Specific Aims

Polygenic signatures of psychiatric drug response

Patient response to prescription medication is often highly unpredictable, ranging from total remission to life-threatening side effects. Response to treatment depends on a variety of genetic, environmental, and lifestyle factors that are often beyond the view of the prescribing physician. Pharmacogenomics is a research field dedicated to uncovering the genetic basis of inter-individual variability in drug response, with the aim of providing personalized interventions. However, currently available pharmacogenomic services are too narrow in their ascertainment of genetic variation to make robust personalized drug recommendations. There is an emphasis on metabolism and known drug targets and mechanisms of action that leads to blind spots relating to both off-target effects as well as additional therapeutic mechanisms. A more holistic integration of genomic information could provide recommendations that maximize therapeutic effect while minimizing side effects.

We propose to address this critical barrier in pharmacogenomics by integrating whole-genome genotype information, expression quantitative trait loci (eQTL), genome-wide association studies (GWAS) of psychiatric traits, and transcriptional perturbation assays that use small molecules as perturbagens. The triangulation of these disparate data sources will allow us to build models with reduced bias, and these models will be able to recommend drugs on an individual basis that "normalize" disease-associated transcriptional signatures. We will also develop new tools, in the form of a new word embedding space, that capture the subjective experience of taking psychoactive medication.

With my strong data analysis and programming background, and under the mentorship of experienced computational biologist Dr. Jacob Michaelson, I am well-positioned to carry out the proposed research on polygenic predictors of drug response.

Our proposed approach will integrate genetic, brain gene expression, behavioral, and subjective data to provide a more comprehensive understanding of drug response and improve personalized drug recommendations. We propose to accomplish this overall goal through the following specific aims:

Aim 1: To enable **personalized drug recommendation**, we will integrate individual-level genotype data with polygenic scores and drug perturbation signatures. We will recommend drugs based on their tendency to "normalize" the disease-associated components of an individual's imputed (using genotype and eQTL) gene expression signature. The recommendation will be based on a consensus of 1) normalizing the individual's gene expression signature while also 2) normalizing the overall gene expression signature associated with the psychiatric trait of interest.

Aim 2: To **identify spatial patterns in susceptibility** to small molecules in the brain, we will integrate high-resolution brain gene expression data and trait maps with drug perturbation signatures. The proposed brain susceptibility map will link specific compounds to their anticipated phenotypic effects based on publicly available meta-analytic brain-trait maps. The map will be used to further prioritize (or de-prioritize) drug recommendations based on anticipated effects in specific brain regions.

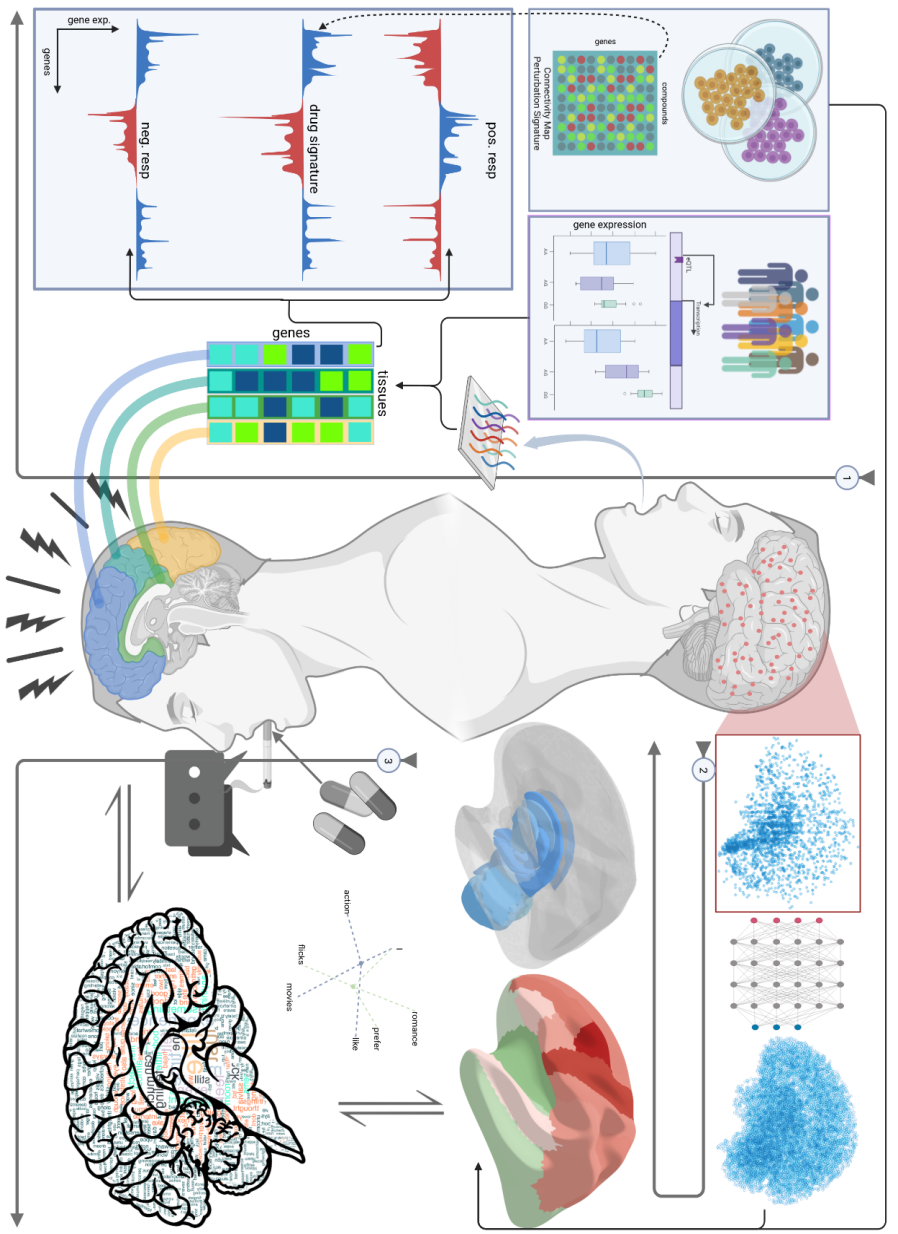
Aim 3: To systematically **characterize the subjective experience** of taking specific psychoactive compounds, we will develop a word embedding space derived from first-person accounts of experience with selected drugs. The proposed approach will identify recurrent patterns and trends

in drug-induced effects on the mind, which can be difficult to estimate using objective measures alone.

Together, the successful achievement of these independent aims will provide new tools to better personalize drug recommendation and prescription. Our preliminary data suggest that individual genotype is a powerful tool in this effort and can provide promising results for drug recommendation.

Table of Figures

Table of Contents



Commented [EMSIM1]: Needs to be updated with the new one. Smaller gap between figures in aim one. More saturated colors in eQTL section. Different background for CMap sig.

1. Introduction

1.1. Background and Significance

Genetics has been one of the most growing science fields, and yet we still know too little about diseases mechanisms or be able to provide proper treatments. Individual's genetics has been known to be a major player in differentiating response to different pharmacological compounds. The field of pharmacogenomics is rapidly growing, with the aim of providing personalized interventions for patients by uncovering the genetic basis of inter-individual variability in drug response. However, currently available pharmacogenomic services are limited in their ascertainment of genetic variation, with an emphasis on metabolism and known drug targets and mechanisms of action. This approach may result in blind spots relating to off-target effects as well as additional therapeutic mechanisms. A more comprehensive understanding of drug response could provide recommendations that maximize therapeutic effect while minimizing side effects.

Commented [EMSIM2]: Ew

Mental disorders are a major challenge for individuals and society. Conditions such as schizophrenia, major depression, and anxiety disorders require long-term treatment with psychoactive drugs. Although there have been more than two-hundred drugs developed in the last six decades, they still can have variable effects between patients due to differences in drug metabolism and action. Consequently, increasing the dosage of medication does not necessarily lead to better treatment outcomes. Early treatment outcomes are frequently poor, with 30 50% of patients failing first-line antidepressant medication due to inefficiency or intolerance, according to estimates (Barak et al., 2011). Moreover, about 25,000 people each year in the United States visit emergency rooms as a result of side effects brought on by antidepressants (Hampton et al., 2014). Before identifying a medication that reduces depressive symptoms with few side effects, patients frequently try with a variety of antidepressant regimens. The psychological and societal costs of repeatedly taking drugs that "do not work" can be distressing for the individual and highlight the need for better drug selection and dosage tactics because antidepressant pharmacotherapy studies frequently take a minimum of 6-8 weeks.

In particular, there is a critical need for personalized drug recommendations for patients with psychiatric disorders. These patients often have complex medication regimens and can experience a wide range of responses to treatment, from total remission to life-threatening side effects. By integrating whole-genome genotype information, expression quantitative trait loci (eQTL), Polygenic Scores (PGS) of psychiatric traits, and transcriptional perturbation assays, it may be possible to develop models with reduced bias that can recommend drugs on an individual basis.

This research proposal aims to address the critical need for personalized drug recommendations for patients with psychiatric disorders by integrating disparate data sources to develop more comprehensive models of drug response. The proposed

research builds on previous studies that have identified genetic and transcriptional signatures associated with psychiatric disorders, as well as studies that have investigated the effects of small molecules on gene expression. By integrating these data sources, we aim to develop a more comprehensive understanding of drug response that takes into account the complex genetic and environmental factors that contribute to inter-individual variability.

The significance of this research lies in its potential to improve patient outcomes by providing personalized drug recommendations that maximize therapeutic effect while minimizing side effects. By developing more comprehensive models of drug response that consider the complex genetic and environmental factors that contribute to inter-individual variability, we hope to provide clinicians with a powerful tool for optimizing patient care. Ultimately, the proposed research could lead to improved treatment outcomes and a better quality of life for patients with psychiatric disorders.

1.2. Objectives of the study

The objectives of this study are to:

- Integrate whole-genome genotype information, eQTL, GWAS, and transcriptional perturbation assays to recommend drugs on an individual basis that "normalize" disease-associated transcriptional signatures.
- Integrate high-resolution brain gene expression data and trait maps with drug perturbation signatures to identify spatial patterns in susceptibility to small molecules in the brain and use them to further prioritize or de-prioritize drug recommendations based on anticipated effects in specific brain regions.
- Develop a new word embedding space that captures the subjective experience of taking psychoactive medication and uses it to improve personalized drug recommendations.

2. Literature Review

2.1. overview of pharmacogenomics and personalized medicine

Pharmacogenomics, also known as pharmacogenetics, is the branch of science that looks at how a person's genes influence how they react to pharmaceuticals. Its long-term objective is to assist physicians in choosing the medications and dosages that are ideal for every patient. It falls under the category of precision medicine, which tries to treat every patient uniquely.

Pharmacogenomics, a rapidly developing field in personalized medicine, aims to elucidate how variations in genes can affect a patient's response to drugs. The influence of genes on drug metabolism and efficacy is well-established, as they encode for enzymes and proteins responsible for the breakdown and uptake of medications in the body.

Of particular interest are genes that encode for enzymes involved in drug metabolism, such as CYP2D6, which acts on a quarter of all prescription drugs. Multiple variations of this gene exist, with some individuals having multiple copies of it. These genetic variations can result in differences in enzyme activity, with some variants leading to a hyperactive enzyme that metabolizes drugs at a faster rate than normal (Ingelman-Sundberg, 2005). This can result in drug overdose, particularly in the case of codeine, which is metabolized by CYP2D6 to produce its active form, morphine. Conversely, some variants of CYP2D6 produce an enzyme that is non-functional or less active, leading to reduced or absent drug efficacy.

Therefore, understanding the impact of genetic variations on drug response is crucial in ensuring safe and effective drug therapy. Pharmacogenomics provides a valuable tool for predicting drug response based on an individual's genetic makeup, enabling personalized medicine approaches for improved patient outcomes.

The FDA has released comments and warning letters on pharmacogenetic testing in response to concerns over the marketing of these tests. The efficacy of clinical pharmacogenetic testing may not be fully supported by clinical data, according to a safety communication released on November 01, 2018. The statement in this safety communication that "the relationship between DNA variations and the effectiveness of antidepressant medication has never been established" particularly emphasized the use of pharmacogenetic testing to guide antidepressant drug prescribing (Shuren, 2018). In keeping with the FDA goal to safeguard and advance the public's health, it's critical to act right away to make sure that the claims being made about the pharmacogenetic tests currently available are supported by reliable research. That can be done by taking measures that safeguard patients while also advancing the creation of analytically and clinically validated pharmacogenetic tests. Recently, the FDA released a new web-based resource that includes a table including some of the pharmacogenetic associations with a last update on October 26, 2022 (FDA, 2022). Some of these have detailed information regarding therapeutic management, but the majority of the associations listed have not been assessed in terms of the effect of genetic testing on clinical outcomes, such as improved therapeutic effectiveness or increased risk of particular adverse events. This version of the table is restricted to pharmacogenetic associations linked to drug transporter, drug metabolizing enzyme, and gene variations associated with a susceptibility for certain adverse outcomes.

Pharmacogenomic testing could be done in two forms: single-gene or a multi-panel testing. Most of the developed pharmacogenetic tests investigate the main drug gene targets, and the advanced ones with a multigene panel investigate different genes that include the ones involved in the process of metabolizing the drug. Yet, none of these tests consider a whole-genome approach, which is one of our main focuses in the proposed study. one of the main advantages of the proposed approach is that we are focusing on the entire genome, not only the genes involved in a drug response or metabolism. The FDA now advises against using direct-to-consumer testing for making

medical choices, even though there has been considerable debate about the use of pharmacogenomics in clinical practice. However, the FDA allowed the marketing of the 23andMe Personal Genome Service Pharmacogenetic Reports test as a direct-to-consumer test with special controls for informing discussions with a healthcare professional about genetic variants that may be related to a patient's capacity to metabolize some medications (FDA, 2018).

2.2. Gene expression, eQTL, and GWAS related to psychiatric disorders:

In recent years, a new area of research has emerged: the investigation of the genetic basis of psychiatric traits. The use of eQTLs, GWAS, and gene expression data is one strategy that has gained interest in this field. Genomic regions known as eQTLs are linked to variations in gene expression levels. Finding eQTLs allows researchers to uncover genetic variations that may have an effect on the transcriptome functionally, perhaps shedding light on the molecular processes behind mental illnesses. For instance, eQTLs for numerous genes, including ITIH4, GLT8D1, GNL3, and NEK4, were discovered to be enriched in schizophrenia-related genetic regions (Kim et al., 2014). This shows that these genes could be involved in the disorder's genesis. eQTLs were also studied for their effects in the developing human brain and their enrichment in neuropsychiatric disorders (Bryois et al., 2022; O'Brien et al., 2018).

The use of GWAS to investigate the genetic foundation of psychiatric characteristics has also become widespread. In these studies, the complete genome is examined in sizable patient cohorts and healthy controls, revealing genetic variations linked to certain disorders. For instance, a recent meta-analysis of three major genome-wide association study (GWAS) of major depressive disorder (MDD) discovered 102 independent variants, 15 gene-sets, and 269 genes to be linked to the condition, many of which are involved in synaptic neurotransmission and synaptic structure (Howard et al., 2019). The functional implications of these polymorphisms are not revealed by GWAS, even though they serve as a useful starting point for the identification of genetic risk factors. The value of these association analysis techniques is increased when evidence of biological pathway enrichment is combined with data on features associated with the trait being studied. This allows for further conclusions regarding similar aetiological processes.

The analysis of different genome variants in the context of their effect on gene has spawned a big field in genetics named as expression quantitative trait loci (eQTLs). An eQTL is a locus that explains a portion of a gene expression phenotype's genetic variation. In a standard eQTL experiment, gene expression levels are often assessed in tens or hundreds of people and genetic variation markers are directly tested for associations. This association study can be performed close to the gene (cis) or far (trans) from it. The *cis*-eQTLs variations are those that are within 1 Mb (megabase) on either side of a gene's transcription start site (TSS). On the other hand, *trans*-eQTLs are those that are at least 5 Mb upstream, downstream, or on a separate chromosome. The same regulatory region or variant can be identified as an eQTL for different genes in different

tissues, suggesting that tissue specificity is a highly important factor to be considered in such analyses.

Commented [EMSIM3]: Add the new figure with Lowe transparency and black outlier points. Also change font size for A&B to be smaller :3

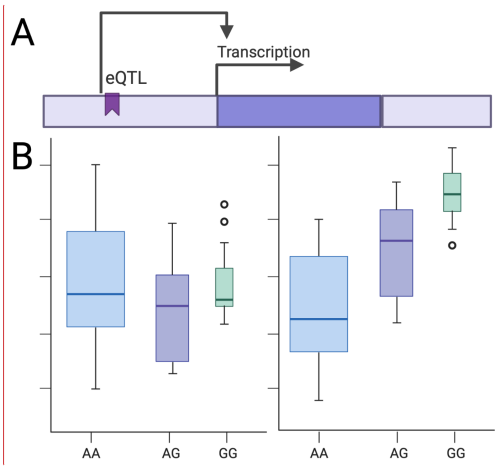


Figure 1 (A) shows a schematic representation of the effect of an eQTL on gene expression. Boxplots in (B) show correlation between genotype variants and gene expression in different populations. The width of a boxplot is proportional to the allele frequency.

2.3. Drug repositioning

Drug repurposing initiatives may benefit from the application of GWAS and eQTL analyses. eQTL analysis can show the functional impact of these variations on gene expression, and GWAS can find genetic variants linked to a specific disease or health issue. Researchers can find medications that may be beneficial for treating diseases other than their intended targets by merging these datasets with drug databases.

A recent article reviewed different methods for drug repurposing that included 5 main approaches utilizing different types of datasets, shown in Figure 1 (Lau & So, 2020). In the first approach "candidate gene approach", which uses functional annotations and eQTL data, the risk loci from GWAS data may be mapped to the genes that are most likely to be important. Drugs may be used as repositioning candidates if the discovered candidate gene is druggable and the medication is not currently prescribed for the condition. Another approach, called "pathway or gene-set analysis approach", states that drugs that target members of the same pathway are prospective drug candidates. The pathways that the selected candidate gene(s) are situated in are next examined. In addition, enrichment tests may be used to find medications whose targets or effector genes obtain higher significance (lower p-values) than anticipated overall. The whole collection of GWAS data may also be utilized to produce gene-based statistics. The third approach is mainly focusing on comparing similarities between drugs and diseases (dx). In that approach, if two medications' indications are sufficiently

similar to one another, they can be relocated. For instance, utilizing cell-line expression data from the Connectivity Map (CMap), one may assess the similarities between the transcriptomes of two medications. According to this, if two diseases are similar, then the medications used to treat one condition may be used to treat the other. The fourth approach is considered by looking for reversed expression patterns between drugs and diseases. The key premise is that a medication may be a candidate for repositioning if it results in an expression profile that is the opposite of that of a disease (owing to its ability to "reverse" expression patterns linked to diseases). The fifth, and last approach, is a network-based analysis approach. The method is based on building biological networks by integrating data from several sources, including interactions between drugs, proteins, genes, and diseases. The idea behind this approach is similar to that of "similarity-based" methods for medication repositioning, however network-based approaches often use a wider range of data.

Commented [EMSIM4]: I don't like the empty space here.

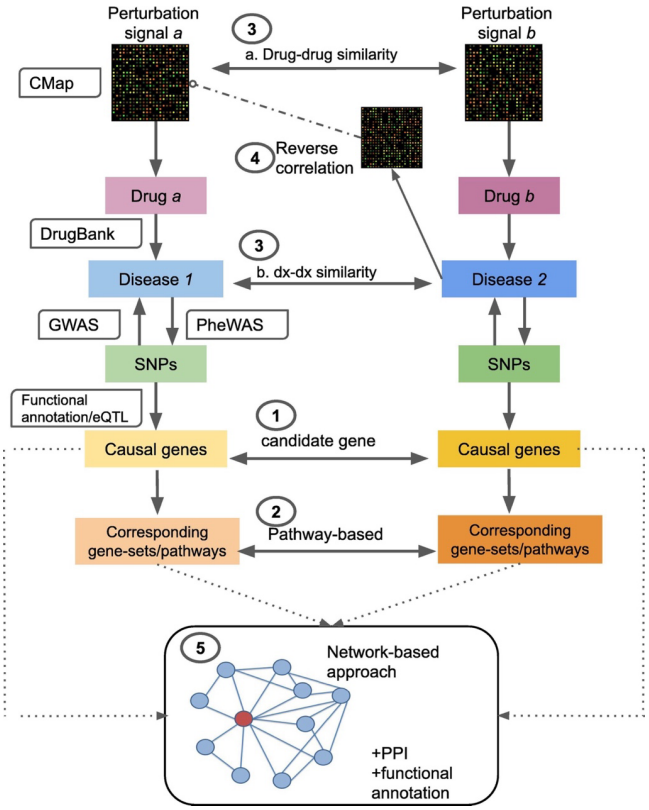


Figure 2: A schematic diagram of drug repurposing approaches using different types of datasets reviewed by (Lau & So, 2020).

2.4. Psychoactive compounds and subjective experience

Psychoactive compounds are those that work on the central nervous system that impact a person's perception, behavior, and mood. Although psychoactive substances have been used for ages for religious, therapeutic, and recreational purposes, their usage can also have unfavorable effects including addiction, psychosis, and other psychological issues. Understanding the subjective experience of consuming psychoactive substances is crucial for producing safer and more potent psychoactive substances as well as for improving the treatment results for psychiatric disorders. The effects of psychoactive substances on the brain and body have been investigated using unbiased techniques including physiological monitoring and brain imaging. These metrics, meanwhile, fall short of properly capturing the user's subjective experience. This is due to the fact that the subjective experience of ingesting psychoactive substances is a complicated phenomenon with many facets that includes a variety of cognitive, emotional, and perceptual processes. As a result, to fully comprehend the subjective experience of ingesting psychoactive substances, it is necessary to employ additional techniques that can adequately capture the subtleties of the effects that drugs have on the brain.

The study of first-person narratives of drug experiences is a potential strategy for comprehending the subjective experience of taking psychoactive substances. We can better grasp the variety of effects that these substances can have as well as the variations in responses across individuals by examining first-person accounts. It can be difficult to analyze unstructured data, such as first-person experiences, thus it's critical to create effective strategies for drawing out useful information from these experiences. Word embedding techniques have been utilized more often recently to analyze unstructured text data and uncover significant trends. A form of natural language processing (NLP) method known as word embeddings may represent words as high-dimensional vectors that capture their semantic and contextual links. A word embedding space on first-person reports of drug experiences can be developed to represents the subjective experience of taking certain psychoactive substances. This method can uncover recurring patterns and trends in the mental side effects of drugs, which can offer insightful information on the subjective experience of taking psychoactive substances. By doing this, we want to improve our comprehension of the intricate and varied phenomena known as the subjective experience of taking psychoactive substances.

3. Aims Overview/Methods

3.1. Aim 1:

We propose to address the critical barrier in pharmacogenomics by integrating tissue-specific gene expression data and transcriptional perturbation assays that use small molecules as perturbagens. The combination of these data sources will allow us to build a model with reduced bias, which will be able to recommend drugs on an individual basis that "normalize" disease-associated transcriptional signatures. For example, if a person is diagnosed with depression and depression caused elevation of gene X

expression, the best drug to be working well with this patient is the one that can neutralize the disease effect on gene expression and bring it back to normal level. Our overall objective in this work is to maximize efficacy of drugs on an individual level, while reducing the negative side effects.

3.1.1. eQTL weights and transcriptome imputation

To identify how different variants on the genome affect gene expression by tissue, the eQTL data from the Genotype-Tissue Expression (GTEx) project will be used to give weights for every variant by gene. The weights are mainly derived from a linear regression model that fits a gene expression in a specific tissue by the known cis variants, close by a 1 Mb distance from the TSS. The linear regression model comes in the form of:

$$Y = B0 + B1\,X1 + B2\,X2 + \ldots + Bn\,Xn \quad \textit{Equation 1}$$

where B1:Bn are the effect sizes that each SNP X has on that gene, and n is the total number of cis-eQTLs of the given gene. By the end of this step, we will have a weights matrix of MxN for every single tissue, where N is the number of genes, and M is the number of SNPs known to have weights for any gene in that tissue. If the SNP is not affecting the gene's expression, the value will be 0. Otherwise, it will have the weight derived from the linear model.

The step of identifying the effect sizes of *cis*-eQTLs per gene in every tissue from GTEx was done different groups of researchers with different approaches (Chen et al., 2023; de Klein et al., 2023; Gamazon et al., 2015; Hu et al., 2019; Liu & Kang, 2022). The model weights from UTMOST by (Hu et al., 2019) will be used for transcriptome imputation in this project. The workflow of getting model weights is shown in the schematic figure below.

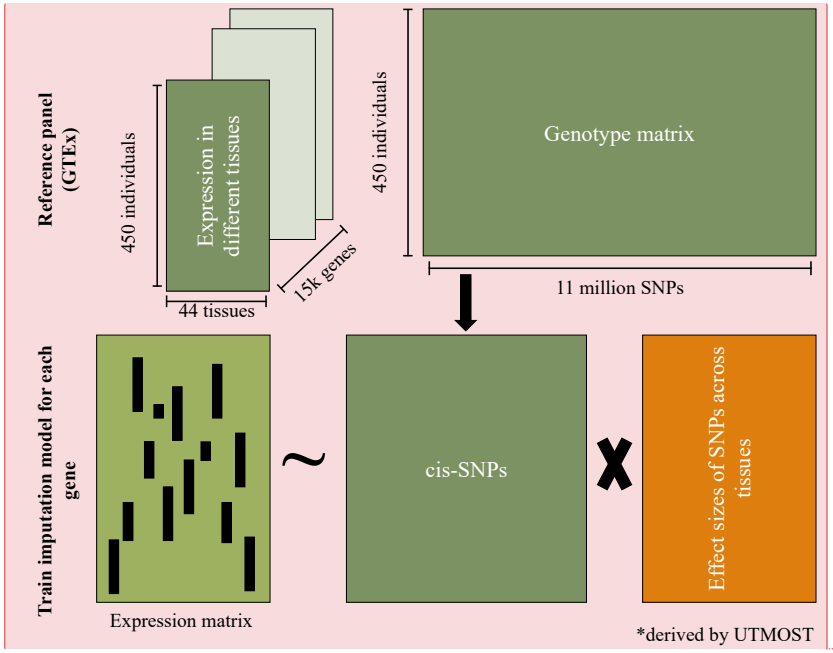


Figure 3 workflow of calculating eQTLs effects on gene expression.

To impute individuals' gene expression in different tissues, a simple matrix multiplication will be done following this formula: (shown in the figure below)

$$Imputed\_tx = Genotypes\_Matrix \times Tissue\_Weights\_Matrix \qquad Equation\ 2$$

The *Tissue_Weights_Matrix* represents the weights matrix, MxN, derived from equation (1). On the other hand, the *Genotypes_Matrix* has rows as individuals, and columns as genotypes. The values in this matrix represent the genotype alleles type found per individuals (i.e., 0 if both alleles are homozygous reference, 1 if both alleles are heterozygous, and 2 if both alleles are homozygous alternative). The *Genotypes_Matrix* used in this project is for participants from the Adolescent Brain Cognitive Development (ABCD) study. The ABCD study is one of the largest long-term NIH-funded study of the brain development and child health in the United States. The study includes ~11,000 children of ages 9-10 years. The study tracks their biological and behavioral development through adolescence into young adulthood.

Commented [EMSIM5]: The colors are fine, don't waste time in changing these. LET IT GO

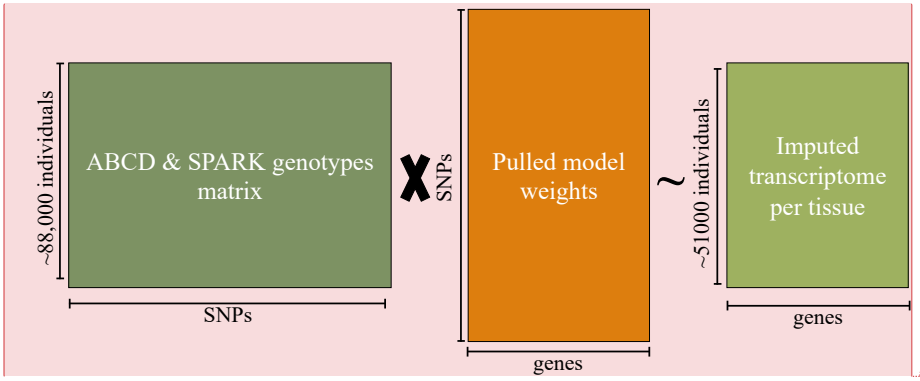


Figure 4 transcriptome imputation illustration

3.1.2. Chemical perturbagens signature

A drug perturbagen signature from the BROAD Institute, called connectivity map (CMap)[1], was developed as a measure of cellular signature for different perturbagens in different cell lines. As of April 7th, 2023, the CMap includes over 1.5M gene expression profiles from ~5,000 small-molecule compounds, and ~3,000 genetic reagents, tested in multiple cell types. A computational pipeline processes measured gene expression data after introducing the chemical compound to a cell line to create signatures from the raw fluorescence intensity.

The signature matrix per drug is a 1xG matrix, where G is the total number of genes with gene expression measured after introducing the drug to the cell line. The values for gene expression values are in a 0-1 range, as a measure of fluorescence intensity. The raw data of the CMap is measured for different doses of drugs with different cell lines. The data I'm using is a fitted data after controlling for the cell line and the dose of the drug, so the signature is unique for the drug effect and not the dose or the cell line. This data is already available and processed by XX. Assuming we have a matrix for all drugs that has X rows corresponding to number of genes, and Y columns corresponding to different doses measured in different cell lines of different drugs (e.g. Venlafaxine_JURKAT_10uM, Venlafaxine_JURKAT_3.33uM, Venlafaxine_THP1_0.25uM, Venlafaxine_THP1_0.08uM, Sertraline_JURKAT_10uM, Sertraline_JURKAT_3.33uM, Sertraline_THP1_0.25uM, and Sertraline_THP1_0.08uM). From that matrix, it can be converted to a tidy format where it only has 5 columns: gene symbol, gene expression, drug, cell line, and dose. To get the effect of a drug on gene expression, we can fit a linear regression model for each gene with the formula:

$$\text{gene\_expression} \sim \text{drug} + \text{cell\_line} + \text{dose} \qquad \textit{Equation 3}$$

[1] https://www.broadinstitute.org/connectivity-map-cmap

Commented [EMSIM6]: It's bothering me, but don't

Commented [EMSIM7]: Get ref here for CMap data that Jake downloaded

From that, we can only keep the *beta-estimate* of the drug variable from the formula, which corresponds to the drug effect on a gene expression after correcting for cell line and dose.

3.1.3. Drug response prediction

To determine individuals' response to different drugs, we will compute the correlation between individual's imputed transcriptome in a brain tissue, and the drug signature (i.e. measured gene expression after introducing the drug to different cell lines). The drug signature data was retrieved from the NIH LINCS program website[2]. A high or low correlation is predictive of how much change the drug is making to the gene expression and in what direction it is in-relation to the disease signature. The correlations calculated in this step are a measure of how different an individual's imputed gene expression than the drug signature. To have a standardized measure of correlation, everyone's predicted correlation will be referenced to correlations measure of European samples from the 1000-genome project. In other words, the final correlation value is a measure of the individual's predicted correlation to the drug and also a measure of how they lie in reference to other individuals of the same population.

After referencing the correlations to the 1000-genomes drug correlations, the individuals will be categorized as positive responder to the drug, if their imputed tissue expression has a negative correlation with the drug perturbation signature, or negative responders if they have a positive correlation. The negative correlation between a drug's signature and an imputed transcriptome means that the drug is expected to reverse the imputed transcriptome, and that will be normalizing the gene expression (i.e., the drug is working by reversing the present state of gene expression).

Commented [EMSIM8]: Change the figure with the new one. Smaller gap between them!

[2] https://lincsproject.org

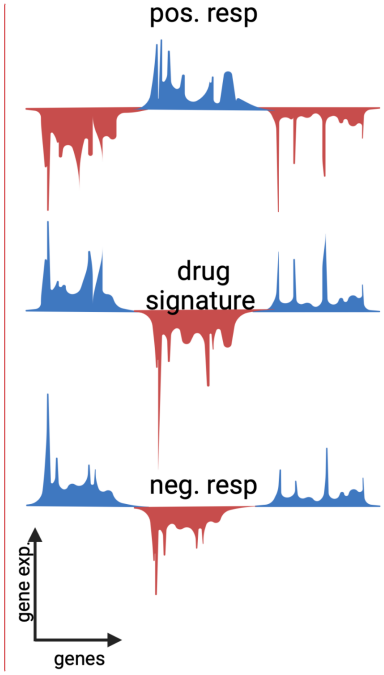


Figure 5 drug response prediction based on CMap drug signature.

3.1.4. Expected results

Commented [EMSIM9]: Probably need to add more? Or discuss?

At timepoints of taking a drug, a positive responder is expected to show less disease symptoms that are expected to be treated by the drug in question. On the other hand, a negative responder is expected to show little enhancement or worse symptoms. We are expecting this approach to be useful in a clinical setting, where a provider has multiple options of treatment for the same symptoms and cannot decide on which one to start with.

3.2. Aim 2:

Drug development is a complex and challenging process that requires a deep understanding of the biological mechanisms underlying drug actions. One of the critical aspects of drug development is the identification of the regions in the brain that are activated by a drug, either for therapeutic or side effects. In this section of the project, we aim to develop a brain activity map that shows the activity score of different brain

regions in response to different drugs. My approach is based on the integration of data from the Allen Institute[3] and the LINCS project[4] using a deep learning model.

The Allen Institute provides a comprehensive dataset of measured gene expression in different brain regions, along with their spatial coordinates in a brain MNI-space. On the other hand, the LINCS project offers a vast collection of drug gene expression signatures that describe the molecular effects of drugs on various cell lines, described in *Aim 1* methods. The integration of these datasets using a deep learning model can help us identify the regions in the brain that are activated by a specific drug.

3.2.1. Allen Institute data

The Allen Institute provides a unique multimodal atlas of the human brain, integrating anatomic and genomic information. We will be mainly using two types of the provided datasets: microarray and MRI data. The microarray dataset provides an "all genes, all structures" survey in multiple adult control brains. It includes measured gene expression using > 62,000 gene probes per profile and ~ 500 samples per hemisphere across cerebrum, cerebellum, and brainstem. The same data was mapped with histology into unified 3-D anatomic framework based on MRI. Six donors have both MRI and microarray data. The table below shows the donors metadata:

Table 1 Donors' metadata from the Allen Institute

Donor	**Age (yrs)**	**Sex**	**Ethnicity**	**PMI**
H0351.1009	57	M	White or Caucasian	26
H0351.1012	31	M	White or Caucasian	17
H0351.1015	49	F	Hispanic	30
H0351.1016	55	M	White or Caucasian	18
H0351.2001	24	M	Black or African American	23
H0351.2002	39	M	Black or African American	10

Commented [EMSIM10]: Recheck these donor IDs again to make sure they're the ones with microarray data used.

After downloading the gene expression data, all gene expression values from probes that map to the same gene will be averaged per sample. Along with the gene expression data, the sample positions were mapped to the MNI-space. All samples collected from six donors are shown in the figure below, based on their spatial xyz MNI-space position.

[3] https://alleninstitute.org
[4] https://lincsproject.org

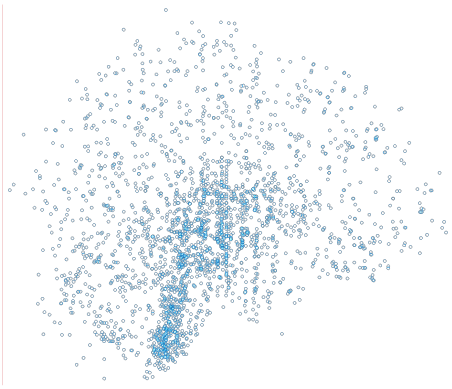


Figure 6 spatial representation of collected transcriptomic samples by the Allen Institute

Commented [EMSIM11]: I don't like the point size here. Whether:
- ignore it
- make it bigger (use the one from below)
- delete the figure and refer to it in the next figure.

3.2.2. Deep learning model for predicting gene expression

Commented [EMSIM12]: Might need to talk about benchmarking and doing a gem or other methods? idk

Although the Allen Institute gene expression data covers many different regions of the brain and considered to be a high-quality data, yet we still don't know the gene expression in other regions that were not measured. We propose to build a deep learning (DL) model that takes xyz MNI-positions as an input and predicts the associated gene expression with these positions. A schematic workflow of this step is shown below.

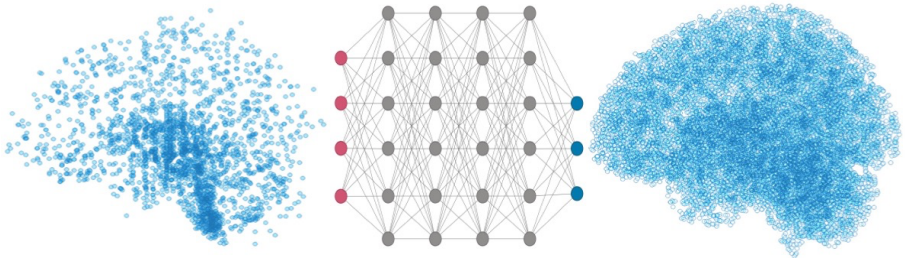


Figure 7 schematic workflow of deep learning model to predict gene expression.

After building the model, we expect to have a vector of predicted gene expression for every voxel position in a brain MNI-space.

3.2.3. Drug activity prediction

After predicting gene expression in different brain positions, they will be correlated to a drug signature as a measure of activity similarity. A positive correlation in this case means the drug is able to give a transcriptomic profile similar to the region's activity profile. It's hypothesized that a positive correlation also means the drug should activate these regions, if taken. An example of predicted activity map of methylphenidate is shown below.

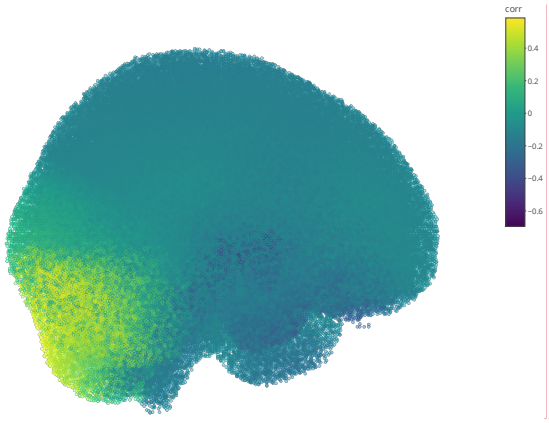


Figure 8 predicted brain activity map of a taking methylphenidate.

Commented [EMSIM13]: Shouldn't I add the correlation color bar here? Figure out what's wrong with the package function

3.2.4. Brain functionality map

Functional magnetic resonance imaging (fMRI) data may be synthesized on a large scale and automatically using the Neurosynth platform[5]. It analyzes thousands of publications that have been published and summarize the findings of fMRI investigations, breaks them up, and then gives out visuals of a brain with colored intensities representing activity. Up to April 11th, 2023, there are 507891 activations reported in 14371 studies, and an interactive, downloadable meta-analyses of 1334 terms. The term meta-analyses pipeline provides activation coordinates for each term of interest (e.g., 'emotion', 'language', etc.). An automated parser is used to extract activation coordinates from neuroimaging articles that have been published. The parser goes through the full text of each article and identifies a set of frequently occurring terms, which are then compiled into a list of several thousand terms that appear in 20 or more studies. To analyze a specific term, the database of coordinates is separated into two groups: one containing articles with the term of interest and the other without. A meta-analysis is then conducted, comparing the reported coordinates for each group. This analysis produces statistical inference maps, including z and p value maps, as well as posterior probability maps, which illustrate the likelihood of a particular term being used in a study if activation is observed at a specific voxel.

With the help of the Neurosynth Image Decoder, we may quantitatively and interactively compare any brain volume in Nifti format to a few chosen images from the Neurosynth database. For this purpose, we can use the brain activity map of a drug that was produced by our pipeline and get the output of Neurosynth Image Decoder. The output should show terms' correlations to the uploaded Nifti image intensities. The

Commented [EMSIM14]: I wish I could give it an abbreviation of NID

[5] https://neurosynth.org

intensities in the uploaded image are a representation of the activity value. A sample of Neurosynth Image Decoder output is shown below.

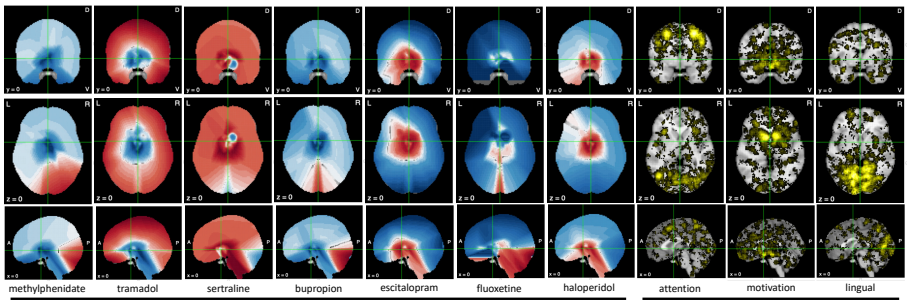


Figure 9 Predicted activity map for different drugs (left) and terms positive activity map from Neurosynth database (right).

As shown in the figure above, Neurosynth compares the uploaded nifti map (left) to different maps for terms (right) and provides computed correlation between them. The terms map can provide a foundation for future research on the mechanisms underlying drug actions in the brain. It also can help in identifying and predicting potential side effects of drugs.

3.2.5. Web-based application development

Commented [EMSIM15]: Give it a name! Hehe

The end goal of this aim is to develop a web-based application to enable researchers and clinicians to access the brain activity maps and predict drug effects in real-time. The application should be able to provide activity maps for most FDA-approved drugs_ only the ones with transcriptomic signatures in the LINCS program database_ and the associated terms from Neurosynth Image Decoder. It is also proposed to provide different visualization outputs for the same results to help researchers identify activity in a brain region as a whole, or a specific xyz position in an MNI-space. Shown below is an example of same results with two different visualizing methods.

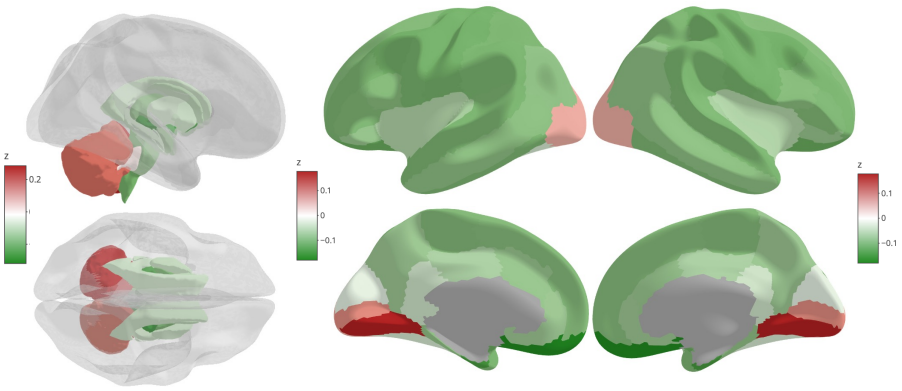


Figure 10 Web-based application visuals for aim 2 results. The map shown is for methylphenidate activity, averaged by anatomical region

Commented [EMSIM16]: Need a longer caption describing the differences? Ugh

3.3. Aim 3:

The subjective effects of psychoactive drugs have been a subject of interest and concern for centuries. The subjective experience of taking these substances is frequently more challenging to record, even if numerous objective measurements, such as physiological and behavioral changes, have been established to research pharmacological effects. In order to better understand how drugs influence the brain and how to enhance therapeutic interventions for people with drug addiction or mental health issues, it is essential to comprehend the subjective experience of psychoactive substances.

3.3.1. Data collection

For data collection, we will be using online discussion forums such as Reddit, which allow individuals to share their experiences and opinions about specific drugs. We will specifically target forums related to drugs commonly used for mental health conditions, such as Venlafaxine, which is used as an antidepressant and anxiety medication. To ensure that we capture a diverse range of experiences, we will use specific keywords and search terms such as "Venlafaxine," "Effexor," and "antidepressant" to locate relevant threads and comments. We will also take care to exclude posts that do not include first-person accounts or are off topic. Once we are able to identify users mentioning they have been on the drug of interest, we will extract their comments for a period of time that includes before taking the drug and after.

Once we have identified relevant threads and comments, we will extract the text data and preprocess it for analysis. This will involve removing any irrelevant or duplicate content, as well as standardizing the text data for consistency (e.g., converting all text to lowercase, removing special characters and punctuation).

3.3.2. Word embedding

The preprocessed data will next be handled by a word embedding method. A natural language processing (NLP) method called word embedding portrays words as high-dimensional vectors in a mathematical space, where each word is connected to a specific set of values that reflect its semantic and contextual significance. By using this method on the text data, we may look at the semantic connections between words and concepts to find patterns and trends in the subjective experience of using Venlafaxine and other chosen medications.

Word embedding is based on the idea that related words will have similar vectors, which may be utilized to understand the relationships and patterns among words. There are limitations to the traditional way of representing words as one-hot vectors, in which each word is represented by a vector with just one non-zero member. One-hot vectors need a lot of processing to process since they are high-dimensional and sparse. They also fail to account for the semantic connections between words. By expressing words as dense vectors in a continuous space, word embedding overcomes these limitations. Large amounts of text data are used to teach these vectors using methods like neural networks. The generated vectors capture both the word co-occurrence patterns and their semantic links in the text.

The word embedding may be used to represent any word as a vector in the high-dimensional space once it has been trained. This enables the comparison of word similarity, the identification of word clusters, and even the application of mathematical operations to words. The word embedding approach will be applied in the context of this research proposal to reflect the language used in first-person reports of experience with certain medications. This will make it possible to spot recurring patterns and trends in the mental side effects of drugs, which can be challenging to gauge using only objective measurements.

We will group similar drug experiences based on their word embeddings using clustering methods like k-means and hierarchical clustering. This will enable us to spot recurring trends and themes in the mental impacts of drugs. To find temporal and contextual changes in drug experiences across time, we will also apply trend analysis approaches including time series analysis and topic modeling.

To validate our approach, we will compare our clustering results with existing taxonomies of drug effects, such as the Altered States Database (ASDB) (Schmidt & Berkemeyer, 2018). We will ensure that all collected data are anonymized and that no personally identifiable information is disclosed. We will also comply with ethical guidelines for research involving data extracted from Reddit.

Proposal Timeline

Commented [EMSIM17]: Lovely optimistic. Can I graduate by 2024? LOL

Table 2 proposed timeline for the project

Aim	Task	2022	2023		2024		2025	
1	get weights of eQTLs							
	impute gene expression for samples							
	predict individuals' drug response							
	validate results using CBCL							
2	build a model for predicting gene expression							
	compute drug activity in brain regions							
	correlate drug activity with brain functionality from neurosynth							
	apply same model on mouse brain and validate							
3	download and preprocess text data of users of interest							
	build a word embedding space for experience reports							
	validate results and reflect on aim 2 brain maps							

points to be mentioned:

- ~~Therapeutic drug monitoring (TDM) has emerged as a promising solution to these challenges, particularly for mood stabilizers, antidepressants, and antipsychotics. TDM has the potential to reduce variability, speed up clinical improvement, and improve drug tolerability and safety.~~
- Link the third aim to the first two aims
- ~~Make a graphical abstract~~
-

References

Barak, Y., Swartz, M., & Baruch, Y. (2011). Venlafaxine or a second SSRI: Switching after treatment failure with an SSRI among depressed inpatients: a retrospective analysis. *Prog Neuropsychopharmacol Biol Psychiatry*, *35*(7), 1744-1747. https://doi.org/10.1016/j.pnpbp.2011.06.007

Bryois, J., Calini, D., Macnair, W., Foo, L., Urich, E., Ortmann, W., Iglesias, V. A., Selvaraj, S., Nutma, E., Marzin, M., Amor, S., Williams, A., Castelo-Branco, G., Menon, V., De Jager, P., & Malhotra, D. (2022). Cell-type-specific cis-eQTLs in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders. *Nat Neurosci*, *25*(8), 1104-1112. https://doi.org/10.1038/s41593-022-01128-z

Chen, F., Wang, X., Jang, S. K., Quach, B. C., Weissenkampen, J. D., Khunsriraksakul, C., Yang, L., Sauteraud, R., Albert, C. M., Allred, N. D. D., Arnett, D. K., Ashley-Koch, A. E., Barnes, K. C., Barr, R. G., Becker, D. M., Bielak, L. F., Bis, J. C., Blangero, J., Boorgula, M. P., . . . Liu, D. J. (2023). Multi-ancestry transcriptome-wide association analyses yield insights into tobacco use biology and drug repurposing. *Nat Genet*, *55*(2), 291-300. https://doi.org/10.1038/s41588-022-01282-x

de Klein, N., Tsai, E. A., Vochteloo, M., Baird, D., Huang, Y., Chen, C. Y., van Dam, S., Oelen, R., Deelen, P., Bakker, O. B., El Garwany, O., Ouyang, Z., Marshall, E. E., Zavodszky, M. I., van Rheenen, W., Bakker, M. K., Veldink, J., Gaunt, T. R., Runz, H., . . . Westra, H. J. (2023). Brain expression quantitative trait locus and network analyses reveal downstream effects and putative drivers for brain-related diseases. *Nat Genet*, *55*(3), 377-388. https://doi.org/10.1038/s41588-023-01300-6

FDA. (2018, October 31). *FDA authorizes first direct-to-consumer test for detecting genetic variants that may be associated with medication metabolism* https://www.fda.gov/news-events/press-announcements/fda-authorizes-first-direct-consumer-test-detecting-genetic-variants-may-be-associated-medication

FDA. (2022). Table of Pharmacogenetic Associations. In.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Consortium, G. T., Nicolae, D. L., Cox, N. J., & Im, H. K. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet*, *47*(9), 1091-1098. https://doi.org/10.1038/ng.3367

Hampton, L. M., Daubresse, M., Chang, H. Y., Alexander, G. C., & Budnitz, D. S. (2014). Emergency department visits by adults for psychiatric medication adverse events. *JAMA Psychiatry*, *71*(9), 1006-1014. https://doi.org/10.1001/jamapsychiatry.2014.436

Howard, D. M., Adams, M. J., Clarke, T. K., Hafferty, J. D., Gibson, J., Shirali, M., Coleman, J. R. I., Hagenaars, S. P., Ward, J., Wigmore, E. M., Alloza, C., Shen,

X., Barbu, M. C., Xu, E. Y., Whalley, H. C., Marioni, R. E., Porteous, D. J., Davies, G., Deary, I. J., . . . McIntosh, A. M. (2019). Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat Neurosci*, *22*(3), 343-352. https://doi.org/10.1038/s41593-018-0326-7

Hu, Y., Li, M., Lu, Q., Weng, H., Wang, J., Zekavat, S. M., Yu, Z., Li, B., Gu, J., Muchnik, S., Shi, Y., Kunkle, B. W., Mukherjee, S., Natarajan, P., Naj, A., Kuzma, A., Zhao, Y., Crane, P. K., Alzheimer's Disease Genetics, C., . . . Zhao, H. (2019). A statistical framework for cross-tissue transcriptome-wide association analysis. *Nat Genet*, *51*(3), 568-576. https://doi.org/10.1038/s41588-019-0345-7

Ingelman-Sundberg, M. (2005). Genetic polymorphisms of cytochrome P450 2D6 (CYP2D6): clinical consequences, evolutionary aspects and functional diversity. *Pharmacogenomics J*, *5*(1), 6-13. https://doi.org/10.1038/sj.tpj.6500285

Kim, Y., Xia, K., Tao, R., Giusti-Rodriguez, P., Vladimirov, V., van den Oord, E., & Sullivan, P. F. (2014). A meta-analysis of gene expression quantitative trait loci in brain. *Transl Psychiatry*, *4*(10), e459. https://doi.org/10.1038/tp.2014.96

Lau, A., & So, H. C. (2020). Turning genome-wide association study findings into opportunities for drug repositioning. *Comput Struct Biotechnol J*, *18*, 1639-1650. https://doi.org/10.1016/j.csbj.2020.06.015

Liu, A. E., & Kang, H. M. (2022). Meta-imputation of transcriptome from genotypes across multiple datasets by leveraging publicly available summary-level data. *PLoS Genet*, *18*(1), e1009571. https://doi.org/10.1371/journal.pgen.1009571

O'Brien, H. E., Hannon, E., Hill, M. J., Toste, C. C., Robertson, M. J., Morgan, J. E., McLaughlin, G., Lewis, C. M., Schalkwyk, L. C., Hall, L. S., Pardinas, A. F., Owen, M. J., O'Donovan, M. C., Mill, J., & Bray, N. J. (2018). Expression quantitative trait loci in the developing human brain and their enrichment in neuropsychiatric disorders. *Genome Biol*, *19*(1), 194. https://doi.org/10.1186/s13059-018-1567-1

Schmidt, T. T., & Berkemeyer, H. (2018). The Altered States Database: Psychometric Data of Altered States of Consciousness. *Front Psychol*, *9*, 1028. https://doi.org/10.3389/fpsyg.2018.01028

Shuren, J. (2018). *Jeffrey Shuren, M.D., J.D., director of the FDA's Center for Devices and Radiological Health and Janet Woodcock, M.D., director of the FDA's Center for Drug Evaluation and Research on agency's warning to consumers about genetic tests that claim to predict patients' responses to specific medications*. Retrieved March 27 from https://www.fda.gov/news-events/press-announcements/jeffrey-shuren-md-jd-director-fdas-center-devices-and-radiological-health-and-janet-woodcock-md