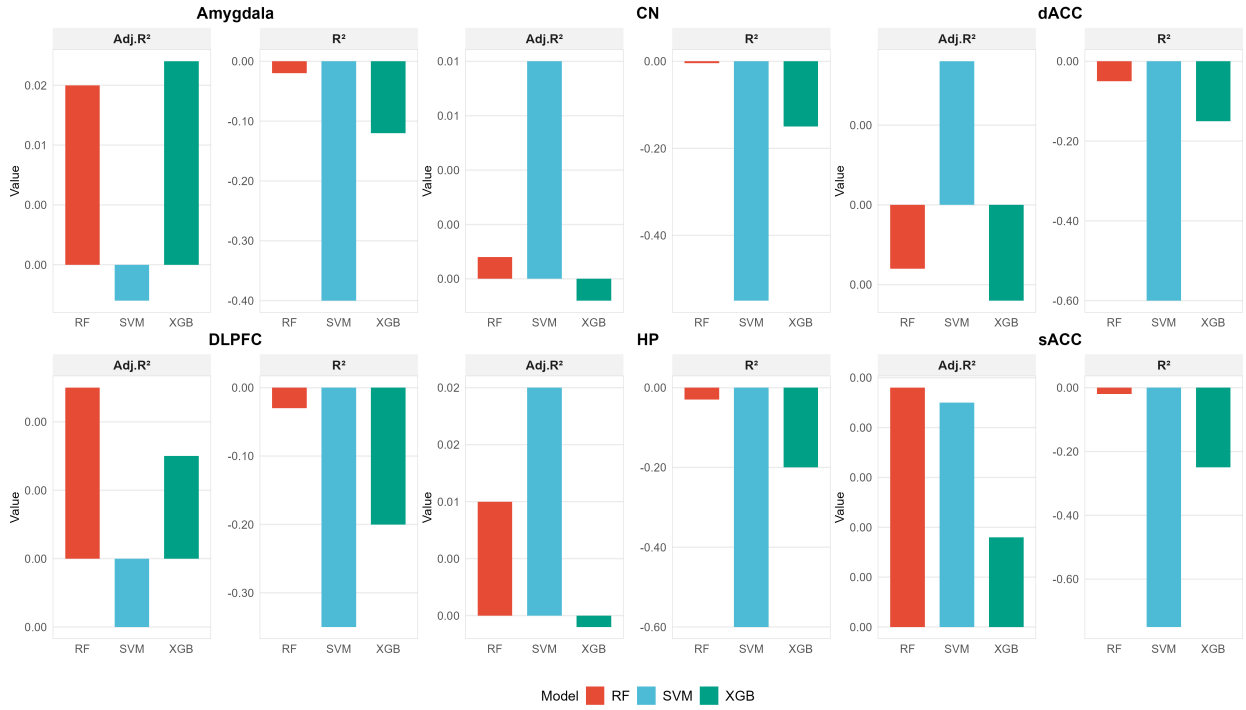

Figure S15: **Model Performance Across Individual Brain Regions.** Each panel displays adjusted R^2, r^2, and raw R^2 values for RF, SVM, and XGBoost across six brain regions: amygdala, caudate, dACC, DLPFC, hippocampus, and sACC. SVM shows marginally higher predictive power in the hippocampus and caudate. Across all regions, however, test set R^2 remains negative, and no model achieves meaningful out-of-sample accuracy.

Bibliography

Akbarian S, Liu C, Knowles JA, Vaccarino FM, Farnham PJ, Crawford GE, Jaffe AE, Pinto D, Dracheva S, Geschwind DH, Mill J, Nairn AC, Abyzov A, Pochareddy S, Prabhakar S, Weissman S, Sullivan PF, State MW, Weng Z, Peters MA, White KP, Gerstein MB, Amiri A, Armoskus C, Ashley-Koch AE, Bae T, Beckel-Mitchener A, Berman BP, Coetzee GA, Coppola G, Francoeur N, Fromer M, Gao R, Grennan K, Herstein J, Kavanagh DH, Ivanov NA, Jiang Y, Kitchen RR, Kozlenkov A, Kundakovic M, Li M, Li Z, Liu S, Mangravite LM, Mattei E, Markenscoff-Papadimitriou E, Navarro FC, North N, Omberg L, Panchision D, Parikshak N, Poschmann J, Price AJ, Purcaro M, Reddy TE, Roussos P, Schreiner S, Scuderi S, Sebra R, Shibata M, Shieh AW, Skarica M, Sun W, Swarup V, Thomas A, Tsuji J, van Bakel H, Wang D, Wang Y, Wang K, Werling DM, Willsey AJ, Witt H, Won H, Wong CC, Wray GA, Wu EY, Xu X, Yao L, Senthil G, Lehner T, Sklar P, Sestan N (2015) The psychencode project. Nat Neurosci 18(12):1707–12

Allen NC, Bagade S, McQueen MB, Ioannidis JP, Kavvoura FK, Khoury MJ, Tanzi RE, Bertram L (2008) Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the szgene database. Nat Genet 40(7):827–34

Altshuler D, Donnelly P, The International HapMap C (2005) A haplotype map of the human genome. Nature 437(7063):1299–1320

Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L, Dermitzakis E, Bonnen PE, Altshuler DM, Gibbs RA,

de Bakker PI, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, Inouye M, Jia X, Palotie A, Parkin M, Whittaker P, Yu F, Chang K, Hawes A, Lewis LR, Ren Y, Wheeler D, Gibbs RA, Muzny DM, Barnes C, Darvishi K, Hurles M, Korn JM, Kristiansson K, Lee C, McCarrol SA, Nemesh J, Dermitzakis E, Keinan A, Montgomery SB, Pollack S, Price AL, Soranzo N, Bonnen PE, Gibbs RA, Gonzaga-Jauregui C, Keinan A, Price AL, Yu F, Anttila V, Brodeur W, Daly MJ, Leslie S, McVean G, Moutsianas L, Nguyen H, Schaffner SF, Zhang Q, Ghori MJ, McGinnis R, McLaren W, Pollack S, Price AL, Schaffner SF, Takeuchi F, Grossman SR, Shlyakhter I, Hostetter EB, Sabeti PC, Adebamowo CA, Foster MW, Gordon DR, Licinio J, Manca MC, Marshall PA, Matsuda I, Ngare D, Wang VO, Reddy D, Rotimi CN, Royal CD, Sharp RR, Zeng C, Brooks LD, McEwen JE (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467(7311):52–8

Arnedo J, Svrakic DM, Del Val C, Romero-Zaliz R, Hernández-Cuervo H, Fanous AH, Pato MT, Pato CN, de Erausquin GA, Cloninger CR, Zwir I (2015) Uncovering the hidden risk architecture of the schizophrenias: confirmation in three independent genome-wide association studies. Am J Psychiatry 172(2):139–53

Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR (2015a) A global reference for human genetic variation. Nature 526(7571):68–74

Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR (2015b) A global reference for human genetic variation. Nature 526(7571):68–74

Avila Cobos F, Alquicira-Hernandez J, Powell JE, Mestdagh P, De Preter K (2020) Benchmarking of cell type deconvolution pipelines for transcriptomics data. Nature Communications 11(1):5650

Bakhshani NM (2014) Impulsivity: A predisposition toward risky behaviors. International Journal of High Risk Behaviors and Addiction 3(2):e20428

Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, Torstenson ES, Shah KP, Garcia T, Edwards TL, Stahl EA, Huckins LM, Nicolae DL, Cox NJ, Im HK (2018) Exploring the phenotypic consequences of tissue specific gene expression variation inferred from gwas summary statistics. Nat Commun 9(1):1825

Batiuk MY, Martirosyan A, Wahis J, de Vin F, Marneffe C, Kusserow C, Koeppen J, Viana JF, Oliveira JF, Voet T, Ponting CP, Belgard TG, Holt MG (2020) Identification of region-specific astrocyte subtypes at single cell resolution. Nat Commun 11(1):1220

Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, Haudenschild CD, Beckman KB, Shi J, Mei R, Urban AE, Montgomery SB, Levinson DF, Koller D (2014) Characterizing the genetic basis of transcriptome diversity through rna-sequencing of 922 individuals. Genome Res 24(1):14–24

Benjamin KJM, Chen Q, Jaffe AE, Stolz JM, Collado-Torres L, Huuki-Myers LA, Burke EE, Arora R, Feltrin AS, Barbosa AR, Radulescu E, Pergola G, Shin JH, Ulrich WS, Deep-Soboslay A, Tao R, Hyde TM, Kleinman JE, Erwin JA, Weinberger DR, Paquola ACM (2022) Analysis of the caudate nucleus transcriptome in individuals with schizophrenia highlights effects of antipsychotics and new risk genes. Nat Neurosci 25(11):1559–1568

Benjamin KJM, Chen Q, Eagles NJ, Huuki-Myers LA, Collado-Torres L, Stolz JM, Pertea G, Shin JH, Paquola ACM, Hyde TM, Kleinman JE, Jaffe AE, Han S, Weinberger DR (2023) Genetic and environmental contributions to ancestry differences in gene expression in the human brain. bioRxiv

Benjamin KJM, Chen Q, Eagles NJ, Huuki-Myers LA, Collado-Torres L, Stolz JM, Pertea G, Shin JH, Paquola ACM, Hyde TM, Kleinman JE, Jaffe AE, Han S, Weinberger DR

(2024) Analysis of gene expression in the postmortem brain of neurotypical black americans reveals contributions of genetic ancestry. Nat Neurosci 27(6):1064–1074

Bernstein HG, Nussbaumer M, Vasilevska V, Dobrowolny H, Nickl-Jockschat T, Guest PC, Steiner J (2024) Glial cell deficits are a key feature of schizophrenia: implications for neuronal circuit maintenance and histological differentiation from classical neurodegeneration. Molecular Psychiatry

Bezdjian S, Baker LA, Tuvblad C (2011) The genetics of impulsivity: evidence for the heritability of delay discounting. Biological Psychiatry 69(3):270–276

Bhattacharya A, Hirbo JB, Zhou D, Zhou W, Zheng J, Kanai M, Pasaniuc B, Gamazon ER, Cox NJ (2022) Best practices for multi-ancestry, meta-analytic transcriptome-wide association studies: Lessons from the global biobank meta-analysis initiative. Cell Genomics 2(10)

Bhattacharya A, Vo DD, Jops C, Kim M, Wen C, Hervoso JL, Pasaniuc B, Gandal MJ (2023) Isoform-level transcriptome-wide association uncovers genetic risk mechanisms for neuropsychiatric disorders in the human brain. Nat Genet 55(12):2117–2128

Birnbaum R, Weinberger DR (2017) Genetic insights into the neurodevelopmental origins of schizophrenia. Nature Reviews Neuroscience 18(12):727–740

Birnbaum R, Weinberger DR (2020) A genetics perspective on the role of the (neuro)immune system in schizophrenia. Schizophr Res 217:105–113

Borcuk C, Parihar M, Sportelli L, Kleinman JE, Shin JH, Hyde TM, Bertolino A, Weinberger DR, Pergola G (2024) Network-wide risk convergence in gene co-expression identifies reproducible genetic hubs of schizophrenia risk. Neuron

Boyle EA, Li YI, Pritchard JK (2017) An expanded view of complex traits: From polygenic to omnigenic. Cell 169(7):1177–1186

Brown AS (2011) The environment and susceptibility to schizophrenia. Prog Neurobiol 93(1):23–58

Bryois J, Skene NG, Hansen TF, Kogelman L, Watson HJ, Liu Z, et al (2022) Cell-type-specific cis-eqtls in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders. Nature Neuroscience 25(8):1104–1112

Buckley PF, Miller BJ (2017) Personalized medicine for schizophrenia. npj Schizophrenia 3(1):2

Burns SB, Szyszkowicz JK, Luheshi GN, Lutz PE, Turecki G (2018) Plasticity of the epigenome during early-life stress. Semin Cell Dev Biol 77:115–132

Burt SA (2009) Are there meaningful etiological differences within antisocial behavior? results of a meta-analysis. Clin Psychol Rev 29(2):163–78

Cai L, Huang T, Su J, Zhang X, Chen W, Zhang F, He L, Chou KC (2018) Implications of newly identified brain eqtl genes and their interactors in schizophrenia. Mol Ther Nucleic Acids 12:433–442

Cameron D, Mi D, Vinh NN, Webber C, Li M, Marín O, O'Donovan MC, Bray NJ (2023) Single-nuclei rna sequencing of 5 regions of the human prenatal brain implicates developing neuron populations in genetic risk for schizophrenia. Biol Psychiatry 93(2):157–166

Caspi A, Moffitt TE (2006) Gene–environment interactions in psychiatry: joining forces with neuroscience. Nature Reviews Neuroscience 7(7):583–590

Caspi A, McClay J, Moffitt TE, Mill J, Martin J, Craig IW, Taylor A, Poulton R (2002) Role of genotype in the cycle of violence in maltreated children. Science 297(5582):851–4

Cavalli G, Heard E (2019) Advances in epigenetics link genetics to the environment and disease. Nature 571(7766):489–499

Chatterjee N, Shi J, García-Closas M (2016) Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nature Reviews Genetics 17(7):392–406

Chen C, Meng Q, Xia Y, Ding C, Wang L, Dai R, Cheng L, Gunaratne P, Gibbs RA, Min S, Coarfa C, Reid JG, Zhang C, Jiao C, Jiang Y, Giase G, Thomas A, Fitzgerald D, Brunetti T, Shieh A, Xia C, Wang Y, Wang Y, Badner JA, Gershon ES, White KP, Liu C (2018) The transcription factor pou3f2 regulates a gene coexpression network in brain tissue from patients with psychiatric disorders. Sci Transl Med 10(472)

Chen J, Cao H, Kaufmann T, Westlye LT, Tost H, Meyer-Lindenberg A, Schwarz E (2020) Identification of reproducible bcl11a alterations in schizophrenia through individual-level prediction of coexpression. Schizophr Bull 46(5):1165–1171

Chun S, Casparino A, Patsopoulos NA, Croteau-Chonka DC, Raby BA, De Jager PL, Sunyaev SR, Cotsapas C (2017) Limited statistical evidence for shared genetic effects of eqtls and autoimmune-disease-associated loci in three major immune-cell types. Nat Genet 49(4):600–605

Collado-Torres L, Burke EE, Peterson A, Shin J, Straub RE, Rajpurohit A, Semick SA, Ulrich WS, Price AJ, Valencia C, Tao R, Deep-Soboslay A, Hyde TM, Kleinman JE, Weinberger DR, Jaffe AE (2019) Regional heterogeneity in gene expression, regulation, and coherence in the frontal cortex and hippocampus across development and schizophrenia. Neuron 103(2):203–216.e8

Collins FS, Morgan M, Patrinos A (2003) The human genome project: lessons from large-scale biology. Science 300(5617):286–90

Congdon E, Canli T (2008) A neurogenetic approach to impulsivity. Journal of Personality 76(6):1447–1484

Connally NJ, Nazeen S, Lee D, Shi H, Stamatoyannopoulos J, Chun S, Cotsapas C, Cassa

CA, Sunyaev SR (2022) The missing link between genetic association and regulatory function. Elife 11

Crow M, Suresh H, Lee J, Gillis J (2022) Coexpression reveals conserved gene programs that co-vary with cell type across kingdoms. Nucleic Acids Research 50(8):4302–4314

Daskalakis NP, Iatrou A, Chatzinakos C, Jajoo A, Snijders C, Wylie D, DiPietro CP, Tsatsani I, Chen CY, Pernia CD, Soliva-Estruch M, Arasappan D, Bharadwaj RA, Collado-Torres L, Wuchty S, Alvarez VE, Dammer EB, Deep-Soboslay A, Duong DM, Eagles N, Huber BR, Huuki L, Holstein VL, Logue MW, Lugenbühl JF, Maihofer AX, Miller MW, Nievergelt CM, Pertea G, Ross D, Sendi MSE, Sun BB, Tao R, Tooke J, Wolf EJ, Zeier Z, Berretta S, Champagne FA, Hyde T, Seyfried NT, Shin JH, Weinberger DR, Nemeroff CB, Kleinman JE, Ressler KJ (2024) Systems biology dissection of ptsd and mdd across brain regions, cell types, and blood. Science 384(6698):eadh3707

Davis J, Eyre H, Jacka FN, Dodd S, Dean O, McEwen S, Debnath M, McGrath J, Maes M, Amminger P, McGorry PD, Pantelis C, Berk M (2016) A review of vulnerability and risks for schizophrenia: Beyond the two hit hypothesis. Neurosci Biobehav Rev 65:185–94

Ding Y, Hou K, Xu Z, Pimplaskar A, Petter E, Boulier K, Privé F, Vilhjálmsson BJ, Loohuis LO, Pasaniuc B (2023) Polygenic scoring accuracy varies across the genetic ancestry continuum. Nature 618(7966):774–781

Domingos P (2012) A few useful things to know about machine learning. Communications of the ACM 55(10):78–87

Duan J, et al. (2003a) Polymorphisms in the 5-untranslated region of the human serotonin receptor 1b (htr1b) gene affect gene expression. Molecular Psychiatry 8:901–910

Duan J, et al. (2003b) Synonymous mutations in the human dopamine receptor d2 (drd2) affect mrna stability and synthesis of the receptor. Human Molecular Genetics 12:205–216

Duarte RRR, Troakes C, Nolan M, Srivastava DP, Murray RM, Bray NJ (2016) Genome-wide significant schizophrenia risk variation on chromosome 10q24 is associated with altered cis-regulation of borcs7, as3mt, and nt5c2 in the human brain. Am J Med Genet B Neuropsychiatr Genet 171(6):806–14

Dudbridge F (2013) Power and predictive accuracy of polygenic risk scores. PLOS Genetics 9(3):e1003348

Edwards SL, Beesley J, French JD, Dunning AM (2013) Beyond gwass: illuminating the dark road from association to function. Am J Hum Genet 93(5):779–97

Fagny M, Paulson JN, Kuijjer ML, Sonawane AR, Chen CY, Lopes-Ramos CM, Glass K, Quackenbush J, Platig J (2017) Exploring regulation in tissues with eqtl networks. Proc Natl Acad Sci U S A 114(37):E7841–e7850

Fazio L, Pergola G, Papalino M, Di Carlo P, Monda A, Gelao B, Amoroso N, Tangaro S, Rampino A, Popolizio T, Bertolino A, Blasi G (2018) Transcriptomic context of drd1 is associated with prefrontal activity and behavior during working memory. Proc Natl Acad Sci U S A 115(21):5582–5587

Feinberg AP, Fallin MD (2015) Epigenetics at the crossroads of genes and the environment. Jama 314(11):1129–30

Fisher RA (1915) Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika 10(4):507–521

Franke B, Buitelaar JK (2018) Gene-environment interactions. In: Textbook of Epigenetics

Friston K, Brown HR, Siemerkus J, Stephan KE (2016) The dysconnection hypothesis (2016). Schizophr Res 176(2-3):83–94

Fromer M, Roussos P, Sieberts SK, Johnson JS, Kavanagh DH, Perumal TM, Ruderfer DM, Oh EC, Topol A, Shah HR, Klei LL, Kramer R, Pinto D, Gümüş ZH, Cicek AE, Dang KK,

Browne A, Lu C, Xie L, Readhead B, Stahl EA, Xiao J, Parvizi M, Hamamsy T, Fullard JF, Wang YC, Mahajan MC, Derry JM, Dudley JT, Hemby SE, Logsdon BA, Talbot K, Raj T, Bennett DA, De Jager PL, Zhu J, Zhang B, Sullivan PF, Chess A, Purcell SM, Shinobu LA, Mangravite LM, Toyoshiba H, Gur RE, Hahn CG, Lewis DA, Haroutunian V, Peters MA, Lipska BK, Buxbaum JD, Schadt EE, Hirai K, Roeder K, Brennand KJ, Katsanis N, Domenici E, Devlin B, Sklar P (2016) Gene expression elucidates functional impact of polygenic risk for schizophrenia. Nat Neurosci 19(11):1442–1453

Fryett JJ, Morris AP, Cordell HJ (2020) Investigation of prediction accuracy and the impact of sample size, ancestry, and tissue in transcriptome-wide association studies. Genet Epidemiol 44(5):425–441

Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, Eyler AE, Denny JC, Nicolae DL, Cox NJ, Im HK (2015) A gene-based association method for mapping traits using reference transcriptome data. Nat Genet 47(9):1091–8

Gandal MJ, Leppa V, Won H, Parikshak NN, Geschwind DH (2016) The road to precision psychiatry: translating genetics into disease mechanisms. Nat Neurosci 19(11):1397–1407

Gandal MJ, Haney JR, Parikshak NN, Leppa V, Ramaswami G, Hartl C, Schork AJ, Appadurai V, Buil A, Werge TM, Liu C, White KP, Horvath S, Geschwind DH (2018a) Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. Science 359(6376):693–697

Gandal MJ, Zhang P, Hadjimichael E, Walker RL, Chen C, Liu S, Won H, van Bakel H, Varghese M, Wang Y, Shieh AW, Haney J, Parhami S, Belmont J, Kim M, Moran Losada P, Khan Z, Mleczko J, Xia Y, Dai R, Wang D, Yang YT, Xu M, Fish K, Hof PR, Warrell J, Fitzgerald D, White K, Jaffe AE, Peters MA, Gerstein M, Liu C, Iakoucheva LM, Pinto D, Geschwind DH (2018b) Transcriptome-wide isoform-level dysregulation in asd, schizophrenia, and bipolar disorder. Science 362(6420)

Gasperini M, Hill AJ, McFaline-Figueroa JL, Martin B, Kim S, Zhang MD, Jackson D, Leith A, Schreiber J, Noble WS, Trapnell C, Ahituv N, Shendure J (2019) A genome-wide framework for mapping gene regulation via cellular genetic screens. Cell 176(1-2):377–390

Gidziela A (2024) Behaviour problems and co-occurring developmental conditions: genes, environments and their interplay. PhD thesis, Queen Mary University of London

Gilad Y, Rifkin SA, Pritchard JK (2008) Revealing the architecture of gene regulation: the promise of eqtl studies. Trends Genet 24(8):408–15

Glinos DA, et al (2022) Transcriptome variation in human tissues revealed by long-read sequencing. Nature 608:353–359

Gottesman I, Shields J (1967) A polygenic theory of schizophrenia. Proc Natl Acad Sci U S A 58(1):199–205

Greenwood TA, Kelsoe JR (2003) Promoter and intronic variants affect the transcriptional regulation of the human dopamine transporter gene. Genomics 82(5):511–520

GTEx (2015) Human genomics. the genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. Science 348(6235):648–60

GTEx (2017) Genetic effects on gene expression across human tissues. Nature 550(7675):204–213

GTEx (2020) The gtex consortium atlas of genetic regulatory effects across human tissues. Science 369(6509):1318–1330

Guloksuz S, Pries LK, Delespaul P, Kenis G, Luykx JJ, Lin BD, Richards AL, Akdede B, Binbay T, Altınyazar V, Yalınçetin B, Gümüş-Akay G, Cihan B, Soygür H, Ulaş H, Cankurtaran E, Kaymak SU, Mihaljevic MM, Petrovic SA, Mirjanic T, Bernardo M, Cabrera B, Bobes J, Saiz PA, García-Portilla MP, Sanjuan J, Aguilar EJ, Santos JL, Jiménez-López E, Arrojo M, Carracedo A, López G, González-Peñas J, Parellada M,

Maric NP, Atbaşog Lu C, Ucok A, Alptekin K, Saka MC, Arango C, O'Donovan M, Rutten BPF, van Os J (2019) Examining the independent and joint effects of molecular genetic liability and environmental exposures in schizophrenia: results from the eugei study. World Psychiatry 18(2):173–182

Gupta S, Gupta A (2019) Dealing with noise problem in machine learning data-sets: A systematic review. Procedia Computer Science 161:466–474

Gurdasani D, Barroso I, Zeggini E (2019) Genomics of disease risk in globally diverse populations. Nature Reviews Genetics 20(9):520–535

Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, Jansen R, de Geus EJC, Boomsma DI, Wright FA, Sullivan PF, Nikkola E, Alvarez M, Civelek M, Lusis AJ, Lehtimäki T, Raitoharju E, Kähönen M, Seppälä I, Raitakari OT, Kuusisto J, Laakso M, Price AL, Pajukanta P, Pasaniuc B (2016) Integrative approaches for large-scale transcriptome-wide association studies. Nature Genetics 48(3):245–252

Gusev A, Mancuso N, Won H, Kousi M, Finucane HK, Reshef Y, Song L, Safi A, McCarroll S, Neale BM, Ophoff RA, O'Donovan MC, Crawford GE, Geschwind DH, Katsanis N, Sullivan PF, Pasaniuc B, Price AL (2018) Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. Nat Genet 50(4):538–548

Habib N, Avraham-Davidi I, Basu A, Burks T, Shekhar K, Hofree M, Choudhury SR, Aguet F, Gelfand E, Ardlie K, Weitz DA, Rozenblatt-Rosen O, Zhang F, Regev A (2017) Massively parallel single-nucleus rna-seq with dronc-seq. Nat Methods 14(10):955–958

Hall LS, Medway CW, Pain O, Pardiñas AF, Rees EG, Escott-Price V, Pocklington A, Bray NJ, Holmans PA, Walters JTR, Owen MJ, O'Donovan MC (2020) A transcriptome-wide association study implicates specific pre- and post-synaptic abnormalities in schizophrenia. Hum Mol Genet 29(1):159–167

Hall LS, Pain O, O'Brien HE, Anney R, Walters JTR, Owen MJ, O'Donovan MC, Bray NJ (2021) Cis-effects on gene expression in the human prenatal brain associated with genetic risk for neuropsychiatric disorders. Molecular Psychiatry 26(6):2082–2088

Hartl CL, Ramaswami G, Pembroke WG, Muller S, Pintacuda G, Saha A, Parsana P, Battle A, Lage K, Geschwind DH (2021) Coexpression network architecture reveals the brain-wide and multiregional basis of disease susceptibility. Nat Neurosci 24(9):1313–1323

Hawkins DM (2004) The problem of overfitting. Journal of Chemical Information and Computer Sciences 44(1):1–12

Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA, van de Lagemaat LN, Smith KA, Ebbert A, Riley ZL, Abajian C, Beckmann CF, Bernard A, Bertagnolli D, Boe AF, Cartagena PM, Chakravarty MM, Chapin M, Chong J, Dalley RA, Daly BD, Dang C, Datta S, Dee N, Dolbeare TA, Faber V, Feng D, Fowler DR, Goldy J, Gregor BW, Haradon Z, Haynor DR, Hohmann JG, Horvath S, Howard RE, Jeromin A, Jochim JM, Kinnunen M, Lau C, Lazarz ET, Lee C, Lemon TA, Li L, Li Y, Morris JA, Overly CC, Parker PD, Parry SE, Reding M, Royall JJ, Schulkin J, Sequeira PA, Slaughterbeck CR, Smith SC, Sodt AJ, Sunkin SM, Swanson BE, Vawter MP, Williams D, Wohnoutka P, Zielke HR, Geschwind DH, Hof PR, Smith SM, Koch C, Grant SGN, Jones AR (2012) An anatomically comprehensive atlas of the adult human brain transcriptome. Nature 489(7416):391–399

Hindley G, O'Connell KS, Rahman Z, Frei O, Bahrami S, Shadrin A, Høegh MC, Cheng W, Karadag N, Lin A, Rødevand L, Fan CC, Djurovic S, Lagerberg TV, Dale AM, Smeland OB, Andreassen OA (2022) The shared genetic basis of mood instability and psychiatric disorders: A cross-trait genome-wide association analysis. Am J Med Genet B Neuropsychiatr Genet 189(6):207–218

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA

(2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106(23):9362–7

Hoffman GE, Ma Y, Montgomery KS, Bendl J, Jaiswal MK, Kozlenkov A, Peters MA, Dracheva S, Fullard JF, Chess A, Devlin B, Sieberts SK, Roussos P (2022) Sex differences in the human brain transcriptome of cases with schizophrenia. Biol Psychiatry 91(1):92–101

van Hogezand L (2016) The role of gene x social environment interactions in psychiatric disorders. PhD thesis, University of Groningen

Howard DM, Adams MJ, Clarke TK, Hafferty JD, Gibson J, Shirali M, Coleman JRI, Hagenaars SP, Ward J, Wigmore EM, Alloza C, Shen X, Barbu MC, Xu E, Whalley HC, Marioni RE, Porteous DJ, Davies G, Deary IJ, Hemani G, Lewis G, McIntosh AM (2019) Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nature Neuroscience 22:343–352

Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5(6):e1000529

Huckins LM, Dobbyn A, Ruderfer DM, Hoffman G, Wang W, Pardiñas AF, Rajagopal VM, Als TD, H TN, Girdhar K, Boocock J, Roussos P, Fromer M, Kramer R, Domenici E, Gamazon ER, Purcell S, Demontis D, Børglum AD, Walters JTR, O'Donovan MC, Sullivan P, Owen MJ, Devlin B, Sieberts SK, Cox NJ, Im HK, Sklar P, Stahl EA (2019) Gene expression imputation across multiple brain regions provides insights into schizophrenia risk. Nat Genet 51(4):659–674

Huo Y, Li S, Liu J, Li X, Luo XJ (2019) Functional genomics reveal gene regulatory mechanisms underlying schizophrenia risk. Nat Commun 10(1):670

Huynh PH, Nguyen VH, Do TN (2020) Improvements in the large p, small n classification issue. SN Computer Science 1

Inouye M, Abraham G, Nelson CP, Wood AM, Sweeting MJ, Dudbridge F, Lai FY, Kaptoge S, Khaw KT, Samani NJ, Butterworth AS, Di Angelantonio E (2018) Genomic risk prediction of coronary artery disease in 480,000 adults. Journal of the American College of Cardiology 72(16):1883–1893

International Human Genome Sequencing C (2004) Finishing the euchromatic sequence of the human genome. Nature 431(7011):931–945

Jaffe AE, Straub RE, Shin JH, Tao R, Gao Y, Collado-Torres L, Kam-Thong T, Xi HS, Quan J, Chen Q, Colantuoni C, Ulrich WS, Maher BJ, Deep-Soboslay A, Cross AJ, Brandon NJ, Leek JT, Hyde TM, Kleinman JE, Weinberger DR (2018) Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. Nat Neurosci 21(8):1117–1125

Jaffe AE, Hoeppner DJ, Saito T, Blanpain L, Ukaigwe J, Burke EE, Collado-Torres L, Tao R, Tajinda K, Maynard KR, Tran MN, Martinowich K, Deep-Soboslay A, Shin JH, Kleinman JE, Weinberger DR, Matsumoto M, Hyde TM (2020) Profiling gene expression in the human dentate gyrus granule cell layer reveals insights into schizophrenia and its genetic risk. Nat Neurosci 23(4):510–519

Jaffe AE, Tao R, Page SC, Maynard KR, Pattie EA, Nguyen CV, Deep-Soboslay A, Bharadwaj R, Young KA, Friedman MJ, Williamson DE, Shin JH, Hyde TM, Martinowich K, Kleinman JE (2022) Decoding shared versus divergent transcriptomic signatures across cortico-amygdala circuitry in ptsd and depressive disorders. Am J Psychiatry 179(9):673–686

Kachuri L, Chatterjee N, Hirbo J, Schaid DJ, Martin I, Kullo IJ, Kenny EE, Pasaniuc B,

Witte JS, Ge T (2024) Principles and methods for transferring polygenic risk scores across global populations. Nat Rev Genet 25(1):8–25

Kendler KS, Baker JH (2007) Genetic influences on measures of the environment: a systematic review. Psychol Med 37(5):615–26

Kolberg L, Kerimov N, Peterson H, Alasoo K (2020) Co-expression analysis reveals interpretable gene modules controlled by trans-acting genetic variants. Elife 9

Kursa MB, Rudnicki WR (2010) Feature selection with the boruta package. Journal of Statistical Software 36(11):1 – 13

Kustatscher G, et al (2022) Co-regulation of protein complexes reveals the function and evolution of cellular modules. Nature Reviews Molecular Cell Biology 23(11):676–689

Langfelder P, Horvath S (2008) Wgcna: an r package for weighted correlation network analysis. BMC Bioinformatics 9:559

Lee C (2022) Towards the genetic architecture of complex gene expression traits: Challenges and prospects for eqtl mapping in humans. Genes (Basel) 13(2)

de Leeuw CA, Mooij JM, Heskes T, Posthuma D (2015) Magma: generalized gene-set analysis of gwas data. PLoS Comput Biol 11(4):e1004219

Lemonde S, et al. (2003) Impaired repression at a 5-hydroxytryptamine 1a receptor gene polymorphism associated with major depression and suicide. Journal of Neuroscience 23(25):8788–8799

Leucht S, Chaimani A, Krause M, Schneider-Thoma J, Wang D, Dong S, Samara M, Peter N, Huhn M, Priller J, Davis JM (2022) The response of subgroups of patients with schizophrenia to different antipsychotic drugs: a systematic review and meta-analysis. Lancet Psychiatry 9(11):884–893

Li B, Ritchie MD (2021) From gwas to gene: Transcriptome-wide association studies and other methods to functionally understand gwas discoveries. Front Genet 12:713230

Li B, Verma SS, Veturi YC, Verma A, Bradford Y, Haas DW, Ritchie MD (2018a) Evaluation of predixcan for prioritizing gwas associations and predicting gene expression. Pacific Symposium on Biocomputing 23:448–459

Li M, Jaffe AE, Straub RE, Tao R, Shin JH, Wang Y, Chen Q, Li C, Jia Y, Ohi K, Maher BJ, Brandon NJ, Cross A, Chenoweth JG, Hoeppner DJ, Wei H, Hyde TM, McKay R, Kleinman JE, Weinberger DR (2016) A human-specific as3mt isoform and borcs7 are molecular risk factors in the 10q24.32 schizophrenia-associated locus. Nat Med 22(6):649–56

Li M, Santpere G, Imamura Kawasawa Y, Evgrafov OV, Gulden FO, Pochareddy S, Sunkin SM, Li Z, Shin Y, Zhu Y, Sousa AMM, Werling DM, Kitchen RR, Kang HJ, Pletikos M, Choi J, Muchnik S, Xu X, Wang D, Lorente-Galdos B, Liu S, Giusti-Rodríguez P, Won H, de Leeuw CA, Pardiñas AF, Hu M, Jin F, Li Y, Owen MJ, O'Donovan MC, Walters JTR, Posthuma D, Reimers MA, Levitt P, Weinberger DR, Hyde TM, Kleinman JE, Geschwind DH, Hawrylycz MJ, State MW, Sanders SJ, Sullivan PF, Gerstein MB, Lein ES, Knowles JA, Sestan N (2018b) Integrative functional genomic analysis of human brain development and neuropsychiatric risks. Science 362(6420)

Li M, Santpere G, Imamura Kawasawa Y, Evgrafov OV, Gulden FO, Pochareddy S, Sunkin SM, Li Z, Shin Y, Zhu Y, Sousa AMM, Werling DM, Kitchen RR, Kang HJ, Pletikos M, Choi J, Muchnik S, Xu X, Wang D, Lorente-Galdos B, Liu S, Giusti-Rodríguez P, Won H, de Leeuw CA, Pardiñas AF, Hu M, Jin F, Li Y, Owen MJ, O'Donovan MC, Walters JTR, Posthuma D, Reimers MA, Levitt P, Weinberger DR, Hyde TM, Kleinman JE, Geschwind DH, Hawrylycz MJ, State MW, Sanders SJ, Sullivan PF, Gerstein MB, Lein ES, Knowles JA, Sestan N (2018c) Integrative functional genomic analysis of human brain development and neuropsychiatric risks. Science 362(6420)

Liu S, Won H, Clarke D, Matoba N, Khullar S, Mu Y, Wang D, Gerstein M (2022) Illuminating links between cis-regulators and trans-acting variants in the human prefrontal cortex. Genome Med 14(1):133

Liu X, Li YI, Pritchard JK (2019) Trans effects on gene expression can drive omnigenic inheritance. Cell 177(4):1022–1034.e6

Lo MT, Hinds DA, Tung JY, Franz C, Fan CC, Wang Y, Smeland OB, Schork A, Holland D, Kauppi K, Sanyal N, Escott-Price V, Smith DJ, O'Donovan M, Stefansson H, Bjornsdottir G, Thorgeirsson TE, Stefansson K, McEvoy LK, Dale AM, Andreassen OA, Chen CH (2017) Genome-wide analyses for personality traits identify six genomic loci and show correlations with psychiatric disorders. Nature Genetics 49(1):152–156

Lu Y, Kowalec K, Song J, Karlsson R, Harder A, Giusti-Rodríguez P, Sullivan PF, Yao S (2023) Subtyping schizophrenia using psychiatric polygenic scores. medRxiv

Luningham JM, Chen J, Tang S, De Jager PL, Bennett DA, Buchman AS, Yang J (2020) Bayesian genome-wide twas method to leverage both cis- and trans-eqtl information through summary statistics. Am J Hum Genet 107(4):714–726

Lupi AS, Vazquez AI, de los Campos G (2024) Mapping the relative accuracy of cross-ancestry prediction. Nature Communications 15:10480

Ma L, Semick SA, Chen Q, Li C, Tao R, Price AJ, Shin JH, Jia Y, Brandon NJ, Cross AJ, Hyde TM, Kleinman JE, Jaffe AE, Weinberger DR, Straub RE (2020) Schizophrenia risk variants influence multiple classes of transcripts of sorting nexin 19 (snx19). Mol Psychiatry 25(4):831–843

Ma L, Shcherbina A, Chetty S (2021) Variations and expression features of cyp2d6 contribute to schizophrenia risk. Mol Psychiatry 26(6):2605–2615

Mai J, Lu M, Gao Q, Zeng J, Xiao J (2023) Transcriptome-wide association studies: recent advances in methods, applications and available databases. Communications Biology 6(1):899

Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ (2019) Clinical use of current polygenic risk scores may exacerbate health disparities. Nature Genetics 51(4):584–591

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, Shafer A, Neri F, Lee K, Kutyavin T, Stehling-Sun S, Johnson AK, Canfield TK, Giste E, Diegel M, Bates D, Hansen RS, Neph S, Sabo PJ, Heimfeld S, Raubitschek A, Ziegler S, Cotsapas C, Sotoodehnia N, Glass I, Sunyaev SR, Kaul R, Stamatoyannopoulos JA (2012) Systematic localization of common disease-associated variation in regulatory dna. Science 337(6099):1190–5

Maynard KR, et al (2021) Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. Nature Neuroscience 24:425–436

McAdams TA, Gregory AM, Eley TC (2013) Genes of experience: explaining the heritability of putative environmental variables through their association with behavioural and emotional traits. Behav Genet 43(4):314–28

McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, Kang HM, Fuchsberger C, Danecek P, Sharp K, Luo Y, Sidore C, Kwong A, Timpson N, Koskinen S, Vrieze S, Scott LJ, Zhang H, Mahajan A, Veldink J, Peters U, Pato C, van Duijn CM, Gillies CE, Gandin I, Mezzavilla M, Gilly A, Cocca M, Traglia M, Angius A, Barrett JC, Boomsma D, Branham K, Breen G, Brummett CM, Busonero F, Campbell H, Chan A, Chen S, Chew E, Collins FS, Corbin LJ, Smith GD, Dedoussis G, Dorr M, Farmaki AE, Ferrucci L, Forer L, Fraser RM, Gabriel S, Levy S, Groop L, Harrison T, Hattersley A, Holmen OL, Hveem K, Kretzler M, Lee JC, McGue M, Meitinger T, Melzer D, Min JL,

Mohlke KL, Vincent JB, Nauck M, Nickerson D, Palotie A, Pato M, Pirastu N, McInnis M, Richards JB, Sala C, Salomaa V, Schlessinger D, Schoenherr S, Slagboom PE, Small K, Spector T, Stambolian D, Tuke M, Tuomilehto J, Van den Berg LH, Van Rheenen W, Volker U, Wijmenga C, Toniolo D, Zeggini E, Gasparini P, Sampson MG, Wilson JF, Frayling T, de Bakker PIW, Swertz MA, McCarroll S, Kooperberg C, Dekker A, Altshuler D, Willer C, Iacono W, Ripatti S, et al. (2016) A reference panel of 64,976 haplotypes for genotype imputation. Nature Genetics 48(10):1279–1283

McGue M, Carey G (2017) Gene–environment interaction in the behavioral sciences: Findings, challenges, and prospects. In: Behavior Genetics of Cognition Across the Lifespan, Springer, pp 49–75

Mill J, Asherson P, Browes C, D'Souza U, Craig I (2002) Expression of the dopamine transporter gene is regulated by the 3' utr vntr: Evidence from brain and lymphocytes using quantitative rt-pcr. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics 114:975–979

Miller GM, Madras BK (2002) Polymorphisms in the 3-untranslated region of human and monkey dopamine transporter genes affect reporter gene expression. Molecular Psychiatry 7:44–55

Miyamoto S, Miyake N, Jarskog LF, Fleischhacker WW, Lieberman JA (2012) Pharmacological treatment of schizophrenia: a critical review of the pharmacology and clinical effects of current and future therapeutic agents. Molecular Psychiatry 17(12):1206–1227

Moeller FG, Barratt ES, Dougherty DM, Schmitz JM, Swann AC (2001) Psychiatric aspects of impulsivity. American Journal of Psychiatry 158(11):1783–1793

Moffitt TE (2005) The new look of behavioral genetics in developmental psychopathology: Gene–environment interplay in antisocial behaviors. Psychological Bulletin 131(4):533–554

Moreno-Estrada A, Gignoux CR, Fernandez-Lopez JC, Zakharia F, Sikora M, Contreras AV, Acuna-Alonzo V, Sandoval K, Eng C, Romero-Hidalgo S, et al. (2013) The genetics of mexico recapitulates native american substructure and affects biomedical traits. Science 344(6189):1280–1285

Mostafavi H, et al (2020) Variable prediction accuracy of polygenic scores within an ancestry group. eLife 9:e48376

Mostafavi H, Spence JP, Naqvi S, Pritchard JK (2023) Systematic differences in discovery of genetic effects on gene expression and complex traits. Nat Genet 55(11):1866–1875

Mostafavi S, Gaiteri C, Sullivan SE, White CC, Tasaki S, Xu J, Taga M, Klein HU, Patrick E, Komashko V, McCabe C, Smith R, Bradshaw EM, Root DE, Regev A, Yu L, Chibnik LB, Schneider JA, Young-Pearse TL, Bennett DA, De Jager PL (2018) A molecular network of the aging human brain provides insights into the pathology and cognitive decline of alzheimer's disease. Nature Neuroscience 21(6):811–819

Nath AP, Ritchie SC, Byars SG, Fearnley LG, Havulinna AS, Joensuu A, Kangas AJ, Soininen P, Wennerström A, Milani L, Metspalu A, Männistö S, Würtz P, Kettunen J, Raitoharju E, Kähönen M, Juonala M, Palotie A, Ala-Korpela M, Ripatti S, Lehtimäki T, Abraham G, Raitakari O, Salomaa V, Perola M, Inouye M (2017) An interaction map of circulating metabolites, immune gene networks, and their genetic regulation. Genome Biology 18(1):146

Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ (2010) Trait-associated snps are more likely to be eqtls: annotation to enhance discovery from gwas. PLoS Genetics 6(4):e1000888

Nilsson K, Åslund C, Comasco E (2018) Gene–environment interaction of monoamine oxidase a in relation to antisocial behaviour: current and future directions. Journal of Neural Transmission

Notter T, Meyer U (2017) Microglia and schizophrenia: where next? Mol Psychiatry 22(6):788–789

Oldham MC, et al (2008) Functional organization of the transcriptome in human brain. Nature Neuroscience 11:1271–1282

Oldham MC, Horvath S, Geschwind DH (2006) Conservation and evolution of gene coexpression networks in human and chimpanzee brains. Proceedings of the National Academy of Sciences 103(47):17973–17978

Ouwens KG, Jansen R, Nivard MG, van Dongen J, Frieser MJ, Hottenga JJ, Arindrarto W, Claringbould A, van Iterson M, Mei H, Franke L, Heijmans BT, A C 't Hoen P, van Meurs J, Brooks AI, Heijmans BT, A C 't Hoen P, van Meurs J, Isaacs A, Jansen R, Franke L, Boomsma DI, Pool R, van Dongen J, Hottenga JJ, van Greevenbroek MMJ, Stehouwer CDA, van der Kallen CJH, Schalkwijk CG, Wijmenga C, Franke L, Zhernakova S, Tigchelaar EF, Slagboom PE, Beekman M, Deelen J, van Heemst D, Veldink JH, van den Berg LH, van Duijn CM, Hofman BA, Isaacs A, Uitterlinden AG, van Meurs J, Jhamai PM, Verbiest M, Suchiman HED, Verkerk M, van der Breggen R, van Rooij J, Lakenberg N, Mei H, van Iterson M, van Galen M, Bot J, Zhernakova DV, Jansen R, van't Hof P, Deelen P, Nooren I, A C 't Hoen P, Heijmans BT, Moed M, Franke L, Vermaat M, Zhernakova DV, Luijk R, Jan Bonder M, van Iterson M, Deelen P, van Dijk F, van Galen M, Arindrarto W, Kielbasa SM, Swertz MA, van Zwet EW, Jansen R, 't Hoen PB, Heijmans BT, Penninx BWJH, Boomsma DI, Consortium B (2020) A characterization of cis- and trans-heritability of rna-seq-based gene expression. European Journal of Human Genetics 28(2):253–263

Panagiotakos G, Pasca SP (2022) A matter of space and time: Emerging roles of disease-associated proteins in neural development. Neuron 110(2):195–208

Pardiñas AF, Holmans P, Pocklington AJ, Escott-Price V, Ripke S, Carrera N, Legge SE,

Bishop S, Cameron D, Hamshere ML, Han J, Hubbard L, Lynham A, Mantripragada K, Rees E, MacCabe JH, McCarroll SA, Baune BT, Breen G, Byrne EM, Dannlowski U, Eley TC, Hayward C, Martin NG, McIntosh AM, Plomin R, Porteous DJ, Wray NR, Caballero A, Geschwind DH, Huckins LM, Ruderfer DM, Santiago E, Sklar P, Stahl EA, Won H, Agerbo E, Als TD, Andreassen OA, Bækvad-Hansen M, Mortensen PB, Pedersen CB, Børglum AD, Bybjerg-Grauholm J, Djurovic S, Durmishi N, Pedersen MG, Golimbet V, Grove J, Hougaard DM, Mattheisen M, Molden E, Mors O, Nordentoft M, Pejovic-Milovancevic M, Sigurdsson E, Silagadze T, Hansen CS, Stefansson K, Stefansson H, Steinberg S, Tosato S, Werge T, Collier DA, Rujescu D, Kirov G, Owen MJ, O'Donovan MC, Walters JTR (2018) Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. Nat Genet 50(3):381–389

Patton JH, Stanford MS, Barratt ES (1995) Factor structure of the barratt impulsiveness scale. Journal of Clinical Psychology 51(6):768–774

Pergola G, Di Carlo P, Andriola I, Gelao B, Torretta S, Attrotto MT, Fazio L, Raio A, Albergo D, Masellis R, Rampino A, Blasi G, Bertolino A (2016) Combined effect of genetic variants in the glun2b coding gene (grin2b) on prefrontal function during working memory performance. Psychol Med 46(6):1135–50

Pergola G, Di Carlo P, D'Ambrosio E, Gelao B, Fazio L, Papalino M, Monda A, Scozia G, Pietrangelo B, Attrotto M, Apud JA, Chen Q, Mattay VS, Rampino A, Caforio G, Weinberger DR, Blasi G, Bertolino A (2017) Drd2 co-expression network and a related polygenic index predict imaging, behavioral and clinical phenotypes linked to schizophrenia. Transl Psychiatry 7(1):e1006

Pergola G, Di Carlo P, Jaffe AE, Papalino M, Chen Q, Hyde TM, Kleinman JE, Shin JH, Rampino A, Blasi G, Weinberger DR, Bertolino A (2019a) Prefrontal coexpression of schizophrenia risk genes is associated with treatment response in patients. Biol Psychiatry 86(1):45–55

Pergola G, Papalino M, Gelao B, Sportelli L, Vollerbergh W, Grattagliano I, Bertolino A (2019b) Evocative gene-environment correlation between genetic risk for schizophrenia and bullying victimization. World Psychiatry 18(3):366–367

Pergola G, Parihar M, Sportelli L, Bharadwaj R, Borcuk C, Radulescu E, Bellantuono L, Blasi G, Chen Q, Kleinman JE, Wang Y, Sripathy SR, Maher BJ, Monaco A, Rossi F, Shin JH, Hyde TM, Bertolino A, Weinberger DR (2023a) Consensus molecular environment of schizophrenia risk genes in coexpression networks shifting across age and brain regions. Sci Adv 9(15):eade2812

Pergola G, Penzel N, Sportelli L, Bertolino A (2023b) Lessons learned from parsing genetic risk for schizophrenia into biological pathways. Biol Psychiatry 94(2):121–130

Pergola G, Rampino A, Sportelli L, Borcuk CJ, Passiatore R, Di Carlo P, Marakhovskaia A, Fazio L, Amoroso N, Castro MN, Domenici E, Gennarelli M, Khlghatyan J, Kikidis GC, Lella A, Magri C, Monaco A, Papalino M, Parihar M, Popolizio T, Quarto T, Romano R, Torretta S, Valsecchi P, Zunuer H, Blasi G, Dukart J, Beaulieu JM, Bertolino A (2023c) A mir-137-related biological pathway of risk for schizophrenia is associated with human brain emotion processing. Biol Psychiatry Cogn Neurosci Neuroimaging

Perälä J, Suvisaari J, Saarni SI, Kuoppasalmi K, Isometsä E, Pirkola S, Partonen T, Tuulio-Henriksson A, Hintikka J, Kieseppä T, Härkänen T, Koskinen S, Lönnqvist J (2007) Lifetime prevalence of psychotic and bipolar i disorders in a general population. Archives of General Psychiatry 64(1):19–28

Pierce BL, Tong L, Chen LS, Rahaman R, Argos M, Jasmine F, Roy S, Paul-Brutus R, Westra HJ, Franke L, Esko T, Zaman R, Islam T, Rahman M, Baron JA, Kibriya MG, Ahsan H (2014) Mediation analysis demonstrates that trans-eqtls are often explained by cis-mediation: a genome-wide analysis among 1,800 south asians. PLoS Genet 10(12):e1004818

Polderman TJC, Benyamin B, de Leeuw CA, Sullivan PF, van Bochoven A, Visscher PM,

Posthuma D (2015) Meta-analysis of the heritability of human traits based on fifty years of twin studies. Nature Genetics 47(7):702–709

Popejoy AB, Fullerton SM (2016) Genomics is failing on diversity. Nature 538(7624):161–164

of the Psychiatric Genomics Consortium CDG (2019) Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. Cell 179(7):1469–1482.e11

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) Plink: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559–75

Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460(7256):748–52

Radulescu E, Jaffe AE, Straub RE, Chen Q, Shin JH, Hyde TM, Kleinman JE, Weinberger DR (2020) Identification and prioritization of gene sets associated with schizophrenia risk by co-expression network analysis in human brain. Mol Psychiatry 25(4):791–804

Raine A (2002) Biosocial studies of antisocial and violent behavior in children and adults: A review. Journal of Abnormal Child Psychology 30(4):311–326

Rajiv Tandon DMBJBREGSHDMMJOSSMTJVOWC Wolfgang Gaebel (2013) Definition and description of schizophrenia in the dsm-5. Schizophrenia Research 150(1):3–10

Ramaswami G, Won H, Gandal MJ, Haney J, Wang JC, Wong CCY, Sun W, Prabhakar S, Mill J, Geschwind DH (2020) Integrative genomics identifies a convergent molecular subtype that links epigenomic with transcriptomic differences in autism. Nat Commun 11(1):4873

Rhee SH, Waldman ID (2002) Genetic and environmental influences on antisocial behavior: a meta-analysis of twin and adoption studies. Psychol Bull 128(3):490–529

Richards J (2015) Plasticity genes, the social environment, and their interplay in adolescents with and without adhd. from behaviour to brain. PhD thesis, Radboud University

Ripke S, Neale BM, Corvin A, Walters JTR, Farh KH, Holmans PA, Lee P, Bulik-Sullivan B, Collier DA, Huang H, Pers TH, Agartz I, Agerbo E, Albus M, Alexander M, Amin F, Bacanu SA, Begemann M, Belliveau Jr RA, Bene J, Bergen SE, Bevilacqua E, Bigdeli TB, Black DW, Bruggeman R, Buccola NG, Buckner RL, Byerley W, Cahn W, Cai G, Campion D, Cantor RM, Carr VJ, Carrera N, Catts SV, Chambert KD, Chan RCK, Chen RYL, Chen EYH, Cheng W, Cheung EFC, Ann Chong S, Robert Cloninger C, Cohen D, Cohen N, Cormican P, Craddock N, Crowley JJ, Curtis D, Davidson M, Davis KL, Degenhardt F, Del Favero J, Demontis D, Dikeos D, Dinan T, Djurovic S, Donohoe G, Drapeau E, Duan J, Dudbridge F, Durmishi N, Eichhammer P, Eriksson J, Escott-Price V, Essioux L, Fanous AH, Farrell MS, Frank J, Franke L, Freedman R, Freimer NB, Friedl M, Friedman JI, Fromer M, Genovese G, Georgieva L, Giegling I, Giusti-Rodríguez P, Godard S, Goldstein JI, Golimbet V, Gopal S, Gratten J, de Haan L, Hammer C, Hamshere ML, Hansen M, Hansen T, Haroutunian V, Hartmann AM, Henskens FA, Herms S, Hirschhorn JN, Hoffmann P, Hofman A, Hollegaard MV, Hougaard DM, Ikeda M, Joa I, et al. (2014) Biological insights from 108 schizophrenia-associated genetic loci. Nature 511(7510):421–427

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) limma powers differential expression analyses for rna-sequencing and microarray studies. Nucleic Acids Res 43(7):e47

Rodriguez-López J, Arrojo M, Paz E, Páramo M, Costas J (2020) Identification of relevant hub genes for early intervention at gene coexpression modules with altered predicted expression in schizophrenia. Progress in Neuro-Psychopharmacology and Biological Psychiatry 98:109815

Romeo R, Knapp M, Scott S (2006) Economic cost of severe antisocial behaviour in children - and who pays it. British Journal of Psychiatry 188(6):547–553

Roussos P, Mitchell AC, Voloudakis G, Fullard JF, Pothula VM, Tsang J, Stahl EA, Georgakopoulos A, Ruderfer DM, Charney A, Okada Y, Siminovitch KA, Worthington J, Padyukov L, Klareskog L, Gregersen PK, Plenge RM, Raychaudhuri S, Fromer M, Purcell SM, Brennand KJ, Robakis NK, Schadt EE, Akbarian S, Sklar P (2014) A role for noncoding variation in schizophrenia. Cell Rep 9(4):1417–29

Ruisch H (2020) Gene-environment interactions in disruptive behaviors

Russell M, Aqil A, Saitou M, Gokcumen O, Masuda N (2023) Gene communities in co-expression networks across different tissues. PLOS Computational Biology 19(11):1–32

Ruzicka WB, Mohammadi S, Fullard JF, Davila-Velderrain J, Subburaju S, Tso DR, Hourihan M, Jiang S, Lee HC, Bendl J, Voloudakis G, Haroutunian V, Hoffman GE, Roussos P, Kellis M (2024) Single-cell multi-cohort dissection of the schizophrenia transcriptome. Science 384(6698):eadg5136

Sanchez-Roige S, Gray JC, MacKillop J, Chen CH, Palmer AA (2018) The genetics of human personality. Genes Brain Behav 17(3):e12439

Schoeler T, Choi SW, Dudbridge F, Baldwin J, Duncan L, Cecil CM, Walton E, Viding E, McCrory E, Pingault JB (2019) Multi–polygenic score approach to identifying individual vulnerabilities associated with the risk of exposure to bullying. JAMA Psychiatry 76(7):730–738

Sekar A, Bialas AR, de Rivera H, Davis A, Hammond TR, Kamitaki N, Tooley K, Presumey J, Baum M, Van Doren V, Genovese G, Rose SA, Handsaker RE, Daly MJ, Carroll MC, Stevens B, McCarroll SA, Schizophrenia Working Group of the Psychiatric Genomics C (2016) Schizophrenia risk from complex variation of complement component 4. Nature 530(7589):177–183

Sharma L, Markon KE, Clark LA (2014) Toward a theory of distinct types of "impulsive" behaviors: A meta-analysis of self-report and behavioral measures. Psychol Bull 140(2):374–408

Shen LX, Basilion JP, Stanton VPJ (1999) Single-nucleotide polymorphisms can cause different structural folds of mrna. Proceedings of the National Academy of Sciences 96(14):7871–7876

Shmueli G (2011) To explain or to predict? Statistical Science 25

Singh T, Poterba T, Curtis D, Akil H, Al Eissa M, Barchas JD, Bass N, Bigdeli TB, Breen G, Bromet EJ, Buckley PF, Bunney WE, Bybjerg-Grauholm J, Byerley WF, Chapman SB, Chen WJ, Churchhouse C, Craddock N, Cusick CM, DeLisi L, Dodge S, Escamilla MA, Eskelinen S, Fanous AH, Faraone SV, Fiorentino A, Francioli L, Gabriel SB, Gage D, Gagliano Taliun SA, Ganna A, Genovese G, Glahn DC, Grove J, Hall MH, Hämäläinen E, Heyne HO, Holi M, Hougaard DM, Howrigan DP, Huang H, Hwu HG, Kahn RS, Kang HM, Karczewski KJ, Kirov G, Knowles JA, Lee FS, Lehrer DS, Lescai F, Malaspina D, Marder SR, McCarroll SA, McIntosh AM, Medeiros H, Milani L, Morley CP, Morris DW, Mortensen PB, Myers RM, Nordentoft M, O'Brien NL, Olivares AM, Ongur D, Ouwehand WH, Palmer DS, Paunio T, Quested D, Rapaport MH, Rees E, Rollins B, Satterstrom FK, Schatzberg A, Scolnick E, Scott LJ, Sharp SI, Sklar P, Smoller JW, Sobell JL, Solomonson M, Stahl EA, Stevens CR, Suvisaari J, Tiao G, Watson SJ, Watts NA, Blackwood DH, Børglum AD, Cohen BM, Corvin AP, Esko T, Freimer NB, Glatt SJ, Hultman CM, McQuillin A, Palotie A, Pato CN, Pato MT, Pulver AE, St Clair D, et al. (2022) Rare coding variants in ten genes confer substantial risk for schizophrenia. Nature 604(7906):509–516

Skene NG, Bryois J, Bakken TE, Breen G, Crowley JJ, Gaspar HA, Giusti-Rodriguez P, Hodge RD, Miller JA, Muñoz-Manchado AB, O'Donovan MC, Owen MJ, Pardiñas AF, Ryge J, Walters JTR, Linnarsson S, Lein ES, Sullivan PF, Hjerling-Leffler J, Major Depres-

sive Disorder Working Group of the Psychiatric Genomics C (2018) Genetic identification of brain cell types underlying schizophrenia. Nature Genetics 50(6):825–833

Sportelli L, Eisenberg DP, Passiatore R, D'Ambrosio E, Antonucci LA, Bettina JS, Chen Q, Goldman AL, Gregory MD, Griffiths K, Hyde TM, Kleinman JE, Pardiñas AF, Parihar M, Popolizio T, Rampino A, Shin JH, Veronese M, Ulrich WS, Zink CF, Bertolino A, Howes OD, Berman KF, Weinberger DR, Pergola G (2024) Dopamine signaling enriched striatal gene set predicts striatal dopamine synthesis and physiological activity in vivo. Nat Commun 15(1):3342

Stanford MS, Mathias CW, Dougherty DM, Lake SL, Anderson NE, Patton JH (2009) Fifty years of the barratt impulsiveness scale: An update and review. Personality and Individual Differences 47(5):385–395

Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, Montgomery S, Tavaré S, Deloukas P, Dermitzakis ET (2007) Population genomics of human gene expression. Nature Genetics 39(10):1217–1224

Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science 302(5643):249–255

Stuart T, et al (2019) Comprehensive integration of single-cell data. Cell 177(7):1888–1902.e21

Sullivan PF, Geschwind DH (2019) Defining the genetic, genomic, cellular, and diagnostic architectures of psychiatric disorders. Cell 177(1):162–183

Sullivan PF, Kendler KS, Neale MC (2003) Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. Arch Gen Psychiatry 60(12):1187–92

Takata A, et al (2017) Integrative analyses of de novo mutations provide deeper biological insights into autism. Nature Genetics 49(7):974–980

Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, Taliun SAG, Corvelo A, Gogarten SM, Kang HM, Pitsillides AN, LeFaive J, Lee SB, Tian X, Browning BL, Das S, Emde AK, Clarke WE, Loesch DP, Shetty AC, Blackwell TW, Smith AV, Wong Q, Liu X, Conomos MP, Bobo DM, Aguet F, Albert C, Alonso A, Ardlie KG, Arking DE, Aslibekyan S, Auer PL, Barnard J, Barr RG, Barwick L, Becker LC, Beer RL, Benjamin EJ, Bielak LF, Blangero J, Boehnke M, Bowden DW, Brody JA, Burchard EG, Cade BE, Casella JF, Chalazan B, Chasman DI, Chen YI, Cho MH, Choi SH, Chung MK, Clish CB, Correa A, Curran JE, Custer B, Darbar D, Daya M, de Andrade M, DeMeo DL, Dutcher SK, Ellinor PT, Emery LS, Eng C, Fatkin D, Fingerlin T, Forer L, Fornage M, Franceschini N, Fuchsberger C, Fullerton SM, Germer S, Gladwin MT, Gottlieb DJ, Guo X, Hall ME, He J, Heard-Costa NL, Heckbert SR, Irvin MR, Johnsen JM, Johnson AD, Kaplan R, Kardia SLR, Kelly T, Kelly S, Kenny EE, Kiel DP, Klemmer R, Konkle BA, Kooperberg C, Köttgen A, Lange LA, Lasky-Su J, Levy D, Lin X, Lin KH, Liu C, Loos RJF, et al. (2021) Sequencing of 53,831 diverse genomes from the nhlbi topmed program. Nature 590(7845):290–299

Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D (2019) Benefits and limitations of genome-wide association studies. Nat Rev Genet 20(8):467–484

Teixeira JR, Szeto RA, Carvalho VMA, Muotri AR, Papes F (2021) Transcription factor 4 and its association with psychiatric disorders. Transl Psychiatry 11(1):19

Tielbeek JJ, Johansson A, Polderman TJC, Rautiainen MR, Jansen P, Taylor M, Tong X, Lu Q, Burt AS, Tiemeier H, Viding E, Plomin R, Martin NG, Heath AC, Madden PAF, Montgomery G, Beaver KM, Waldman I, Gelernter J, Kranzler HR, Farrer LA, Perry JRB, Munafò M, LoParo D, Paunio T, Tiihonen J, Mous SE, Pappa I, de Leeuw C, Watanabe K, Hammerschlag AR, Salvatore JE, Aliev F, Bigdeli TB, Dick D, Faraone SV, Popma A, Medland SE, Posthuma D (2017) Genome-wide association studies of a broad spectrum of antisocial behavior. JAMA Psychiatry 74(12):1242–1250

Tkachev D, Mimmack ML, Ryan MM, Wayland M, Freeman T, Jones PB, Starkey M, Webster MJ, Yolken RH, Bahn S (2003) Oligodendrocyte dysfunction in schizophrenia and bipolar disorder. Lancet 362(9386):798–805

Tokuhiro S, et al. (2003) An intronic snp in a runx1 binding site of slc22a4, encoding an organic cation transporter, is associated with rheumatoid arthritis. Nature Genetics 35:341–348

Torretta S, Rampino A, Basso M, Pergola G, Di Carlo P, Shin JH, Kleinman JE, Hyde TM, Weinberger DR, Masellis R, Blasi G, Pennuto M, Bertolino A (2020) Nurr1 and err1 modulate the expression of genes of a drd2 coexpression network enriched for schizophrenia risk. J Neurosci 40(4):932–941

Trubetskoy V, Pardiñas AF, Qi T, Panagiotaropoulou G, Awasthi S, Bigdeli TB, Bryois J, Chen CY, Dennison CA, Hall LS, Lam M, Watanabe K, Frei O, Ge T, Harwood JC, Koopmans F, Magnusson S, Richards AL, Sidorenko J, Wu Y, Zeng J, Grove J, Kim M, Li Z, Voloudakis G, Zhang W, Adams M, Agartz I, Atkinson EG, Agerbo E, Al Eissa M, Albus M, Alexander M, Alizadeh BZ, Alptekin K, Als TD, Amin F, Arolt V, Arrojo M, Athanasiu L, Azevedo MH, Bacanu SA, Bass NJ, Begemann M, Belliveau RA, Bene J, Benyamin B, Bergen SE, Blasi G, Bobes J, Bonassi S, Braun A, Bressan RA, Bromet EJ, Bruggeman R, Buckley PF, Buckner RL, Bybjerg-Grauholm J, Cahn W, Cairns MJ, Calkins ME, Carr VJ, Castle D, Catts SV, Chambert KD, Chan RCK, Chaumette B, Cheng W, Cheung EFC, Chong SA, Cohen D, Consoli A, Cordeiro Q, Costas J, Curtis C, Davidson M, Davis KL, de Haan L, Degenhardt F, DeLisi LE, Demontis D, Dickerson F, Dikeos D, Dinan T, Djurovic S, Duan J, Ducci G, Dudbridge F, Eriksson JG, Fañanás L, Faraone SV, Fiorentino A, Forstner A, Frank J, Freimer NB, Fromer M, Frustaci A, Gadelha A, Genovese G, Gershon ES, et al. (2022) Mapping genomic loci implicates genes and synaptic biology in schizophrenia. Nature 604(7906):502–508

Tuvblad C, Beaver KM (2013) Genetic and environmental influences on antisocial behavior. J Crim Justice 41(5):273–276

Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, Martin HC, Lappalainen T, Posthuma D (2021) Genome-wide association studies. Nature Reviews Methods Primers 1(1):59

Umans BD, Battle A, Gilad Y (2021) Where are the disease-associated eqtls? Trends Genet 37(2):109–124

van der Sijde MR, Ng A, Fu J (2014) Systems genetics: From gwas to disease pathways. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease 1842(10):1903–1909

Vasconcelos AG, Malloy-Diniz LF, Correa H (2015) Systematic review of psychometric properties of the barratt impulsiveness scale in neuropsychiatric samples. Clinical Neuropsychiatry 12(2):45–54

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J (2017) 10 years of gwas discovery: Biology, function, and translation. Am J Hum Genet 101(1):5–22

Võsa U, Claringbould A, Westra HJ, Bonder MJ, Deelen P, Zeng B, Kirsten H, Saha A, Kreuzhuber R, Yazar S, Brugge H, Oelen R, de Vries DH, van der Wijst MGP, Kasela S, Pervjakova N, Alves I, Favé MJ, Agbessi M, Christiansen MW, Jansen R, Seppälä I, Tong L, Teumer A, Schramm K, Hemani G, Verlouw J, Yaghootkar H, Sönmez Flitman R, Brown A, Kukushkina V, Kalnapenkis A, Rüeger S, Porcu E, Kronberg J, Kettunen J, Lee B, Zhang F, Qi T, Hernandez JA, Arindrarto W, Beutner F, Dmitrieva J, Elansary M, Fairfax BP, Georges M, Heijmans BT, Hewitt AW, Kähönen M, Kim Y, Knight JC, Kovacs P, Krohn K, Li S, Loeffler M, Marigorta UM, Mei H, Momozawa Y, Müller-Nurasyid M, Nauck M, Nivard MG, Penninx B, Pritchard JK, Raitakari OT, Rotzschke O, Slagboom EP, Stehouwer CDA, Stumvoll M, Sullivan P, t Hoen PAC, Thiery J, Tönjes A, van Dongen J, van Iterson M, Veldink JH, Völker U, Warmerdam R, Wijmenga C,

Swertz M, Andiappan A, Montgomery GW, Ripatti S, Perola M, Kutalik Z, Dermitzakis E, Bergmann S, Frayling T, van Meurs J, Prokisch H, Ahsan H, Pierce BL, Lehtimäki T, Boomsma DI, Psaty BM, Gharib SA, Awadalla P, Milani L, Ouwehand WH, Downes K, Stegle O, et al. (2021) Large-scale cis- and trans-eqtl analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet 53(9):1300–1310

Wainberg M, Merico D, Delong A, Frey BJ (2018) Deep learning in biomedicine. Nature Biotechnology 36(9):829–838

Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, Ermel R, Ruusalepp A, Quertermous T, Hao K, Björkegren JLM, Im HK, Pasaniuc B, Rivas MA, Kundaje A (2019) Opportunities and challenges for transcriptome-wide association studies. Nature Genetics 51(4):592–599

Walker RL, Ramaswami G, Hartl C, Mancuso N, Gandal MJ, de la Torre-Ubieta L, Pasaniuc B, Stein JL, Geschwind DH (2019) Genetic control of expression and splicing in developing human brain informs disease mechanisms. Cell 179(3):750–771.e22

Wang X, Christian KM, Song H, Ming GL (2018) Synaptic dysfunction in complex psychiatric disorders: from genetics to mechanisms. Genome Med 10(1):9

Watanabe K, Stringer S, Frei O, Umicevic Mirkov M, de Leeuw C, Polderman TJC, van der Sluis S, Andreassen OA, Neale BM, Posthuma D (2019) A global overview of pleiotropy and genetic architecture in complex traits. Nature Genetics 51(9):1339–1348

Weaver ICG, Cervoni N, Champagne FA, D'Alessio AC, Sharma S, Seckl JR, Dymov S, Szyf M, Meaney MJ (2004) Epigenetic programming by maternal behavior. Nature Neuroscience 7(8):847–854

Weinstein A, Dannon P (2015) Is impulsivity a male trait rather than female trait? exploring the sex difference in impulsivity. Current Behavioral Neuroscience Reports 2(1):9–14

Werling DM, Pochareddy S, Choi J, An JY, Sheppard B, Peng M, Li Z, Dastmalchi C, Santpere G, Sousa AMM, Tebbenkamp ATN, Kaur N, Gulden FO, Breen MS, Liang L, Gilson MC, Zhao X, Dong S, Klei L, Cicek AE, Buxbaum JD, Adle-Biassette H, Thomas JL, Aldinger KA, O'Day DR, Glass IA, Zaitlen NA, Talkowski ME, Roeder K, State MW, Devlin B, Sanders SJ, Sestan N (2020) Whole-genome and rna sequencing reveal variation and transcriptomic coordination in the developing human prefrontal cortex. Cell Rep 31(1):107489

van der Wijst MGP, Brugge H, de Vries DH, Deelen P, Swertz MA, Franke L, LifeLines Cohort S, Consortium B (2018) Single-cell rna sequencing identifies celltype-specific cis-eqtls and co-expression qtls. Nature Genetics 50(4):493–497

Wilks C, Zheng SC, Chen FY, Charles R, Solomon B, Ling JP, Imada EL, Zhang D, Joseph L, Leek JT, Jaffe AE, Nellore A, Collado-Torres L, Hansen KD, Langmead B (2021) recount3: summaries and queries for large-scale rna-seq expression and splicing. Genome Biol 22(1):323

Woodward AA, Urbanowicz RJ, Naj AC (2022) Genetic heterogeneity: Challenges, impacts, and methods through an associative lens. Genetic Epidemiology 46(6):595–611

Wray NR, Ripke S, Mattheisen M, Trzaskowski M, Byrne EM, Abdellaoui A, Adams MJ, Agerbo E, Air TM, Andlauer TMF, Bacanu SA, Bækvad-Hansen M, Beekman AFT, Bigdeli TB, Binder EB, Blackwood DRH, Bryois J, Buttenschøn HN, Bybjerg-Grauholm J, Cai N, Castelao E, Christensen JH, Clarke TK, Coleman JIR, Colodro-Conde L, Couvy-Duchesne B, Craddock N, Crawford GE, Crowley CA, Dashti HS, Davies G, Deary IJ, Degenhardt F, Derks EM, Direk N, Dolan CV, Dunn EC, Eley TC, Eriksson N, Escott-Price V, Kiadeh FHF, Finucane HK, Forstner AJ, Frank J, Gaspar HA, Gill M, Giusti-Rodríguez P, Goes FS, Gordon SD, Grove J, Hall LS, Hannon E, Hansen CS, Hansen TF, Herms S, Hickie IB, Hoffmann P, Homuth G, Horn C, Hottenga JJ, Hougaard DM, Hu M, Hyde CL, Ising M, Jansen R, Jin F, Jorgenson E, Knowles JA, Kohane IS, Kraft

J, Kretzschmar WW, Krogh J, Kutalik Z, Lane JM, Li Y, Li Y, Lind PA, Liu X, Lu L, MacIntyre DJ, MacKinnon DF, Maier RM, Maier W, Marchini J, Mbarek H, McGrath P, McGuffin P, Medland SE, Mehta D, Middeldorp CM, Mihailov E, Milaneschi Y, Milani L, Mill J, Mondimore FM, Montgomery GW, Mostafavi S, Mullins N, Nauck M, Ng B, et al. (2018) Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nature Genetics 50(5):668–681

Yao C, Joehanes R, Johnson AD, Huan T, Liu C, Freedman JE, Munson PJ, Hill DE, Vidal M, Levy D (2017) Dynamic role of trans regulation of gene expression in relation to complex traits. Am J Hum Genet 100(4):571–580

Yao DW, O'Connor LJ, Price AL, Gusev A (2020) Quantifying genetic effects on disease mediated by assayed gene expression levels. Nat Genet 52(6):626–633

Yarkoni T, Westfall J (2017) Choosing prediction over explanation in psychology: Lessons from machine learning. Perspectives on Psychological Science 12(6):1100–1122

Yuan Z, Zhu H, Zeng P, Yang C, Li X, Li Z, Yu W, Liu J, Sun S, Yang C, He D, Yang J, Wang X, Wang H, Xia K, Liu J, Xu H, Liu C, Sun L, Jia Z, Shi J, Yang J, Wu Y, Qian W (2020) Testing and controlling for horizontal pleiotropy with probabilistic mendelian randomization in transcriptome-wide association studies. Nature Communications 11(1):3861

Zandi PP, Jaffe AE, Goes FS, Burke EE, Collado-Torres L, Huuki-Myers L, Seyedian A, Lin Y, Seifuddin F, Pirooznia M, Ross CA, Kleinman JE, Weinberger DR, Hyde TM (2022) Amygdala and anterior cingulate transcriptomes from individuals with bipolar disorder reveal downregulated neuroimmune and synaptic pathways. Nat Neurosci 25(3):381–389

Zaykin DV (2011) Optimally weighted z-test is a powerful method for combining probabilities in meta-analysis. J Evol Biol 24(8):1836–41

Zeng B, Bendl J, Kosoy R, Fullard JF, Hoffman GE, Roussos P (2022) Multi-ancestry eqtl

meta-analysis of human brain identifies candidate causal variants for brain-related traits. Nature Genetics 54(2):161–169

Zhang W, Voloudakis G, Rajagopal VM, Readhead B, Dudley JT, Schadt EE, Björkegren JLM, Kim Y, Fullard JF, Hoffman GE, Roussos P (2019) Integrative transcriptome imputation reveals tissue-specific and shared biological mechanisms mediating susceptibility to complex traits. Nat Commun 10(1):3834

Zhang Y, et al (2019) Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. Nature 572:576–580

Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, Montgomery GW, Goddard ME, Wray NR, Visscher PM, Yang J (2016) Integration of summary data from gwas and eqtl studies predicts complex trait gene targets. Nat Genet 48(5):481–7

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B: Statistical Methodology 67(2):301–320