BIOLOGICALLY INFORMED FEATURE SELECTION IN LARGE SCALE GENOMICS

by

William Stone

Submitted in partial fulfillment of the requirements
for the degree of Master of Computer Science

at

Dalhousie University
Halifax, Nova Scotia
December 2019

Table of Contents

List of Tables

List of Figures

Abstract

Predictive genetics is a promising field of research, particularly in medical science where the ability to identify disease or treatment response could provide novel methods of mitigating their negative effects. Machine learning represents the most obvious tool that can be used to this end, however a notable property of genetic data that proves difficult for machine learning is a significant imbalance between samples and features, indicating the need for feature selection. The dataset we used was collected from multiple international centres and includes subjects with bipolar disorder, some of whom respond to the drug lithium and some who do not. We first select the features that were measured jointly by each data collection centre and show that above chance classification is possible with these data, despite significant overfitting which indicated the need for further feature space reduction. We then introduce a novel method capable of reducing the number of features even further so as to be bounded by the number of subjects. This method uses the hierarchical structure of genetic data to select feature subsets and evaluate their fitness individually before including the best ones in the final feature set. We show that our method improves on the first method while maintaining biological interpretability.

Acknowledgements

Thomas, thanks for you advice and help during the last two years and for the opportunity to be a part of a great research group. Abraham, your wisdom, expertise, and friendship have been an invaluable resource. As well, thanks to my committee members, Dr. Sageev Oore and Dr. Martin Alda, for taking the time to offer your advice and help me through this process.

To all of my friends and family, thank-you for your support. Jon, we made it! Thanks for always being up for bouncing ideas around. Annelise, you're simply the best. Thanks for everything.

Mom, your support throughout all of my academic endeavors has been essential to my success and I don't know where I would be without it. Dad, from a young age you have inspired within me the insatiable curiosity that has driven many late nights of thinking and tinkering. Both of you, as well as Kathryn, have also imparted to me a general outlook on life that has similarly been crucial to my academic success.

Chapter 1

Introduction

1.1 Motivation

The ability to predict disease and treatment response in advance could allow novel ways of managing or eradicating their negative effects. Given that many phenotypes of interest have a genetic underpinning, using genetic data has been a growing topic of interest. This has been spurred both by falling costs of sequencing and a growing list of discoveries [33]. A common method of measuring genetic data is known as a SNP micro-array, which measures individual genetic deviants that are commonly referred to as single nucleotide polymorphisms (SNP, pronounced 'snip'). This technique can yield hundreds of thousands or even millions of measurements across the genome for each individual.

The current standard method of analyzing micro-array data utilizes a technique called logistic regression analysis (LRA) and is most often referred to as a genome-wide association study (GWAS). LRA seeks to determine the degree of statistical association between each individual SNP and the phenotype in question. This method is limited in that it cannot capture multivariate effects between SNPs and the phenotype since it applies a classifier to many features individually. Interactions between variants are of considerable interest however given that many complex disease phenotypes are polygenic in nature. Moreover, interactions between variants and complex phenotypes cannot be

meaningfully modeled using individual features with small effect sizes. In contrast to LRA, many other supervised machine learning (ML) techniques are capable of modeling multivariate interactions between features and a dependent variable, thus making more advanced ML an attractive alternative.

A common attribute of micro-array data that is relevant from a ML perspective is the imbalance between the number of features and the number of samples. The number of features can be on the order of millions whereas the number of samples, in the largest of studies, is only on the order of several tens of thousands [27]. Moreover, effect sizes of individual SNPs are small which affects the statistical power of the analysis. In addition, other strong signals such as population structure (genetic variations between populations that are unrelated to the phenotype) may obscure the effect being studied. These properties of SNP micro-array data thus make a naive application of ML techniques, where we use the entire genome as a feature vector, computationally intractable.

We have so far outlined two possible methods of analyzing micro-array data that exist at two extremes: analyzing variants one at a time by applying LRA, and all variants at once by applying an ML classifier to all features naively. Both of these methods have properties that make them undesirable. Fortunately, there exists a hierarchical organization of genetic structures that range in between these two extremes, which is digitally captured by tools such as Gene Ontology [2] that can be used for data analysis. Using such tools, we may organize the individual variants into chromosomes, genes or functional classes that are involved in certain biological processes. In this work, we propose a framework with which it is possible to both take advantage of multivariate interactions between SNPs and avoid the intractability of analyzing the entire genome at once by using these semantic structures. We argue that this

technique is capable of providing competitive classification results while still maintaining biological interpretability.

1.2 Overview

In this work we used machine learning (ML) in an attempt to predict whether or not subjects with bipolar disorder responded to lithium as a mood stabilizer. To do this, we used a SNP micro-array dataset that was collected from different international centres. On top of the large sample/feature imbalance that is typical of SNP micro-array data, the international nature of this dataset posed additional challenges. Namely, the labeling of the phenotype is not perfectly consistent between sites, and some sites use different data collection platforms that do not measure exactly the same features and so an imputation method was applied to fill in the spaces.

Given that this dataset was initially untouched by ML, we first took an approach to reducing the feature space that eschewed the use of sophisticated feature selection or biological hypotheses. Instead, we simply relied on the fact that complete data is better than imputed data and selected all features that were measured in common by each platform. This process so happened to leave us with a feature space that, while large, was within the realm of computational tractability. With this so-called "non-imputed" dataset, we attempted to determine if a) any detectable signal related to lithium response could be found and b) if so, to what degree is this signal affected by data collection site.

The experiment using the non-imputed dataset showed that prediction is not possible on the entire aggregated dataset, but that it may be possible to predict response slightly above chance for a subset of the sites. This above

chance performance was first evinced by comparing the Cohen's kappa statistic to that which would be expected by a null classifier. Later, we further back up this claim by showing that the SNPs deemed most important by the trained classifier are over-represented in genes that have been previously and independently associated with bipolar disorder. Finally, we show with this experiment that each classifier we train perfectly (or near perfectly) overfits the training set, indicating the need for a more significant reduction in the size of the feature space.

Directed by the previous experiment, we next looked to apply feature selection to the entire imputed dataset. For this task we required a feature selection method must balance effectiveness, tractability, and intepretability, but found existing methods that have been applied to this form of data lacking. We therefore introduce a novel method of feature selection for SNP micro-array data that we call gene-wise feature selection. Our proposed method utilizes the hierarchical structure of genetic data to identify biologically relevant feature subsets, and analyzes the predictive capacity of each in order to aggregate useful signal into a final feature set.

With this feature selection method, we performed the same classification analyses as in the previous experiment. Similarly, we found that it was not possible to predict response when the sample was combined. Also, we showed that slightly above chance prediction was possible in one of the two sites that contained a large enough sample for gene-wise selection work. Fortunately, a subset of samples from this site overlapped with a dataset of clinical features which a separate work has used to identify a group of more easily predictable subjects. We wished to see if the difference in the ability to predict these subjects with clinical data extended to genetic data, and so we applied our

method to these subjects separately. This analysis showed a significantly improved ability to genetically predict the more clinically predictable subjects.

Finally, we conclude that there is indeed some genetic signal involved with this phenotype and that this signal is stronger in at least one subset of samples from one of the sites. We argue that our gene-wise feature selection method could be of use for other datasets in that it has been shown to be effective, it is computationally tractable, all while maintaining biological interpretability, but also that there are possible improvements. We also suggest that any future genetic study of lithium response in bipolar disorder be coupled with the collection of ancillary clinical data.

1.3 Thesis Outline

This document is organized as follows. Chapter 2 details relevant background material including the representation of SNP micro-array data, the standard method of GWAS, discussion of the need for and limitations of various feature selection methods, and the details of the dataset that we use in this work. Chapter 3 outlines exploratory work we have done using a subset of the feature space which guides our future work. In Chapter 4 we detail a novel method of constructing a genetic feature set that is tractable for ML methods and present the results of applying this method to our dataset. In Chapter 5 we summarize the results of Chapters 3 and 4 and discuss potential future directions this research could take.

Chapter 2

Background

2.1 Representation of Genetic Variation

Genetic variation influences traits such as hair color, height, and even various diseases. The commonly used quantum of genetic variation is known as a Single Nucleotide Polymorphism (SNP, pronounced "snip"). As its name suggests, a SNP represents a single nucleotide (A, C, T, G) in a strand of deoxyribonucleic acid (DNA) that differs with respect to that same location on the DNA of other members of the given species. For each location that is measured, we refer to the more common variant as 'major' and the less common as 'minor'. Microarray data are thus represented as the integer number of minor alleles at each locus. This count can have a value of 0 (homozygous major), 1 (heterozygous) or 2 (homozygous minor) since there are two copies of each chromosome. A feature set $\mathbf{X} = (x_{ij})_{i=1..n_s}^{j=1..n_g}$ consisting of n_s subjects and n_g SNPs is therefore a set of binomial counts $x_{ij} \in \{0, 1, 2\}$. The target variable for a binary phenotype is simply represented as a binary vector, $\mathbf{y} = (y_i)_{i=1..n_s} \in \{0, 1\}^{n_s}$.

In this work we make frequent reference to hierarchically organized structures in the genome, and we provide a visual representation for this hierarchy in Figure 2.1. As described above, SNPs represent the smallest quantum of genetic variation and as such exist at the bottom of the hierarchy. A SNP may lie in an area of the genome known as a gene which, as a unit, encodes the information necessary to create a protein. SNPs that are part of genes are

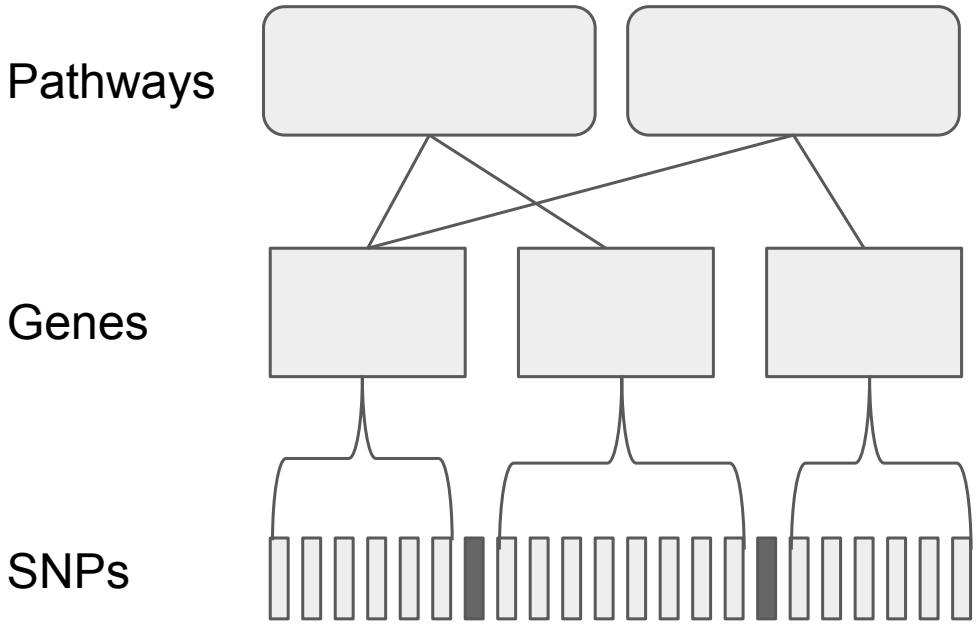

Figure 2.1: A simplified representation of the hierarchical organization of genetic structures. At the bottom, SNPs signify the fundamental quanta of genetic variation. Some SNPs are associated with blocks of the genome called genes (lighter colored) which encode information necessary for creating proteins, whereas others are not (darker colored). Genes may be part of higher level pathways which, for example, are involved in certain activities of the cell or other biological processes.

referred to as *intragenic* whereas SNPs outside of genes are *intergenic*. Proteins that are encoded by separate genes may interact in ways that give rise to higher order biological processes. When one or more genes are involved with such a process, we group these genes into structures known as pathways.

2.2 Standard Association Analysis for Micro-Array Studies

The standard method for studying SNP micro-array data focuses on logistic regression analysis of each SNP. In the case of a binary phenotype, each SNP is entered into a logistic regression model along with other covariates that may correlate with the phenotype. Upon fitting of the logistic model,

a degree of association is calculated by way of a p-value that represents the probability that the given result would be observed from the null distribution by random chance. A value of 0.05 is widely accepted threshold for significance in many statistical studies, however it is insufficiently strict in the regime where hundreds of thousands, if not millions, of such statistical tests are being performed. For this reason, a Bonferroni corrected significance threshold is typically adopted in GWA studies thus resulting in a widely used significance threshold of $5e^{-8}$ [7]. This method has the ability to identify genotype-phenotype associations involving single SNPs that are unlikely to arise by chance, but fails to account for any SNP-SNP interactions. For a more detailed description of logistic regression analysis please see the following section (2.3.3).

Most notable among the covariates that GWA study practitioners commonly include are the first several principal components (PC) of the n_s by n_g feature array. Price *et al.* suggest that the inclusion of PCs in the logistic model reduce the rate of false positives as population structure, which is captured by these first few principal components, can account for some of the variance in phenotype [25]. In studies that collect data using multiple different genotyping platforms, the platform is sometimes also included as a covariate as it can be a confounding factor that is unrelated to the underlying biology.

2.3 Classification Models

2.3.1 Logistic Regression

The logistic regression classifier is a linear method used for estimating the probability of a binary variable, $\hat{y}$, given some observed data, $\mathbf{x}$. This estimate is represented as,

$$\hat{y} = p(y|\mathbf{x}, \beta) = \frac{1}{1 + exp(\beta^{\intercal}\mathbf{x})}, \tag{2.1}$$

where β is a parameter vector with length equal to the number of measurements for each sample. Let the function $\mathcal{L}(\beta, \mathbf{x})$ represent the likelihood of seeing parameters β given some measurements, $\mathbf{x}$. The optimal parameters of the logistic model are those that maximize the likelihood function (or equivalently its logarithm), $\mathcal{L}$. For the logistic model, the log-likelihood function is represented as,

$$\log \mathcal{L} = \sum_{i}^{n} y_i \log \hat{y} + (1 - y_i) \log(1 - \hat{y}), \tag{2.2}$$

where $\hat{y}$ represents the logistic model.

In this work, we use the scikit-learn implementation of the LR classifier [24]. For minimizing the loss function, we use the limited-memory Broyden–Fletcher–Goldfarb–Shanno (l-BFGS) optimization algorithm. This algorithm is a second order gradient method that approximates the Hessian of the function it is optimizing, a process which runs more efficiently when the feature space is large. We otherwise use the default parameters, which includes an l_2 penalty on the loss function.

One of the benefits of a linear model comes from its interpretability. For the logistic model, the magnitude of each coefficient, β_j, informs us of the size of the effect that the corresponding feature has on the prediction. We note also that each coefficient carries a sign, which in our case tells which class the presence of minor alleles affects.

2.3.2 Extreme Gradient Boosting

XGBoost (Extreme Gradient Boosting, XGB) is a widely used gradient tree boosting algorithm that can be applied to both classification and regression problems [3]. While similar to other tree ensemble methods such as random forests in that multiple weak models with high variance are combined in order to create a single stronger model, the way that a gradient boosted tree model combines weak learners differs. Namely, the gradient boosting method trains models sequentially and combines them additively such that the k^{th} model is represented by,

$$F_k(\mathbf{x}) = F_{k-1}(\mathbf{x}) + h_k(\mathbf{x}), \tag{2.3}$$

where $h_k(\mathbf{x})$ is the model that is trained at the k^{th} iteration. The gradient boosting model gets its name from the fact that it treats the addition of new models as a gradient based optimization,

$$F_k(\mathbf{x}) = F_{k-1}(\mathbf{x}) + \alpha \frac{\partial \mathcal{L}(\mathbf{x}, F(\mathbf{x}))}{\partial F(\mathbf{x})}\Big|_{F(\mathbf{x})=F_{k-1}(\mathbf{x})}, \tag{2.4}$$

where α is a learning rate, and $\mathcal{L}$ is a loss function. For simplicity we assume a regression problem here and use the mean squared error as the loss function, though we note that the process is analogous for classification but with a different loss function. The loss is therefore given by,

$$\mathcal{L}(\mathbf{x}, F(\mathbf{x})) \propto (\mathbf{y} - F(\mathbf{x}))^2. \tag{2.5}$$

Differentiating with respect to F (rather than the model parameters), we obtain,

$$\frac{\partial \mathcal{L}(\mathbf{x}, F(\mathbf{x}))}{\partial F(\mathbf{x})} \propto \mathbf{y} - F(\mathbf{x}). \tag{2.6}$$

Thus at each iteration the new model, $h_k(\mathbf{x})$, is trained to predict the mistakes that the previous model has made and is scaled by the learning rate for stability. The final model can be represented as,

$$F(\mathbf{x}) = \sum_{k}^{K} h_k(\mathbf{x}). \tag{2.7}$$

The gradient boosted tree model provides a feature importance metric for each individual feature. For an individual tree, the number of splits for which a given feature causes an improvement is counted and weighted by the number of observations that split is responsible for. This is then and normalized by the number of features,such that the sum of all feature importances is one, and then averaged over all trees in the ensemble. We note that a key difference between the feature importances of this model and of the logistic model is that this metric carries no sign.

The XGBoost algorithm uses the same model as outlined above, but makes several modifications to the training algorithm in comparison to the original gradient boosting method that improve both classification performance and scalability. Namely, XGB implements regularization that penalizes model complexity and prevent overfitting, as well as novel data structures that reduce memory utilization and improve parallelization in training.

2.3.3 Logistic Regression Analysis

Logistic regression analysis (LRA) uses the same model and cost function (equations 2.1 and 2.2 respectively) as the LR classifier, however with a different end goal. Rather than performing predictions, LRA seeks to determine the

degree of statistical association between a measured feature and a dependent variable.

In the context of a GWA study the measurement, $\mathbf{x}$, corresponds to a SNP as well as any other covariates that may be of importance (i.e. population structure). The objective of association analysis with logistic regression is to test the following null hypothesis: the SNP in question is unrelated to the phenotype. With this hypothesis, we also assume that the coefficient in the logistic model that corresponds to the SNP variable, β_{SNP}, is drawn from a normal distribution centered at zero. To reject the null hypothesis and say that it is likely that a SNP is indeed associated with the phenotype, we must observe a value of β_{SNP} that is far enough from zero that it is unlikely to be due to random chance.

To calculate the probability that we may reject the null hypothesis we calculate the t-statistic,

$$t_{SNP} = \frac{\beta_{SNP}}{\sigma_{SNP}}, \tag{2.8}$$

where σ_{SNP} is the standard error of the normal distribution from which we assume β_{SNP} was drawn. We then square the t-statistic and measure its probability using the Chi-squared distribution. For this computation we use the newton-raphson python package [23].

2.4 Lithium Response in Bipolar Disorder

Bipolar Disorder (BD) is a neuropsychiatric illness characterized by recurring episodes of mania and depression separated by periods of partial or even full recovery. While it is possible to provide effective treatment, illness course can differ significantly between patients and it can take up to ten years to find an

appropriate medication [8]. Given that BD is also associated with significantly increased risk of death by suicide, particularly in the early course of the illness, reducing the time taken to achieve an effective intervention is an important endeavour [17].

The procedure for finding the optimal medication course is effectively trial and error. A contributing factor to the delay in treatment is the number of trial medications given to a patient before finding the optimal drug. A potential avenue for reducing the time delay in treatment is to predict treatment response in advance.

Lithium is a commonly used medication for treating BD, though only approximately 30% will be good responders [29]. Response to lithium as a mood stabilizer among patients with BD has been shown to aggregate within families, thus indicating lithium response may have a genetic factor [11]. To this end the International Consortium on Lithium Genetics has collected the largest micro-array dataset on lithium response to date [13]. With this dataset, it is possible to perform the first ever predictive analyses on the lithium response phenotype using genetic data.

2.5 Dataset

The International Consortium on Lithium Genetics (ConLiGen) is an organization spanning multiple institutions and is dedicated to studying the genetic origins of lithium response among patients with BD [30]. ConLiGen is responsible for constructing the largest ever data set of lithium responders/non-responders. Genotype and phenotype data were collected from subjects living in Europe, the Americas, Asia, and Australia in the interval between 2008 and 2013. The genotype data were collected on several different brands of micro-array platforms. More details on data collection can be found in Hou *et al.*

[13].

Lithium response, the target variable in this study, was rated on a scale from 0-10 commonly referred to as the Alda Scale [16]. Unless otherwise stated, we define a lithium responder as a subject with a score $\geq$7. In the aggregated sample from all sites, 29% of subjects are classed as responders and the rest non-responders. Each site individually differs in it's proportions of responders and non-responders, and these distributions are summarized in Table 2.1.

Institution	**Responders**	**Non-Responders**
University of Cagliari, Italy	55 (28%)	141 (72%)
Dalhousie University, Canada	159 (45%)	194 (55%)
University of NSW, Australia	13 (20%)	50 (80%)
Poznan University of Medical Sciences, Poland	47 (48%)	50 (52%)
UC San Diego, USA	23 (11%)	192 (89%)
RIKEN Brain Institute, Japan	31 (24%)	97 (76%)
Mayo Clinic, USA	22 (23%)	72 (77%)
University of Würzburg, Germany	30 (17%)	145 (83%)
Karolinska Institutet, Sweden	138 (45%)	166 (55%)
National Taiwan University, Taiwan	13 (14%)	79 (86%)
Obregia Hospital, Romania	32 (21%)	120 (79%)
University of Geneva, Switzerland	13 (23%)	44 (77%)
University of Barcelona, Spain	20 (27%)	54 (73%)
INSERM, France	38 (18%)	172 (82%)
ALL	634 (29%)	1576 (71%)

Table 2.1: The distribution of lithium responders and non-responders by each of the 14 sites.

In consortium level genomic studies, it is not uncommon for different data collection centers to genotype subjects on different micro-array platforms, which sometimes assay different sets of SNPs. Figure 2.2 shows a simplified representation of this for three different platforms and five SNPs. In such cases, it is also not uncommon for researchers to perform a statistical imputation that fills in areas of the genome that are not measured for some platforms using a set of densely sampled reference genomes. Hou *et al.* performed such an imputation to create a larger set of SNPs that are common to all subjects which we refer to as the **imputed data set**. Across all platforms there were

Figure 2.2: A simplified representation of the overlap in SNP measurements between three different platforms. Darker squares signify that the given SNP on the given platform was measured, whereas lighter colored squares indicate that it was not. In this diagram the non-imputed dataset would contain SNPs 2 and 5. To construct an imputed dataset, certain SNPs would have to be imputed for all of the subjects on each platform (e.g. SNP 4 would have to be imputed for all subjects genotyped using Platform 1).

a number of directly genotyped SNPs in common. We refer to the overlapping set of directly genotyped SNPs as the **non-imputed data set**.

Upon receiving the imputed dataset from ConLiGen, it was represented as categorical probabilities for the locus being homozygous on the major allele (p_0, zero minor alleles), and the probability of the locus being heterozygous (p_1, one minor allele). For each SNP in the dataset, we calculated the probability of the locus being homozygous on the minor allele (two minor alleles) as $p_2 = 1 - p_0 - p_1$. In order to represent the dataset in terms of minor allele counts, we simply took the max probability value at each loci for each subject and assumed the corresponding zygostity.

Hou et al. [13] imposed quality control measures that are typical for GWA studies for retaining both SNPs and subjects. These included: per subject genotype missingness, control for autosomal heterozygosity rate, minor allele

frequency pruning, Hardy-Weinberg equilibrium pruning, and linkage disequilibrium pruning using PLINK v1.07 [26]. This set of quality control procedures is typical of genome-wide association studies [18]. On top of the original quality control, we also removed sites that had fewer than 50 total samples or fewer than ten lithium responders. We removed these sites as they would give rise to sample size problems in some of the analyses per perform (namely, the site-level analysis outlined in Section 3.2.2). The imputed data set contained 2210 subjects and 5,795,772 SNPs after quality control, whereas the non-imputed set contained the same number of subjects and 47,465 SNPs. Further details regarding quality control and imputation can be found in [13].

2.6 Feature Selection

Modern micro-array platforms are capable of measuring over a million SNPs for each subject, whereas even the most notable studies that have used micro-array data contain only on the order of several tens of thousands of samples [33]. Moreover, modern genetic studies are typically international efforts carried out by consortia of data collection centres that do not always use the same platforms. These platforms may only measure some SNPs in common, leaving significant missingness in a platform-correlated pattern. In such studies, researchers are likely to perform a statistical imputation in order to maximize the coverage of SNPs over the genome which results in an even larger number of features for each sample.

End-to-end learning defines a process in which a model is used to learn the relationship between an entire unmodified feature space and a dependent variable. Usage of this technique has been a prominent trend in machine learning research in recent years which has mainly been facilitated by widespread access to “big data”. Two notable areas of end-to-end learning success

are computer vision (CV) and natural language processing (NLP). The data in each of these areas exist in very large feature spaces, but large numbers of samples are also available, which is not true for micro-array data. Moreover, both image and language data are highly strucutred in that individual features (pixels, words, etc.) are not positionally invariant within the feature vector. Data that lack positional invariance allow for the application of models such as convolutional and recurrent neural networks which efficiently share parameters across input features. In contrast, micro-array data are positionally invariant and thus it is not obvious that they will benefit from such parameter sharing models.

The properties of micro-array data discussed above clearly indicate the need for feature selection. At a high level, we place feature selection techniques into three categories: online methods, embedding methods, and input variable selection (IVS) methods.

2.6.1 Online Feature Selection

Online methods can incorporate feature selection into the learning algorithm itself. An example of this is penalized regression, wherein an extra constraint is added to the original model's cost function. This constraint serves to force model weights associated with useless features towards zero. One commonly used penalty method is known as the least absolute shrinkage and selection operator (LASSO) penalty [32]. With this method, we represent the cost function as follows,

$$J(\beta, \mathbf{X}) = \mathcal{L}(\beta, \mathbf{X}) + \lambda||\beta||_1, \tag{2.9}$$

where β represents model weights, $\mathbf{X}$ represents training data, $\mathcal{L}$ is the unmodified cost function, and λ is a parameter that dictates the strength of the penalty. Other penalty methods work in a similar fashion but with different terms. For example, ridge regression uses the square of the weights rather than the absolute value as in the LASSO penalty, and elastic net regression combines both the LASSO and ridge regression penalty terms [12, 37].

Kohannim *et al.* propose a method for out of sample prediction of a phenotype with micro-array data using Elastic net regression [14]. They apply this technique to predict temporal lobe volume and are able to recover previously identified genetic relationships to the phenotype by finding which SNPs consistently had the largest regression coefficients. Their method however also uses an IVS feature selection technique before using Elastic net. More specifically, they apply LRA to each SNP and retain only those that obtain a p-value below some threshold for the next phase of their analysis.

2.6.2 Embedding Methods

Embedding methods seek to construct an alternative representation of the data in a lower dimensional space. A deep learning technique for creating an embedded data representation is an autoencoder. While there are several variations on autoencoding networks (most notably denoising and variational autoencoders), they can generally be broken down into two modules: an encoder and a decoder. The encoding module seeks to compress a feature vector into a smaller space, while the decoder learns to reconstruct the encoded feature vector into its original form.

Romero *et al.* make use of a denoising autoencoder in order to predict the parameters of a secondary neural network which they train to classify ancestry in the 1000 genomes dataset [28, 6]. This had the effect of reducing the

number of trainable parameters in the classification network by a factor of 600. However, the signal related to ancestry in genetic data is notoriously strong and so their model architecture may not generalize to more subtle genetic signals (we were also unable to independently reproduce their findings).

Fergus *et al.* and Abdulaimma *et al.* both use stacked autoencoders to create compressed representations of genetic data for the purposes of predicting preterm birth and type-2 diabetes respectively [9, 1]. Each of these works, however, does not solely rely upon embedding method. Instead, they apply LRA in much the same fashion as Kohannim et al. [14] as discussed above. This was likely done in order to remove large signals in the genome, such as population structure, which an autoencoder could learn to rely on more heavily than the smaller signal from the target phenotype.

In each of the above cases, the application of embedding methods came with assumptions: a) the phenotype signal is very strong, and b) the embedding method was applied on top of another feature selection method. We therefore conclude that it is not immediately obvious that we can rely solely on embedding methods to solve the problem of the sample and feature imbalance.

2.6.3 Input Variable Selection

Input variable selection (IVS) methods select features before training the final model. We can further break IVS down into model based and model free methods. Model based methods use a preliminary model that can measure the relative importance of each feature and retain only the best. Alternatively, model free methods rely on statistical measurements or heuristic knowledge of the features in order to pre-select useful features.

An example of a model based method uses LRA to measure a p-value for each feature and retain only the features with p-values beneath some threshold.

Several recent works have used this technique for feature selection on micro-array data with various phenotypes. Fergus *et al.* and Abdulaimma *et al.* select features before employing an embedding method to further constrain the feature space before applying a multi-layer perceptron (MLP) to perform classification [9, 1]. Montanez *et al.* follow feature selection directly using an MLP to classify obesity and Maciukiewicz *et al.* similarly follow feature selection using both decision trees and support vector machines to classify response to duloxetine in subjects with major depressive disorder [21, 15].

There are several points worth noting about these four works. Each study used datasets that were collected using a single platform, meaning that the feature sets were not as large as can be expected in a consortia level study. The scalability of this technique as far as we are aware has not been addressed. Also, the initial feature selection is purely based upon an associative analysis of single SNPs, thus disallowing the possibility of capturing SNPs with epistatic interactions that don't have significant enough individual associations from the outset. Most importantly is the issue of out of sample feature selection. Three of the four works discussed above ([9, 1, 21]) do not make it explicitly clear that they performed LRA *only* in their training sets. To highlight the importance of cross-validation in feature selection, we perform a simple experiment using a small subset of our data which can be found in Section 2.6.5.

Another example of model based IVS, called a "wrapper method", uses a predictive model to iteratively train and evaluate subsets of the total feature set to determine if features within each subset should be retained. Pahikkala *et al.* introduce a method they call "Greedy Regularized Least-Squares (RLS)" and claim that it is the first application of a wrapper based feature selection method to SNP micro-array data. They begin with an empty feature set, train separate models for each feature, add the best performing feature according

to area under the receiver operator curve, and repeat this procedure until a predefined quota of features has been filled. The main drawbacks of this method however are that it is only tractable for selecting small sets of SNPs from relatively small datasets, and that it is intractable when using more complicated models.

Yin *et al.* attempt to predict amyotrophic lateral sclerosis using microarray data [34]. In an initial phase of feature selection on their data, they use heuristic knowledge of the phenotype to pre-select SNPs. More specifically, they extract fixed size groups of SNPs that fall within the promoter regions (regions that dictate the degree to which the corresponding genes are transcribed) of each gene in their dataset. The authors then use a deep learning classifier to determine the importance of each promoter region, and combine the most important regions into a final dataset.

2.6.4 Our Proposed Method

In our proposed feature selection technique, we combine several of the above mentioned IVS methods. We start by using independently discovered heuristic knowledge and pre-select SNPs that are within a biological pathway that is associated with lithium response. We next apply a "biologically informed" wrapper method by grouping SNPs according to gene and measuring the importance of each gene by its out of sample classification performance. Finally, we use feature importance metrics from the gene-wise classifiers to determine the most important SNPs in each selected gene and combine these SNPs to form the final feature set.

2.6.5 Proper Cross-Validation with Logistic Regression Analysis

To highlight the importance of proper cross validation when applying feature selection, we performed a simple experiment using the Halifax sample, the non-imputed dataset and the logistic regression classifier. We performed LRA both with and without proper five fold stratified cross validation and set varying p-value thresholds to create datasets. The results of this experiment are shown in Figure 2.3. Here we see a stark difference between the different feature selection protocols, with the improperly cross validated method achieving perfect performance at the largest threshold. This is a clear demonstration of the importance of proper cross validation when performing feature selection.

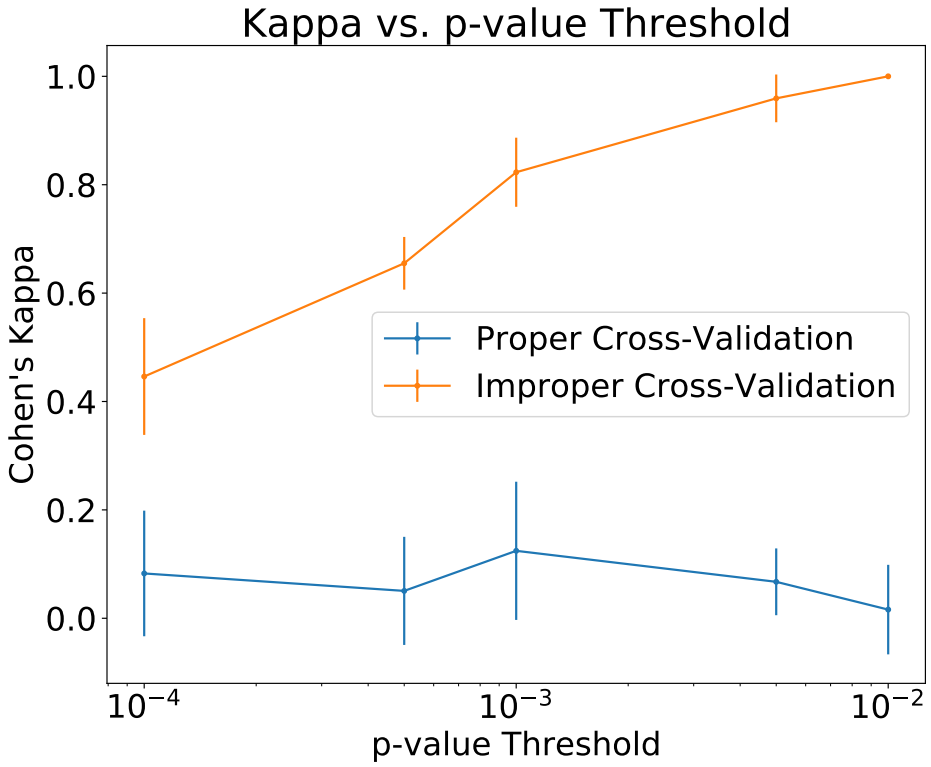

Figure 2.3: p-value inclusion threshold vs. Cohen's kappa for the case where feature selection is done out of sample (blue) and within sample (orange). In each case, the logistic regression classifier (with the same default parameters as in all other experiments) was used to perform the final prediction. Error bars represent the standard deviation across the five cross-validation folds.

Chapter 3

Platform Overlap Feature Selection

3.1 Motivation

The ConLiGen dataset is the largest micro-array dataset in existence that addresses lithium response in subjects with BD. To the best of our knowledge, no out of sample predictive analyses have been performed on these data. We therefore take a simplistic approach to this problem so that the results of this chapter may serve as a benchmark for future studies.

In this chapter we perform a basic multi-site classification study where we aim to determine the degree to which we can predict lithium response, and the degree to which the site of data collection confounds these results. For the sake of simplicity, we eschew the use of any feature selection techniques or biological hypotheses about the phenotype and use the non-imputed dataset. In a post-hoc analysis we analyze the biological relevance of the features deemed by the classifiers to be most important by means of gene set analysis.

3.2 Methods

3.2.1 Dataset

In these analyses we use the non-imputed dataset, which includes 2210 subjects from 14 different centers where each subject has a total of 47,465 directly genotyped SNPs. We used the non-imputed dataset for two main reasons: the imputed dataset would be intractable without using some method of feature

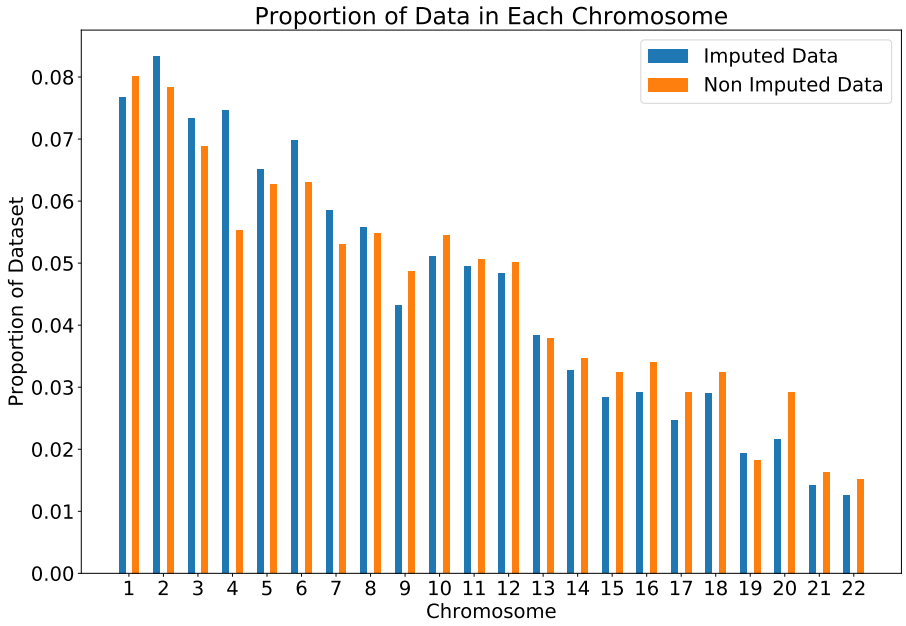

Figure 3.1: Proportions of both the imputed and non-imputed datasets that belong to each chromosome. The most significant imbalance is on chromosome 4 where the non-imputed dataset has around 2% less coverage.

selection and the non-imputed dataset contains only complete genotype information. For more details on data collection and quality control see Section 2.5.

Figure 3.1 compares the proportions of both the imputed and non-imputed datasets that come from each chromosome. The non-imputed dataset provides relatively similar coverage across the genome as the imputed dataset.

3.2.2 Classification Analyses

We performed four sets of classification analyses. These are: i) aggregate analysis, ii) site-level analysis, iii) leave one site out analysis, and iv) predict one site out analysis. The aggregate analysis is performed using all subjects at once and aims to measure the overall classification accuracy that can be achieved on the whole sample, whereas the site-level analysis is performed on

each site separately. The leave one site out analysis is performed using all of the data at once save for one site, and is repeated consecutively for each site. This aims to determine the effect that each site has on the aggregate performance. In the predict one site out analysis we train a classifier on all but one site and test on the left out site, with the goal of determining how much signal from the remainder of the sample can generalize to the left out site.

3.2.3 Classifiers

In each analysis we used two classifiers: logistic regression (LR) and XGBoost (XGB). We chose these classifiers for several reasons, namely they are relatively simple and do not require significant hyperparameter optimization, and because we wish to compare a linear classifier (LR) to a non-linear classifier (XGB). For more detailed descriptions of each classifier, see Section 2.3.

3.2.4 Model Criticism

We assessed classification performance using: accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and Cohen's kappa. We viewed Cohen's kappa as a particularly strong metric in this work as it is typically the most conservative under class imbalance, and applied a simulation technique to determine a probability, p, that a trivial classifier (biased coin toss) would produce a better result. This was done by using the number of samples, n_s, in the dataset and the class proportion, $\alpha = \frac{1}{n_s}\sum_i^{n_s} \mathcal{I}[y_i = 1]$, to model simulated sensitivities, β and specificities, γ, using Beta distributions. Then, representing kappa as a function of these parameters, $\kappa(\alpha, \beta, \gamma)$, M trials are performed and the p-value is estimated as the proportion of trials for which the simulated kappa exceed the observed value,

$$p = \frac{1}{M} \sum_{j}^{M} \mathcal{I}[\kappa_{sim} > \kappa_{obs}] \tag{3.1}$$

We deem results with $p < 0.01$ to be statistically significant. We used the criticism package to run this procedure, which can be found at (along with a more detailed description) https://www.github.com/abrahamnunes/criticism.

Due to the sensitive and high risk nature of relying on a genetic prediction of a disease phenotype in a clinical setting, we stress the importance of out of sample testing. To address this concern, we perform all experiments using five fold cross-validation. Because of the class imbalance in this dataset, we specifically apply stratified cross-validation such that both the training and testing sets have the same proportions of positive and negative classes. This process yields a set of five estimates of the true classification performance.

3.2.5 Feature Importance and Gene Set Analysis

Both the LR and XGB classifiers provide feature importance metrics that can be used to interpret which dimensions of the feature space the classifier is most focusing on to make its prediction. By determining the SNPs that contribute most to an above chance prediction, we may gain biological insight into the phenotype.

For the LR classifier trained on a single site, we generate a set of the most important features (hereafter referred to as an *effect set*) by first extracting those for which the sign is the same across all five folds. We then rank these coefficients according to the median of their absolute value, and select the upper quartile. Similarly for the feature importances from the XGB classifier, we determine the set of SNPs which have non-zero feature importance across all folds and again rank them according to their median. To create an effect

set for models trained separately on different sites, we take the SNPs for which the signs (or non-zero importances) agree across all folds for both models.

For the method of analyzing biological importance of features we wish to use, we must represent the effect set in terms of genes as opposed to SNPs (for a visual representation of this distinction, see Figure 2.1). To do this, we use the *biopython* package to annotate each SNP as being part of a gene, omitting those that are intergenic and counting each gene only once [4]. We refer to the entire set of genes covered by our dataset as the reference gene set, and the genes covered by the effect set of SNPs the effect gene set.

The PANTHER gene ontology (GO) database organizes genes according to larger scale functional pathways, such as individual cell components or protein pathways [19]. PANTHER also offers a tool for comparing a reference and effect set of genes that determines whether or not any functional pathways are statistically over-represented in the effect set with respect to the reference set. To do this, PANTHER annotates each gene with an ontology label and measures the proportions of each that appear in the reference set. Given these proportions and the size of the effect set, it then computes the expected number of genes that would appear in a randomly sampled effect set. It then compares the observed number of genes that appears in the effect gene set for each ontology label against the expected number. To determine the significance of the result, PANTHER generates a p-value using Fisher's exact test, and a false discovery rate (FDR) using the Benjamini-Hochberg correction for multiple comparisons. Results are deemed significant where the FDR falls below 0.05.

3.3 Results

3.3.1 Classification Analyses

Table 3.1 shows results for the analysis on the aggregate sample using both the LR and XGB classifiers. Neither the LR classifier (Kappa 0.02 (-0.01, 0.04), p = 0.2) nor the XGB classifier (Kappa 0.0 (-0.02, 0.01), p = 0.4) was able to discriminate response above random chance. We also note that in each case the classifiers were able to discriminate either perfectly or near perfectly between responders and non-responders within the training sets.

Table 3.1 also shows the most notable results for the site level analysis (see Appendix A.1 for the full set of site level results). The LR classifier applied to the Halifax and Würzburg samples showed Cohen's kappa of 0.15 (95% CI (0.07, 0.21), $p = 0.0019$) and 0.2 (95% CI (0.1,0.3), $p = 0.0006$) respectively. Once again we note that both the LR and XGB classifiers achieve perfect discrimination of lithium response in the training sets.

Table 3.2 displays the subset of the results for the leave one site out analysis using the LR classifier. These were the results for which the simulated p-value for kappa fell below the preset bound for statistical significance, indicating that exclusion of these sites brought classification to an above chance level. This occurred for Barcelona (Kappa 0.05, (0.04, 0.06), $p = 0.007$), Romania (Kappa 0.05 (0.02, 0.09), $p = 0.003$), San Diego (Kappa 0.06 (0.04, 0.08), $p = 0.002$), and Würzburg (Kappa 0.05 (0.04, 0.07), $p = 0.003$).

We found no notable results for the predict one site out analysis. For the full set of results for this analysis, see Table A.3.

Table 3.1: Results for the aggregate and site level analysis on both the training and testing sets for each classifier. Table columns represent various classification statistics where each cell contains the mean across five cross validation folds along with a 95% confidence interval. *Abbreviations*: area under the curve (AUC), positive predictive value (PPV) negative predictive value (NPV).

Model	Center	Set	Accuracy	AUC	Sensitivity	Specificity	PPV	NPV	Kappa
LR	All	Test	0.7 (0.7, 0.71)	0.57 (0.55, 0.6)	0.04 (0.03, 0.06)	0.97 (0.97, 0.98)	0.36 (0.27, 0.46)	0.72 (0.71, 0.72)	0.02 (-0.01, 0.04)
LR	All	Train	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (1, 1)
XGB	All	Test	0.7 (0.7, 0.71)	0.57 (0.55, 0.6)	0.03 (0.01, 0.04)	0.98 (0.97, 0.98)	0.3 (0.14, 0.45)	0.71 (0.71, 0.72)	0.0 (-0.02, 0.03)
XGB	All	Train	0.98 (0.98, 0.98)	0.98 (0.97, 0.98)	0.94 (0.93, 0.94)	1 (1, 1)	1 (1, 1)	0.97 (0.97, 0.98)	0.95 (0.95, 0.96)
LR	Halifax	Test	0.61 (0.57, 0.64)	0.62 (0.58, 0.65)	0.23 (0.16, 0.31)	0.91 (0.87, 0.95)	0.68 (0.58, 0.78)	0.59 (0.57, 0.62)	0.15 (0.07, 0.24)
LR	Halifax	Train	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (1, 1)
XGB	Halifax	Test	0.56 (0.52, 0.6)	0.57 (0.52, 0.63)	0.41 (0.33, 0.49)	0.69 (0.59, 0.78)	0.52 (0.46, 0.58)	0.59 (0.56, 0.61)	0.1 (0.02, 0.17)
XGB	Halifax	Train	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (1, 1)
LR	Würzburg	Test	0.85 (0.84, 0.86)	0.6 (0.51, 0.69)	0.13 (0.07, 0.2)	1 (1, 1)	0.8 (0.41, 1)	0.85 (0.84, 0.86)	0.2 (0.1, 0.3)
LR	Würzburg	Train	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (1, 1)
XGB	Würzburg	Test	0.85 (0.83, 0.86)	0.6 (0.56, 0.65)	0.13 (0.07, 0.2)	0.99 (0.98, 1)	0.7 (0.31, 1)	0.85 (0.84, 0.86)	0.19 (0.09, 0.28)
XGB	Würzburg	Train	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (1, 1)

Table 3.2: Results for the leave one site out analysis using the logistic regression classifier for which the simulated p-value for the kappa statistic fell below the preset bound for statistical significance of 0.01 (for the full result set, see Table A.2). Each cell shows the mean value of the statistic over five folds along with a 95% confidence interval. *Abbreviations*: positive predictive value (PPV) negative predictive value (NPV).

Centre	Accuracy	AUC	NPV	PPV	Sensitivity	Specificity	Kappa
Barcelona	0.72 (0.71, 0.72)	0.59 (0.58, 0.6)	0.72 (0.72, 0.72)	0.58 (0.53, 0.63)	0.05 (0.04, 0.07)	0.98 (0.98, 0.99)	0.05 (0.04, 0.06)
Romania	0.71 (0.7, 0.72)	0.6 (0.58, 0.62)	0.72 (0.71, 0.72)	0.59 (0.39, 0.8)	0.05 (0.03, 0.08)	0.99 (0.98, 0.99)	0.05 (0.02, 0.09)
San Diego	0.7 (0.69, 0.7)	0.6 (0.59, 0.61)	0.7 (0.7, 0.71)	0.54 (0.46, 0.62)	0.07 (0.06, 0.09)	0.97 (0.96, 0.98)	0.06 (0.04, 0.08)
Würzburg	0.71 (0.7, 0.71)	0.6 (0.57, 0.62)	0.71 (0.71, 0.72)	0.55 (0.48, 0.63)	0.06 (0.04, 0.08)	0.98 (0.97, 0.99)	0.05 (0.04, 0.07)

3.3.2 Gene Set Analyses

We performed gene set analysis using the effect sets obtained from the site level analysis of Halifax and Würzburg separately, as well as the effect set for these two sites combined. Table 3.3 shows a list of the over-represented cellular component classes using the combination of Halifax and Würzburg feature importances. We see here that the postsynaptic membrane class is over-represented by a factor of 1.71 with an FDR of $8.14e^{-3}$, which meets the bounds of statistical significance. Table A.4 shows a list of all 97 genes that are a part of the postsynaptic membrane class that were in the effect set.

Functional Class	N_{Ref}	N_{Obs}	N_{Exp}	Factor	p-value	FDR
postsynaptic membrane (GO:0045211)	190	97	56.79	1.71	2.40E-05	8.14E-03
synaptic membrane (GO:0097060)	252	124	75.32	1.65	6.28E-06	3.55E-03
synapse (GO:0045202)	618	253	184.72	1.37	1.61E-05	6.83E-03
synapse part (GO:0044456)	494	201	147.66	1.36	1.78E-04	3.78E-02
cell junction (GO:0030054)	634	254	189.5	1.34	4.84E-05	1.17E-02
neuron part (GO:0097458)	871	335	260.34	1.29	3.56E-05	1.01E-02
cell projection (GO:0042995)	1055	387	315.34	1.23	2.47E-04	4.65E-02
cell periphery (GO:0071944)	2461	866	735.59	1.18	4.15E-07	7.04E-04
plasma membrane (GO:0005886)	2407	843	719.45	1.17	1.34E-06	1.14E-03

Table 3.3: Set of PANTHER functional classes that were found to have a statistically significant over-representation when comparing the effect set of genes generated from the Halifax and Würzburg samples to the overall reference set. N_{Ref} signifies the number of genes with the given ontology label in the reference set, N_{Obs} signifies the number of genes in the effect set, and N_{Exp} signifies the number of genes expected to be in the effect set if it were randomly sampled from the reference set. The Factor is the ratio of N_{Obs} to N_{Exp}, p-value is the outcome of Fisher's exact test, and the FDR is the outcome from the Benjamini-Hochberg correction.

3.4 Discussion

This study is, to the best of our knowledge, the first that attempts to perform out of sample classification of lithium response using only genetic data. We found that it was possible to classify response with above random chance

performance only on a subset of the 14 sites, namely Halifax and Würzburg, but that classification on the aggregated sample was trivial. The ability to predict response in only a small subset of the sites suggests a strong degree of between site heterogeneity. While this heterogeneity could stem from many sources, one clear source in this case is the differences in the number of samples and class balances between sites. We also observed a slight, but non-trivial, increase in performance when we left certain sites out of the aggregate sample. Overall, we found that above chance classification of lithium response using genetic data is indeed possible, though significant challenges remain.

We performed a gene set analysis using the SNPs that we found to be most informative between the Halifax and Würzburg samples in the site level analysis. This analysis showed a statistical over-representation of several biological pathways. Some genes in the over-represented pathways (Table A.4), such as ANK3, have been previously and independently associated with BD [10, 22, 31]. We have shown that classification methods are indeed able to recover biologically relevant aspects of the feature space despite the fact that the classification performance is relatively poor.

In each experiment we saw perfect (or near perfect) discriminability in the training sets. In other words, this indicates that a linear decision surface can always be found in this feature space that perfectly separates responders from non-responders, regardless of whether or not the classifier achieves any generalizability. We suspect that this is a result of the dimensionality of the feature space; given so many degrees of freedom, there is always a way for even a linear classifier to overfit the training set without identifying a generalizable decision rule. This suggests that a) cross validation is a very important practice in predictive genetics, and b) future work on these data should aim to reduce the size of the feature space even further.

In this chapter we have performed a set of exploratory experiments on the non-imputed lithium response dataset from ConLiGen. We have shown that it is possible to predict lithium response non-trivially on two subsets of the data from that come from the Halifax and Würzburg sites. Also, using a combination of the most important features from the two classifiable sites, we have shown that even despite relatively poor classification performance it is possible to recover biological information that has been previously associated with the phenotype. This is further evidence that our classification results are not random noise. Given these results along with the clearly demonstrated issue of overfitting, this work provides a suitable benchmark with which we can compare future experiments.

Chapter 4

Gene-Wise Selection of G-Protein Coupled Receptor SNPs

4.1 Motivation

In Chapter 3 we used a small subset of the total dataset to predict lithium response. We showed that it is possible to achieve above chance classification with the non-imputed dataset, and that the most relevant features could be used to identify biological pathways that have previously been associated with BD. The above chance classification, however, was localized to only two of the 14 sites and in both cases left ample room for improvement.

Selecting input features according to the measurement overlap between genotyping platforms, while mildly effective in this case, is a method that may not be appropriate for all GWA datasets. For example, if the overlap is significantly larger and thus intractable, or is not representative of the entire genome. We are interested in finding a more general way to select features that will: a) make use of the entire (imputed) dataset, b) can be applied to any dataset regardless of the mixture of genotyping platforms, c) will result in an even more significant reduction in the size of the feature space compared to the benchmark and d) maintains a high degree of biological interpretability.

In this chapter we employ two feature selection methods simultaneously. First, we use heuristic information about the phenotype in our dataset to preselect SNPs that come from a specific biological pathway that has been

previously shown to be related to BD and lithium response. Next, we introduce a novel two tiered feature selection method inspired by the biological structure of the data that we call 'gene-wise' feature selection. The gene-wise method first identifies informative sets of features at the level of genes, and then further selects the most important SNPs within the most relevant genes. We argue that this method is not only advantageous for selecting important feature sets that result in improved classification accuracy, but is also capable of identifying biologically relevant information about the phenotype.

4.2 Methods

4.2.1 Dataset

In this set of experiments we used the imputed dataset, which consisted of 2210 subjects each consisting of 5,795,772 SNPs. For more information on the imputed dataset, see Section 2.5.

4.2.2 Classifiers

For these analyses we again used the logistic regression (LR) and XGBoost (XGB) classifiers. For more information on these classifiers, see Section 2.3.

4.2.3 A Priori Selection of Genes Related to G-Protein Coupled Receptors

G-protein coupled receptors (GPCR) are a class of proteins that exist on the membranes of cells and serve to transmit chemical signals from molecules that bind to the outside of the cell to the inside of the cell. Previous work by other researchers (using separate data) has provided substantial evidence linking parts of the GPCR pathway to lithium response [5, 20, 35]. We limit our

analysis to SNPs contained within genes that are associated to the GPCR pathway. While our gene-wise feature selection method can in theory scale to analyze every gene within dataset, we focus on features related to GPCR related genes due to the clinical interests surrounding them.

4.3 Model Criticism

We use the same set of performance metrics as outlined in Section 3.2.4. Similarly to the work of Chapter 3, we use cross validation stratified by lithium response in order to prevent overfitting. However, as discussed further in Section 4.3.1, we must take further steps in to prevent overfitting by also performing cross validation in the feature selection step. We therefore employ a "nested" cross validation technique in this chapter where we divide each "outer" training fold into sets of "inner" training and testing folds (also stratified by lithium response). For the outer cross validation procedure we use ten folds, whereas for the inner cross validation we use five. Lastly, we also use the simulation procedure exactly as it was described in Chapter 3 for generating p-values for the kappa statistic.

4.3.1 Gene-Wise Feature Selection

At a high level, our proposed gene-wise feature selection method first filters out only the most predictive genes, and then selects only the best SNPs from within each. We do this under cross-validation in order to prevent overfitting during the feature selection step. This general process is outlined in Algorithm 4.1.

To construct the GPCR gene-wise dataset, we first queried a list of 814 GPCR related genes from the AmiGO database [2]. We next used *pyensembl*

to determine the start and end locations for each of the 814 genes, and selected SNPs that were within 50,000 base pairs of the start or end of the gene [36]. In total, this process produced a total of 297,178 SNPs with some genes accounting for as few as two SNPs while others as many as 2,767.

To measure the "fitness" of each gene we wish to consider both it's classification performance and the reliability of this performance. For each gene, we train a model (either LR or XGB) to predict the phenotype under stratified five fold cross validation. We select Cohen's kappa to be the primary measure of classification performance, and define a metric we call the representativeness that is a function of the kappa values across folds to measure the reliability. The representativeness is based on the exponential of the Shannon entropy,

$$r = \exp\left(-\sum_{i}^{n_{\text{fold}}} p_i \log p_i\right), \tag{4.1}$$

where n_{fold} is the number of inner folds, $\mathbf{p}$ is the normalized set of rectified kappa values across folds ($\kappa_+ = max(\kappa, 0)$), and $r \in [1, n_{\text{fold}}]$. We standardize the representativeness to the interval $[0, 1]$ such that deviations in the values of kappa across folds will push the standardized representativeness, $\tilde{r}$, towards zero whereas perfect agreement in kappa scores across folds will result in $\tilde{r} = 1$. Finally, we combine the representativeness and the maximum value of kappa across folds to create a metric we call the "gene-wise exemplar score" given by,

$$E = \frac{\tilde{r} + max(\kappa)}{2}, \tag{4.2}$$

which is also defined on the interval $[0, 1]$. We refer to genes with a gene-wise exemplar score above a pre-defined threshold as exemplary genes.

Within each exemplary gene we do a further selection of the most informative SNPs. To create an importance ranking we first extract all SNPs for which the regression coefficients have the same sign across folds (or all feature importances being non-zero for the XGB classifier). We then order the remaining features according to the absolute value of their median across folds. In each of the classification analyses, we wish to select a number of features that does not exceed the number of samples in the training set minus one, $n_f = n_{train} - 1$. We preferentially select SNPs from the genes with higher scores by computing a softmax distribution over the exemplar scores, and then taking $n_i = \lfloor softmax(e_i) \cdot n_f \rfloor$ SNPs from the i^{th} gene. Lastly, we use this feature set to train the inference model, which we use to report the final results.

Algorithm 4.1 Pseudocode outlining our gene-wise feature selection method experiment. D represents the dataset, whereas T represents the exemplar score threshold above which a gene is considered to be exemplary. The main loop is over the *outer* cross validation folds, whereas the inner cross validation loops over each gene (on the outer loop training sets) occur in the "train_genes_with_CV" function. There is one "outer_results" variable created per outer fold.

```
procedure EXPERIMENT(D, T)
    for D_train, D_test in Outer_CV do
        n_f = len(D_train) - 1
        gene_results, gene_importances = genewise_testing_with_CV(D_train)
        gene_scores = measure_genewise_exemplar_scores(gene_results)
        exemplar_genes = select_and_sort(gene_scores, T)
        softmax_dist = softmax(gene_scores[exemplar_genes])
        SPNs = [ ]
        for gene in exemplar_genes do
            importance_mtx = gene_importances[gene]
            n_snps = ⌊softmax_dist[gene] · n_f⌋
            best_features = get_best_features(importance_mtx, n_snps)
            SNPs.append(best_features)
        D_train = extract_features(D_train, SNPs)
        D_test = extract_features(D_test, SNPs)
        outer_results = train_and_test(D_train, D_test)
```

4.3.2 Classification Analyses

Similar to Chapter 3, we perform aggregate, site level, leave one site out, and predict one site out analyses. In Section 4.3.1 we outline our feature selection method which involves performing a nested cross-validation procedure. This technique is data intensive and is impractical to perform on some sites that contain small numbers of subjects and responders. We therefore only perform site level analyses on the Halifax and Swedish samples, and we perform a secondary aggregate analysis by combining samples from these two sites.

In these experiments, we use an exemplar threshold of 0.51 to define exemplary genes. We use this threshold as it is only just above the expected score of one of two trivial classifiers. That is, each of the following scenarios would achieve a gene-wise exemplar score of 0.5: a) the gene was perfectly even across folds but with a maximum kappa score of zero, or b) the gene was maximally imbalanced with all folds achieving a kappa score of zero save for one which achieves a perfect kappa score. As well, when computing the softmax distribution across exemplar scores we use an inverse temperature of five. We selected this value due to the fact that any smaller value failed to produce significant unevenness in the softmax distribution.

4.3.3 Subject-Wise Exemplar Score

The gene-wise exemplar score is inspired by the exemplar scoring technique introduced by Nunes *et al.* (work awaiting publication). This work was performed on a dataset that was collected from multiple sites and had the same phenotype as our dataset, but contained clinical data as features as opposed to genetic data. The exemplar scoring technique was designed to measure the accuracy and consistency with which clinical variables could classify a subject

independent of the site from which their data was collected. That is, a higher score indicates that a subject is more "exemplary" of their phenotype given that they are robustly classifiable regardless of which site the a classifier was trained on. Conversely, subjects with lower scores exhibit clinical features that are less consistent across sites and cannot be as easily classified. This analysis showed that the Halifax site had a disproportionately high number of strong exemplars in comparison to other sites in the dataset. We seek to determine whether the limitations in response prediction might be related to between site differences in clinical features.

A set of 320 subjects from the Halifax site overlapped between the clinical and genetic datasets. This was the only portion of the two datasets that overlapped. We performed an experiment to assess the difference in genetic classifiability between subjects with lower and higher "subject-wise" exemplar scores. To do this, we separated the 320 subjects for whom subject-wise exemplar scores were available into two datasets: the bottom 50^{th} percentile and top 50^{th} percentile. We then applied our feature selection method to each dataset separately.

In this experiment, we use a gene-wise exemplar threshold of 0.7. Due to the fact that the sample size is smaller, lower thresholds can lead to feature sets much larger than the sample size. As well, for simplicity, we select only the best SNP from each exemplar gene instead of using the softmax distribution method used for the previous experiments.

4.4 Results

4.4.1 Classification Analyses

Table 4.1 displays the results for both the total aggregate analysis and the aggregated Halifax and Swedish samples using each of the four combinations of LR and XGB as feature selection model and inference model. The best performance on the total aggregate sample in terms of Cohen's kappa came from the combination of feature selection with LR and inference with LR achieving median kappa 0.07 (lower/upper decile (0.00, 0.17)). We also note that in both feature selection with LR and with XGB there were some of the ten cross-validation folds for which no genes were found to be exemplary and thus classification performance for these folds was not defined.

For the combined Halifax and Swedish samples we see that the combination of models that achieves the highest median value of kappa is XGB/XGB (kappa 0.17 (0.05, 0.27)), and that the median value is in excess of 0.1 for each other combination. The only selection/inference combination for which the lower bound of kappa fell below zero was LR/XGB (kappa 0.1 (-0.04, 0.29)). We note as well that there were no folds for which exemplary genes were not found for either feature selection method.

Table 4.1 shows the results for the site level analysis on both Halifax and Sweden. The maximum value of kappa for the Halifax sample was achieved using XGB for both inference and feature selection (0.26 (0.08, 0.41)) whereas the median value of kappa remained near zero in each scenario for the Swedish sample. We note however that the bounds on kappa for the Swedish sample span a wide range, and in all cases intersect with zero which indicates large disagreement between folds.

In Figure 4.1 we highlight both the differences in performance between

Table 4.1: Results for the aggregate and site level analyses using both the LR and XGB classifiers in the feature selection protocol. The values in each cell represents the median of the statistic as well as the lower and upper deciles across ten cross validation folds. *Abbreviations*: area under the curve (AUC) positive predictive value (PPV) negative predictive value (NPV).

Classifier	Site	Accuracy	AUC	Sensitivity	Specificity	PPV	NPV	Kappa
Feature Selection with LR								
LR	ALL	0.66 (0.62, 0.69)	0.58 (0.53, 0.62)	0.27 (0.09, 0.36)	0.81 (0.76, 0.92)	0.36 (0.27, 0.42)	0.73 (0.71, 0.76)	0.07 (0.00, 0.17)
XGB	ALL	0.71 (0.70, 0.72)	0.54 (0.50, 0.56)	0.05 (0.03, 0.05)	0.98 (0.97, 0.99)	0.50 (0.30, 0.72)	0.72 (0.71, 0.72)	0.02 (0.00, 0.06)
LR	Halifax/Sweden	0.57 (0.51, 0.62)	0.57 (0.48, 0.63)	0.53 (0.43, 0.56)	0.61 (0.57, 0.67)	0.52 (0.46, 0.58)	0.61 (0.56, 0.66)	0.14 (0.02, 0.25)
XGB	Halifax/Sweden	0.56 (0.49, 0.66)	0.60 (0.47, 0.69)	0.47 (0.38, 0.54)	0.62 (0.55, 0.84)	0.51 (0.43, 0.71)	0.59 (0.53, 0.66)	0.10 (-0.04, 0.29)
LR	Halifax	0.63 (0.55, 0.72)	0.67 (0.54, 0.78)	0.59 (0.49, 0.70)	0.72 (0.54, 0.79)	0.62 (0.50, 0.68)	0.64 (0.61, 0.75)	0.25 (0.11, 0.42)
LR	Sweden	0.50 (0.42, 0.61)	0.47 (0.39, 0.70)	0.43 (0.28, 0.59)	0.58 (0.47, 0.76)	0.45 (0.33, 0.60)	0.54 (0.47, 0.64)	-0.01 (-0.19, 0.20)
XGB	Halifax	0.59 (0.52, 0.73)	0.67 (0.58, 0.80)	0.53 (0.46, 0.62)	0.67 (0.49, 0.81)	0.55 (0.47, 0.73)	0.63 (0.58, 0.73)	0.18 (0.05, 0.44)
XGB	Sweden	0.47 (0.35, 0.73)	0.46 (0.31, 0.67)	0.29 (0.15, 0.70)	0.62 (0.46, 0.76)	0.37 (0.25, 0.69)	0.51 (0.42, 0.75)	-0.11 (-0.32, 0.44)
Feature Selection with XGB								
LR	ALL	0.71 (0.70, 0.71)	0.52 (0.52, 0.54)	0.02 (0.00, 0.03)	0.99 (0.98, 0.99)	0.27 (0.06, 0.57)	0.71 (0.71, 0.72)	-0.00 (-0.02, 0.03)
XGB	ALL	0.70 (0.69, 0.70)	0.52 (0.50, 0.54)	0.03 (0.00, 0.08)	0.97 (0.96, 0.97)	0.27 (0.05, 0.41)	0.71 (0.71, 0.72)	-0.00 (-0.04, 0.05)
LR	Halifax/Sweden	0.56 (0.53, 0.64)	0.57 (0.52, 0.60)	0.53 (0.36, 0.63)	0.57 (0.52, 0.78)	0.52 (0.48, 0.63)	0.60 (0.57, 0.65)	0.13 (0.06, 0.26)
XGB	Halifax/Sweden	0.60 (0.53, 0.64)	0.59 (0.47, 0.70)	0.43 (0.38, 0.57)	0.72 (0.64, 0.81)	0.57 (0.48, 0.67)	0.61 (0.57, 0.65)	0.17 (0.05, 0.27)
LR	Halifax	0.59 (0.47, 0.66)	0.58 (0.51, 0.72)	0.50 (0.37, 0.63)	0.62 (0.50, 0.79)	0.54 (0.41, 0.67)	0.61 (0.53, 0.67)	0.16 (-0.07, 0.30)
LR	Sweden	0.51 (0.39, 0.63)	0.52 (0.38, 0.62)	0.43 (0.28, 0.65)	0.53 (0.42, 0.66)	0.46 (0.33, 0.60)	0.54 (0.45, 0.64)	0.01 (-0.22, 0.26)
XGB	Halifax	0.63 (0.54, 0.72)	0.66 (0.59, 0.74)	0.56 (0.48, 0.69)	0.67 (0.57, 0.85)	0.59 (0.49, 0.76)	0.68 (0.58, 0.70)	0.26 (0.08, 0.41)
XGB	Sweden	0.48 (0.42, 0.55)	0.47 (0.41, 0.56)	0.39 (0.29, 0.55)	0.54 (0.43, 0.76)	0.44 (0.37, 0.52)	0.53 (0.47, 0.59)	-0.04 (-0.16, 0.08)

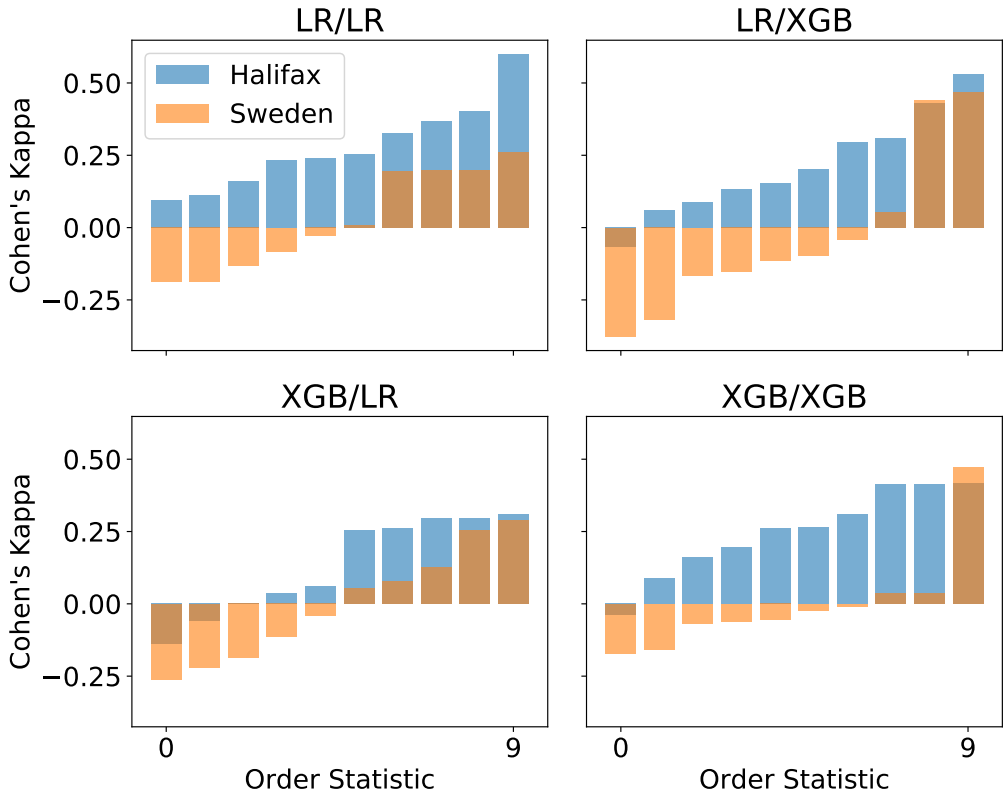

Figure 4.1: Bar plots showing the classification performance in terms of Cohen's kappa for the Halifax and Sweden sites. Each bar represents the kappa value for one fold, and bars are sorted in increasing order. Titles for each pane represent "selection model / inference model". Abbreviations: logistic regression (LR), XGBoost (XGB).

the Halifax and Swedish samples, and the differences each site has internally between folds. While we see notable imbalance in kappa across folds for the Halifax site, this imbalance is far more pronounced for Sweden where, under each combination of selection and inference model, at least half of the folds have negative kappa values.

In Table 4.2 we show the results of the leave one site out analysis for which at least one exemplary gene was found in the feature selection step. That is, all other combinations of site, selection, and inference models that are not shown failed to find a single exemplar gene in one or more folds. Only the

removal of the San Diego site's data in concert with LR feature selection and an LR inference model achieved above chance classification performance, with kappa 0.07 (0.02, 0.14) and a simulated p-value of $p = 0.0005$.

Table 4.3 shows the results for the predict one site out analysis for which the feature selection method did not fail to find any exemplar genes. Using LR for both feature selection and inference we see that a model trained on all sites but Geneva was able to predict the Geneva sample with kappa 0.4 ($p = 0.0002$) and sensitivity 0.62. This suggests the performance was not solely due to predicting non-response. Similarly, using XGB for both models, the classifier achieved kappa 0.19 ($p = 0.003$). In this case, however, the specificity (0.95) more significantly outweighed the sensitivity (0.2). Lastly, we note that every experiment for which the data from Barcelona, Halifax, or Poznan were omitted yielded no exemplary genes.

4.4.2 Exemplary Subject Analyses

Figure 4.2 shows the kappa values for each fold sorted in increasing order for each model combination for both the top and bottom 50% of samples according to their subject-wise (clinical) exemplar score. Both datasets achieved maximal classification accuracy when using XGB for both feature selection and inference with kappa 0.62 (0.28, 0.73) for the top 50% of exemplars and 0.23 (-0.07, 0.37) for the bottom 50%. In the case of feature selection with the LR classifier, there were four folds for which no genes met the exemplar threshold in the bottom 50% dataset. For the full set of results for this experiment including the remaining metrics, see Table B.2.

Table 4.2: Results for the leave one site out analysis for which exemplary genes were found in every cross validation fold. *Abbreviations*: area under the curve (AUC) positive predictive value (PPV) negative predictive value (NPV).

Classifier	Centre	Accuracy	AUC	Sensitivity	Specificity	NPV	PPV	Kappa
Feature Selection with LR								
LR	Halifax	0.69 (0.65, 0.74)	0.52 (0.47, 0.55)	0.13 (0.08, 0.23)	0.89 (0.82, 0.95)	0.27 (0.19, 0.47)	0.75 (0.73, 0.76)	0.01 (-0.05, 0.09)
LR	Japan	0.65 (0.62, 0.70)	0.53 (0.50, 0.56)	0.14 (0.03, 0.25)	0.86 (0.77, 0.92)	0.28 (0.16, 0.34)	0.71 (0.70, 0.72)	-0.00 (-0.08, 0.03)
LR	Paris	0.64 (0.60, 0.66)	0.52 (0.47, 0.55)	0.20 (0.14, 0.31)	0.81 (0.76, 0.85)	0.32 (0.25, 0.40)	0.71 (0.69, 0.73)	0.03 (-0.05, 0.11)
LR	Romania	0.62 (0.58, 0.69)	0.51 (0.46, 0.58)	0.22 (0.16, 0.30)	0.78 (0.73, 0.89)	0.31 (0.23, 0.44)	0.71 (0.69, 0.73)	0.02 (-0.07, 0.10)
LR	San Diego	0.60 (0.56, 0.64)	0.54 (0.49, 0.56)	0.29 (0.21, 0.34)	0.76 (0.69, 0.80)	0.31 (0.28, 0.40)	0.70 (0.68, 0.72)	0.01 (-0.03, 0.12)
LR	Würzburg	0.63 (0.59, 0.65)	0.53 (0.49, 0.59)	0.24 (0.11, 0.36)	0.78 (0.73, 0.88)	0.33 (0.27, 0.37)	0.71 (0.70, 0.73)	0.04 (-0.02, 0.09)
XGB	San Diego	0.68 (0.67, 0.71)	0.56 (0.50, 0.61)	0.12 (0.06, 0.18)	0.94 (0.92, 0.96)	0.44 (0.35, 0.61)	0.71 (0.70, 0.72)	0.07 (0.02, 0.14)
XGB	Würzburg	0.69 (0.69, 0.70)	0.50 (0.47, 0.58)	0.05 (0.02, 0.08)	0.97 (0.95, 0.98)	0.38 (0.25, 0.51)	0.71 (0.70, 0.71)	0.02 (-0.01, 0.05)
XGB	Romania	0.69 (0.67, 0.71)	0.50 (0.45, 0.54)	0.02 (0.01, 0.07)	0.96 (0.94, 0.99)	0.24 (0.09, 0.45)	0.71 (0.70, 0.71)	0.00 (-0.05, 0.04)
XGB	Halifax	0.74 (0.73, 0.75)	0.50 (0.45, 0.55)	0.04 (0.00, 0.07)	0.99 (0.97, 0.99)	0.50 (0.00, 0.75)	0.75 (0.74, 0.75)	0.03 (-0.01, 0.08)
XGB	Japan	0.71 (0.68, 0.71)	0.49 (0.43, 0.54)	0.04 (0.02, 0.07)	0.97 (0.96, 0.99)	0.44 (0.14, 0.55)	0.72 (0.70, 0.72)	0.04 (-0.03, 0.07)
Feature Selection with XGB								
LR	San Diego	0.62 (0.57, 0.65)	0.53 (0.48, 0.56)	0.23 (0.14, 0.33)	0.79 (0.74, 0.86)	0.31 (0.29, 0.38)	0.69 (0.69, 0.71)	0.01 (-0.02, 0.08)
XGB	San Diego	0.67 (0.65, 0.68)	0.55 (0.50, 0.58)	0.07 (0.05, 0.15)	0.92 (0.90, 0.95)	0.33 (0.26, 0.41)	0.70 (0.69, 0.70)	0.01 (-0.02, 0.06)

Table 4.3: Results for the predict one site out analysis which did not fail to find any exemplar genes in the feature selection step. *Abbreviations*: area under the curve (AUC) positive predictive value (PPV) negative predictive value (NPV).

Classifier	Site	Accuracy	AUC	Sensitivity	Specificity	PPV	NPV	Kappa
Feature Selection with LR								
LR	Cagliari	0.63	0.59	0.20	0.80	0.28	0.72	0.00
XGB	Cagliari	0.66	0.52	0.05	0.90	0.18	0.71	-0.06
LR	Geneva	0.77	0.63	0.62	0.82	0.50	0.88	0.40
XGB	Geneva	0.75	0.42	0.00	0.98	0.00	0.77	-0.03
LR	Japan	0.66	0.58	0.23	0.80	0.27	0.76	0.03
XGB	Japan	0.76	0.43	0.00	1.00	0.00	0.76	0.00
LR	Mayo	0.69	0.52	0.23	0.83	0.29	0.78	0.07
XGB	Mayo	0.77	0.55	0.00	1.00	0.00	0.77	0.00
LR	Paris	0.62	0.45	0.18	0.72	0.13	0.80	-0.08
XGB	Paris	0.82	0.47	0.03	0.99	0.50	0.82	0.03
LR	Romania	0.73	0.54	0.19	0.88	0.29	0.80	0.07
XGB	Romania	0.78	0.48	0.03	0.98	0.33	0.79	0.02
LR	San Diego	0.60	0.51	0.30	0.64	0.09	0.88	-0.03
XGB	San Diego	0.74	0.50	0.17	0.81	0.10	0.89	-0.01
LR	Sweden	0.53	0.53	0.07	0.92	0.39	0.54	-0.02
XGB	Sweden	0.55	0.52	0.01	1.00	1.00	0.55	0.01
LR	Sydney	0.78	0.55	0.08	0.96	0.33	0.80	0.05
XGB	Sydney	0.81	0.62	0.08	1.00	1.00	0.81	0.12
LR	Taiwan	0.80	0.61	0.15	0.91	0.22	0.87	0.07
XGB	Taiwan	0.86	0.39	0.00	1.00	0.00	0.86	0.00
Feature Selection with XGB								
LR	Cagliari	0.69	0.54	0.16	0.90	0.39	0.73	0.08
XGB	Cagliari	0.74	0.56	0.20	0.95	0.61	0.75	0.19
LR	Japan	0.77	0.54	0.06	0.99	0.67	0.77	0.08
XGB	Japan	0.75	0.48	0.00	0.99	0.00	0.76	-0.02
LR	Mayo	0.77	0.52	0.00	1.00	0.00	0.77	0.00
XGB	Mayo	0.74	0.57	0.00	0.97	0.00	0.76	-0.04
LR	Paris	0.81	0.53	0.00	0.99	0.00	0.82	-0.01
XGB	Paris	0.79	0.52	0.03	0.95	0.11	0.82	-0.03
LR	San Diego	0.58	0.46	0.30	0.61	0.09	0.88	-0.04
XGB	San Diego	0.68	0.52	0.30	0.73	0.12	0.90	0.02
LR	Sydney	0.79	0.49	0.00	1.00	0.00	0.79	0.00
XGB	Sydney	0.78	0.38	0.00	0.98	0.00	0.79	-0.03
LR	Taiwan	0.84	0.63	0.00	0.97	0.00	0.86	-0.04
XGB	Taiwan	0.86	0.62	0.00	1.00	0.00	0.86	0.00
LR	wuerzburg	0.81	0.51	0.00	0.97	0.00	0.82	-0.04
XGB	wuerzburg	0.79	0.41	0.03	0.95	0.13	0.83	-0.02

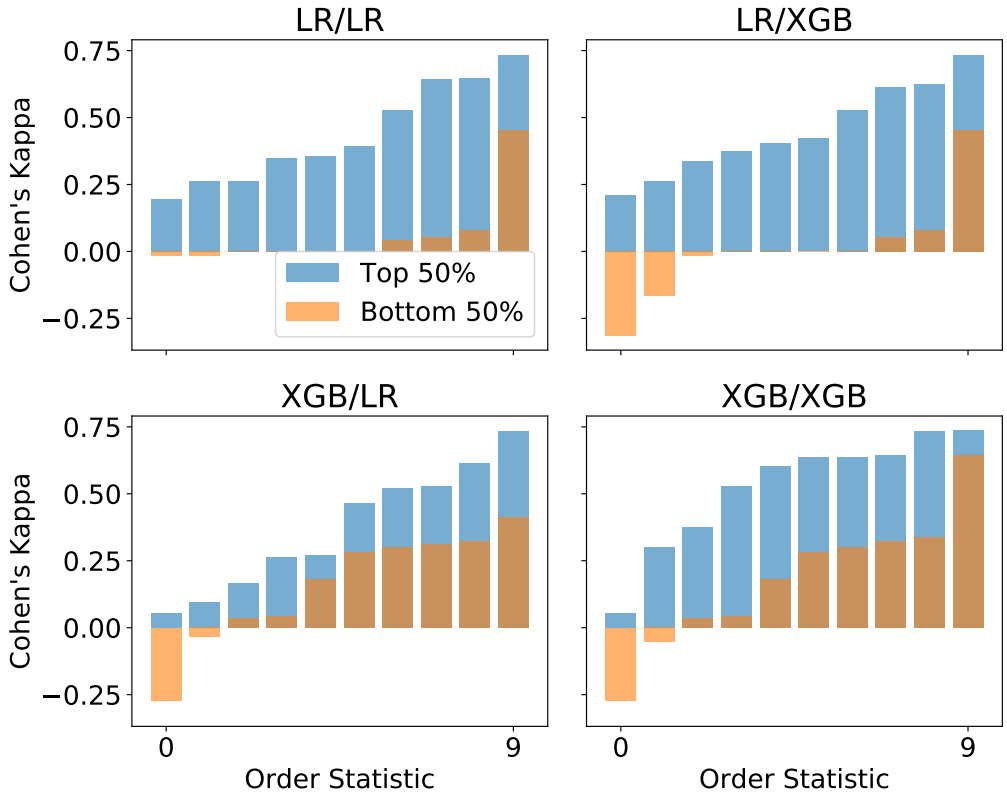

Figure 4.2: Bar plots showing the classification performance in terms of Cohen's kappa for the top 50th percentile along with the bottom 50th percentile of subjects by (subject-wise) exemplar score from the Halifax site. Each bar represents the kappa value for one fold, and bars are sorted in increasing order. Titles for each pane represent "selection model / inference model". Abbreviations: logistic regression (LR), XGBoost (XGB).

4.5 Discussion

Driven by the limitations of the naive feature selection method observed in Chapter 3, we have created a novel method of feature selection that is capable of reducing the feature space so that the number of subjects and features are roughly equivalent. Our method is roughly the equivalent of a wrapper method, but uses the inherent biological structure of genetic data in order to create candidate feature sets which can be tested individually. This has the advantage of being computationally tractable, biologically interpretable, and also accounts for multi-variate effects. In this section we discuss the results of a basic classification study that uses this method, the importance of how our method performs on both more and less clinically exemplary subjects, and some of the benefits and limitations of the method.

In our basic classification study we found no meaningful prediction of lithium response in the aggregate sample. While in one case the median value of kappa was above zero along with a notable increase in sensitivity, selecting features with both LR and XGB on the aggregate sample had some cross validation folds for which no genes exceeded the gene-wise exemplar threshold. In cases where no exemplar genes are identified, we argue that classification performance is undefined, as this scenario is notably different than a classifier that simply predicts each sample incorrectly. We therefore cannot say that our method generalized in this case given that there were folds for which it failed to produce a feature set.

We found it was possible to classify subjects in the Halifax sample above random chance although there was some imbalance in performance across folds. When all folds were taken into account, the classification performance on the Swedish sample was trivial. However, upon closer inspection of the Swedish

sample results, we saw an even more extreme imbalance in performance across folds with some achieving a kappa score as high as 0.45 while many others obtained negative scores. This effect may be due to within site heterogeneity, with some samples being more easily classifiable than others.

Only the omission of one site (San Diego) caused there to be a statistically significant prediction of lithium response in the leave one site out analysis. However, we see what may be a meaningful pattern by looking at which sites for which our method always or never fails. The omission of San Diego never affects the ability of either classifier to find exemplar genes in the remaining data, regardless of which feature selection model is used. Conversely, when the Geneva, Mayo Clinic, Cagliari, Swedish, and Sydney samples are omitted we are never able to find exemplar genes in all ten cross validation folds. This could be an indication that certain sites are particularly rich in signal, whereas others have a particularly detrimental effect when added to the sample.

In the predict one site out analysis, there were only two sites (Geneva and Cagliari) whose signal could be identified using the rest of the data above random chance. Similarly to the leave one site out analysis however, there are patterns in the inability to discover exemplar genes. Specifically, under no combination of models was our method able to discover exemplar genes in the remaining data when the samples from Barcelona, Halifax, or Poznan were omitted.

Nunes et al. proposed a method by which a sample could be rated by the degree to which it's clinical features are representative of it's phenotype across different data collection sites. Their metric showed the existence of subjects that are both more and less "exemplary" of their phenotype, and that more exemplary subjects are over-represented within the Halifax site. Using

samples that overlapped between the clinical and genetic datasets, we separated subjects into two equally sized groups with lower and higher exemplar scores and applied our feature selection method to each. This analysis showed a significantly stronger ability to genetically classify subjects that were more clinically exemplary. This result provides strong evidence that limitations in genetic classification performance could be related to heterogeneity in clinical features. Moreover, the presence of a significantly more classifiable group in the Halifax sample explains the improved accuracy we have observed in comparison to other groups.

One benefit of our method comes from the fact that, provided there are genes with predictive power present, the exemplar gene threshold and upper bound on the number of features provides fine control over the feature space dimension in an interpretable way. Also, our method maintains a degree of biological interpretability that other techniques, such as embedding methods, lack. For example, even if we were to use a gene selection model that lacks an interpretable feature importance metric (e.g. multi-layer perceptron) we can still see which genes have been selected and whether or not their selection is consistent across folds. Lastly, our method has a clear advantage over the most common feature selection method in genomics, logistic regression analysis, in that we are able to take into account multi-variate effects between SNPs.

A notable limitation of the gene-wise selection method is that, due to the necessity of nested cross validation, it requires there to be a relatively large number of subjects. As well, our method largely ignores intergenic SNPs that are not within 50,000 base pairs of a gene. Lastly, our method may be missing some gene-gene interactions due to the fact that we are only selecting the (relatively few) SNPs from each gene with the largest feature importances. It is conceivable that there are cases in which SNPs that are not among the most

important in two separate genes could interact with each other in a stronger way.

Chapter 5

Discussion and Future Work

Genetic prediction of complex phenotypes is an important area of research as it could lead to novel methods of treatment or pharmaceutical intervention for various diseases. The application of machine learning is a fast growing trend in this field, however effective feature selection is a significant barrier to robust and interpretable phenotype classification. In Chapter 3 we show that above chance classification of lithium response in subjects with bipolar disorder is possible using the largest micro-array dataset in existence that addresses this phenotype. We then introduce a novel method that we call gene-wise feature selection in Chapter 4 and demonstrate its benefits. These results are notable because this is the first work in which lithium response has been predicted with genetic data alone, and our gene-wise selection method addresses many concerns that are present with other techniques that are commonly applied to micro-array data.

In this section we discuss the moderate success of the naive feature selection method applied in Chapter 3, how our gene-wise selection method addresses the drawbacks of other feature selection methods, and how we have shown a significant difference in the predictability of more and less clinically exemplary subjects.

In Chapter 3 we attempted to predict lithium response using both logistic regression (LR) and XGBoost (XGB) classifiers on the non-imputed dataset. The non-imputed dataset consisted of features that were measured in common

by every data collection centre, and represented a relatively even sampling of SNPs across all chromosomes. This analysis showed that, while prediction was not possible on the entire sample, the classifiers we able to predict lithium response above random chance on the samples that came from the Halifax and Würzburg sites. To back up these analyses, we used the feature importance metrics of the classifiers trained on Halifax and Würzburg to extract the most informative SNPs and used these in a gene-set analysis using the PANTHER gene ontology tool [19]. This analysis showed that the set of most important SNPs were statistically over-represented in genes that have previously (and independently) been associated with bipolar disorder, which further confirms that our prediction was indeed not random.

Although above chance prediction was found to be possible in Chapter 3, we also found that both the LR and XGB classifiers achieved perfect (or near perfect) accuracy when applied to the training sets. This indicated the need for further feature selection. However, as outlined in Chapter 2, methods of feature selection that are commonly applied to micro-array data have significant drawbacks. In Chapter 4 we detail our gene-wise feature selection method which addresses these drawbacks, and apply it to the lithium response dataset. We show that, by first filtering out genes that lack reliable predictive capacity, we are able to obtain feature sets of the desired size by taking only the most important SNPs from the selected genes. Moreover, the selected feature sets are able to classify lithium response in the Halifax dataset with higher accuracy than with the non-imputed dataset.

The work of Chapter 4 also identified a significantly more classifiable subset of samples from the Halifax site. These samples are those for which Nunes et al. have shown are more *clinically* representative of their phenotypes (both response and non-response). We note that, while the work of Nunes et al

included some of the same subjects as are in the genetic dataset, the measurement of the "exemplaryness" of these subjects was done using entirely non-genetic features. As the clinical dataset only shared overlap with the Halifax sample from the genetic dataset, we were unable to perform this analysis on any other subjects. However, a general finding in the work of Nunes et al. was that Halifax contained significantly more exemplars than other sites. The genetic classifiability of clinical exemplars may therefore be responsible for the improved performance in the Halifax dataset in comparison to other sites.

A main limitation that we observed in applying the gene-wise feature selection method in this work was the requirement that there be enough samples to perform nested cross-validation. With increased access and incentive for micro-array studies however, this sample requirement will not always be a limiting factor for other datasets. Another limitation is the fact that we select a relatively sparse set of SNPs from each exemplar gene, and thus potentially left out important interactions. This could be solved however with a more sophisticated method of SNP selection, for example we could apply an embedding method such as an autoencoder to learn a fixed size representation for every exemplary gene that would include information for all SNPs within the gene.

In this work we have provided a background on the problem genetic prediction of phenotypes using machine learning, including the drawbacks of some commonly used methods. We also showed that classification of lithium response in subjects with bipolar disorder is possible with genetic data. Inspired by the hierarchical nature of genetic data, we have introduced a gene-wise feature selection method capable of significantly reducing the genetic feature space in a biologically interpretable way. We note several limitations of the

gene-wise feature selection method, however we believe the more serious limitations are not unsolvable. Namely, we suggest that the adoption of gene-wise embeddings, which would capture information from every SNP within a gene, could alleviate the problem that comes from the sparse coverage of each exemplar gene. We also note that we have only applied our method to the subset of our dataset that is in the GPCR pathway for reasons of clinical interest. In future work it would be possible to apply the method to the entire dataset. Lastly, we have shown that heterogeneity in clinical variables can be strongly correlated with genetic classification performance, and thus suggest that future genetic data collection for the lithium response phenotype be coupled with collection of clinical data.

Appendix A

Supplementary Materials: Naive Approach Paper

Table A.4: The list of genes contained within the combined Halifax-Würzburg effect set that are over-represented in the postsynaptic membrane functional class along with their associated PANTHER protein class.

Gene ID	**PANTHER Protein Class**	**Gene ID**	**PANTHER Protein Class**
LRFN2		ANK3	
RGS9		PTPRT	protein phosphatase (PC00195); receptor(PC00197)
DLG1	transmembrane receptor regulatory/adaptor protein (PC00226)	ADCY8	
CTNNA2	cell adhesion molecule (PC00069); non-motor actin binding protein (PC00165)	SEMA4F	membrane-bound signaling molecule (PC00152)

Continuation of Table A.4			
Gene ID	**PANTHER Protein Class**	**Gene ID**	**PANTHER Protein Class**
GRIK1		GABRR1	GABA receptor (PC00023); acetylcholine receptor (PC00037)
CADPS2	calcium-binding protein (PC00060)	LRRC4C	
GRIP2		SLC6A11	cation transporter (PC00068)
NTRK3		ADGRB1	G-protein coupled receptor (PC00021); antibacterial response protein (PC00051); protease (PC00190)
DCC		SHISA6	
CDH9		SLC8A3	
CHRM3	G-protein coupled receptor (PC00021)	KCNH1	

Continuation of Table A.4			
Gene ID	**PANTHER Protein Class**	**Gene ID**	**PANTHER Protein Class**
LRRC7		GABRB1	GABA receptor (PC00023); acetylcholine receptor (PC00037)
SYNDIG1		GRIK2	
NEURL1	ubiquitin-protein ligase (PC00234)	SLC8A1	
ANKS1B	transmembrane receptor regulatory/adaptor protein (PC00226)	ATAD1	
SLC6A6	cation transporter (PC00068)	SORCS3	receptor (PC00197); transporter (PC00227)
KCNB1		GABRG2	GABA receptor (PC00023); acetylcholine receptor (PC00037)
DGKI	kinase (PC00137)	GRID1	
IGSF21		HOMER1	
DISC1		BAALC	
CNTN1		ANK1	

Continuation of Table A.4			
Gene ID	**PANTHER Protein Class**	**Gene ID**	**PANTHER Protein Class**
GABRR2	GABA receptor (PC00023); acetylcholine receptor (PC00037)	FARP1	
LIN7A	cell adhesion molecule (PC00069); cell junction protein (PC00070)	GRIA1	
GRIN2B		CPEB4	mRNA polyadenylation factor (PC00146)
ACTN2		GABRB2	GABA receptor (PC00023); acetylcholine receptor (PC00037)
DRD3	G-protein coupled receptor (PC00021)	PSD3	

Continuation of Table A.4			
Gene ID	**PANTHER Protein Class**	**Gene ID**	**PANTHER Protein Class**
TENM2		DNM3	hydrolase (PC00121); microtubule family cytoskeletal protein (PC00157); small GTPase (PC00208)
NRP1		GABRG3	GABA receptor (PC00023); acetylcholine receptor (PC00037)
CDH10		GABBR2	
RGS7BP		GRIK2	
KCNC2		GRIN2A	
EPHA4		PTPRO	
NRCAM		CLSTN2	calcium-binding protein (PC00060); cell adhesion molecule (PC00069)
GRM5	G-protein coupled receptor (PC00021)	GRIK4	

Continuation of Table A.4			
Gene ID	**PANTHER Protein Class**	**Gene ID**	**PANTHER Protein Class**
GSG1L	cytoskeletal protein (PC00085)	ARHGAP32	
GRID2		GRIP1	
ADRA1A	G-protein coupled receptor (PC00021)	LRRC4B	
LZTS1		DLG2	transmembrane receptor regulatory/adaptor protein (PC00226)
KCNMA1		SHANK2	
SRGAP2	G-protein modulator (PC00022)	SYNE1	
DLGAP1	transmembrane receptor regulatory/adaptor protein (PC00226)	CACNG4	voltage-gated calcium channel (PC00240)
SHISA9		GRM7	G-protein coupled receptor (PC00021)
ANK2		TMEM108	

Continuation of Table A.4			
Gene ID	**PANTHER Protein Class**	**Gene ID**	**PANTHER Protein Class**
NRG1	growth factor (PC00112)	DLGAP2	transmembrane receptor regulatory/adaptor protein (PC00226)
CACNA1C		NRP2	
OPRM1	G-protein coupled receptor (PC00021)	TIAM1	
LRFN5		GPHN	
OPRD1	G-protein coupled receptor (PC00021)	ATP2B2	cation transporter (PC00068); hydrolase (PC00121); ion channel (PC00133)
ERBB4			
End of Table			

Table A.1: Remaining site-level results for the LR and XGB classifiers. Each cell shows the mean value of the statistic over five folds along with a 95% confidence interval. *Abbreviations*: area under the curve (AUC) positive predictive value (PPV) negative predictive value (NPV).

Site	Accuracy	AUC	Sensitivity	Specificity	PPV	NPV	Kappa
Logistic Regression							
Barcelona	0.73 (0.72, 0.74)	0.4 (0.26, 0.53)	0.0 (0.0, 0.0)	1.0 (1.0, 1.0)	0.73 (0.72, 0.74)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)
Cagliari	0.72 (0.72, 0.72)	0.49 (0.46, 0.52)	0.0 (0.0, 0.0)	1.0 (1.0, 1.0)	0.72 (0.72, 0.72)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)
Geneva	0.77 (0.74, 0.8)	0.49 (0.28, 0.7)	0.0 (0.0, 0.0)	1.0 (1.0, 1.0)	0.77 (0.74, 0.8)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)
Japan	0.76 (0.75, 0.77)	0.6 (0.45, 0.75)	0.0 (0.0, 0.0)	1.0 (1.0, 1.0)	0.76 (0.75, 0.77)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)
Mayo	0.77 (0.75, 0.78)	0.32 (0.21, 0.44)	0.0 (0.0, 0.0)	1.0 (1.0, 1.0)	0.77 (0.75, 0.78)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)
Paris	0.82 (0.81, 0.83)	0.4 (0.29, 0.51)	0.0 (0.0, 0.0)	1.0 (1.0, 1.0)	0.82 (0.81, 0.83)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)
Poznan	0.51 (0.39, 0.62)	0.62 (0.44, 0.8)	0.24 (0.03, 0.45)	0.76 (0.6, 0.92)	0.52 (0.44, 0.61)	0.36 (0.08, 0.64)	0.0 (-0.23, 0.23)
Romania	0.79 (0.78, 0.8)	0.57 (0.51, 0.64)	0.0 (0.0, 0.0)	1.0 (1.0, 1.0)	0.79 (0.78, 0.8)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)
San Diego	0.89 (0.88, 0.9)	0.54 (0.5, 0.58)	0.0 (0.0, 0.0)	1.0 (1.0, 1.0)	0.89 (0.88, 0.9)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)
Sweden	0.55 (0.54, 0.56)	0.52 (0.46, 0.58)	0.15 (0.06, 0.25)	0.88 (0.81, 0.94)	0.56 (0.54, 0.57)	0.49 (0.41, 0.58)	0.03 (-0.0, 0.07)
Sydney	0.78 (0.75, 0.81)	0.32 (0.18, 0.47)	0.0 (0.0, 0.0)	0.98 (0.94, 1.0)	0.79 (0.76, 0.82)	0.0 (0.0, 0.0)	-0.03 (-0.07, 0.02)
Taiwan	0.86 (0.84, 0.88)	0.49 (0.29, 0.69)	0.0 (0.0, 0.0)	1.0 (1.0, 1.0)	0.86 (0.84, 0.88)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)
XGBoost							
Barcelona	0.69 (0.65, 0.73)	0.52 (0.42, 0.61)	0.05 (0.0, 0.15)	0.93 (0.89, 0.96)	0.72 (0.7, 0.75)	0.1 (0.0, 0.3)	-0.04 (-0.16, 0.08)
Cagliari	0.71 (0.7, 0.73)	0.53 (0.46, 0.59)	0.02 (0.0, 0.05)	0.99 (0.96, 1.0)	0.72 (0.72, 0.72)	0.07 (0.0, 0.2)	0.0 (-0.0, 0.01)
Geneva	0.74 (0.68, 0.8)	0.8 (0.65, 0.94)	0.0 (0.0, 0.0)	0.96 (0.9, 1.0)	0.76 (0.73, 0.8)	0.0 (0.0, 0.0)	-0.06 (-0.13, 0.01)
Japan	0.73 (0.71, 0.76)	0.42 (0.33, 0.51)	0.0 (0.0, 0.0)	0.97 (0.94, 0.99)	0.75 (0.74, 0.76)	0.0 (0.0, 0.0)	-0.04 (-0.08, -0.01)
Mayo	0.78 (0.76, 0.79)	0.71 (0.56, 0.85)	0.04 (0.0, 0.12)	1.0 (1.0, 1.0)	0.77 (0.76, 0.79)	0.2 (0.0, 0.59)	0.05 (-0.05, 0.16)
Paris	0.82 (0.81, 0.83)	0.52 (0.43, 0.6)	0.0 (0.0, 0.0)	1.0 (1.0, 1.0)	0.82 (0.81, 0.83)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)
Poznan	0.61 (0.5, 0.72)	0.6 (0.5, 0.7)	0.51 (0.34, 0.69)	0.7 (0.58, 0.8)	0.61 (0.51, 0.72)	0.62 (0.5, 0.75)	0.21 (0.0, 0.43)
Romania	0.79 (0.78, 0.8)	0.47 (0.35, 0.6)	0.0 (0.0, 0.0)	1.0 (1.0, 1.0)	0.79 (0.78, 0.8)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)
San Diego	0.89 (0.88, 0.9)	0.44 (0.38, 0.5)	0.0 (0.0, 0.0)	1.0 (1.0, 1.0)	0.89 (0.88, 0.9)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)
Sweden	0.52 (0.47, 0.57)	0.52 (0.48, 0.56)	0.33 (0.22, 0.44)	0.68 (0.63, 0.73)	0.55 (0.51, 0.6)	0.45 (0.37, 0.52)	0.01 (-0.1, 0.12)
Sydney	0.78 (0.73, 0.83)	0.41 (0.22, 0.59)	0.0 (0.0, 0.0)	0.98 (0.94, 1.0)	0.79 (0.76, 0.83)	0.0 (0.0, 0.0)	-0.03 (-0.08, 0.03)
Taiwan	0.86 (0.84, 0.88)	0.54 (0.33, 0.75)	0.0 (0.0, 0.0)	1.0 (1.0, 1.0)	0.86 (0.84, 0.88)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)

Table A.2: Remaining leave one site out results for the LR classifier, and the entire result set for the XGB classifier. Each cell shows the mean value of the statistic over five folds along with a 95% confidence interval. *Abbreviations*: area under the curve (AUC) positive predictive value (PPV) negative predictive value (NPV).

Centre	Accuracy	AUC	Sensitivity	Specificity	NPV	PPV	Kappa
Logistic Regression							
Cagliari	0.72 (0.71, 0.73)	0.59 (0.57, 0.61)	0.04 (0.01, 0.08)	0.99 (0.98, 0.99)	0.72 (0.71, 0.73)	0.51 (0.22, 0.8)	0.05 (0.01, 0.09)
Geneva	0.71 (0.71, 0.72)	0.59 (0.57, 0.6)	0.05 (0.04, 0.05)	0.98 (0.98, 0.99)	0.72 (0.72, 0.72)	0.55 (0.47, 0.63)	0.04 (0.03, 0.05)
Halifax	0.75 (0.74, 0.75)	0.59 (0.56, 0.62)	0.01 (0, 0.03)	1 (1, 1)	0.75 (0.74, 0.75)	0.7 (0.31, 1)	0.02 (0, 0.04)
Japan	0.71 (0.71, 0.72)	0.59 (0.56, 0.62)	0.05 (0.03, 0.07)	0.98 (0.98, 0.99)	0.72 (0.71, 0.72)	0.54 (0.44, 0.63)	0.04 (0.02, 0.07)
Mayo	0.71 (0.71, 0.72)	0.58 (0.57, 0.6)	0.05 (0.04, 0.06)	0.98 (0.97, 0.99)	0.72 (0.72, 0.72)	0.51 (0.44, 0.59)	0.04 (0.03, 0.05)
Paris	0.7 (0.7, 0.71)	0.59 (0.57, 0.61)	0.04 (0.03, 0.05)	0.98 (0.98, 0.99)	0.71 (0.71, 0.71)	0.52 (0.43, 0.61)	0.03 (0.02, 0.05)
Poznan	0.73 (0.72, 0.73)	0.58 (0.57, 0.59)	0.04 (0.02, 0.05)	0.99 (0.99, 1)	0.73 (0.73, 0.73)	0.61 (0.51, 0.71)	0.04 (0.02, 0.06)
Sweden	0.74 (0.74, 0.75)	0.54 (0.51, 0.57)	0.03 (0.01, 0.04)	1 (0.99, 1)	0.74 (0.74, 0.75)	0.55 (0.27, 0.82)	0.03 (0.01, 0.05)
Sydney	0.71 (0.71, 0.72)	0.59 (0.57, 0.61)	0.04 (0.04, 0.05)	0.98 (0.98, 0.99)	0.72 (0.71, 0.72)	0.52 (0.41, 0.63)	0.04 (0.02, 0.05)
Taiwan	0.71 (0.7, 0.71)	0.59 (0.58, 0.59)	0.05 (0.03, 0.06)	0.98 (0.97, 0.98)	0.71 (0.71, 0.72)	0.48 (0.39, 0.57)	0.04 (0.01, 0.06)
XGBoost							
Barcelona	0.71 (0.7, 0.71)	0.56 (0.55, 0.56)	0.03 (0.01, 0.05)	0.98 (0.97, 0.99)	0.71 (0.71, 0.72)	0.29 (0.11, 0.46)	0.01 (-0.01, 0.03)
Cagliari	0.7 (0.69, 0.7)	0.56 (0.54, 0.57)	0.03 (0.01, 0.05)	0.97 (0.96, 0.98)	0.71 (0.71, 0.71)	0.28 (0.22, 0.33)	0 (-0.01, 0.01)
Geneva	0.71 (0.7, 0.72)	0.55 (0.52, 0.59)	0.04 (0.02, 0.05)	0.98 (0.97, 0.99)	0.72 (0.71, 0.72)	0.43 (0.29, 0.58)	0.03 (0, 0.05)
Halifax	0.75 (0.74, 0.75)	0.55 (0.52, 0.58)	0.02 (0.01, 0.02)	1 (0.99, 1)	0.75 (0.75, 0.75)	0.63 (0.42, 0.85)	0.02 (0.01, 0.03)
Japan	0.71 (0.7, 0.71)	0.58 (0.55, 0.6)	0.04 (0.02, 0.05)	0.98 (0.98, 0.99)	0.71 (0.71, 0.72)	0.42 (0.28, 0.57)	0.02 (0, 0.05)
Mayo	0.71 (0.7, 0.72)	0.57 (0.54, 0.59)	0.04 (0.01, 0.07)	0.98 (0.97, 0.99)	0.72 (0.71, 0.72)	0.42 (0.2, 0.63)	0.03 (-0.01, 0.07)
Paris	0.69 (0.69, 0.7)	0.55 (0.53, 0.57)	0.04 (0.03, 0.06)	0.97 (0.97, 0.98)	0.7 (0.7, 0.71)	0.37 (0.24, 0.5)	0.02 (-0.01, 0.04)
Poznan	0.72 (0.72, 0.72)	0.54 (0.53, 0.56)	0.03 (0.03, 0.04)	0.98 (0.98, 0.99)	0.73 (0.72, 0.73)	0.47 (0.4, 0.54)	0.03 (0.02, 0.03)
Romania	0.7 (0.7, 0.71)	0.56 (0.53, 0.6)	0.03 (0.02, 0.04)	0.98 (0.97, 0.99)	0.71 (0.71, 0.71)	0.45 (0.33, 0.57)	0.02 (0, 0.04)
San Diego	0.69 (0.68, 0.7)	0.58 (0.55, 0.61)	0.05 (0.03, 0.07)	0.97 (0.97, 0.98)	0.7 (0.7, 0.7)	0.47 (0.35, 0.59)	0.03 (0.01, 0.06)
Sweden	0.74 (0.73, 0.75)	0.57 (0.55, 0.6)	0.03 (0.01, 0.05)	0.99 (0.98, 1)	0.74 (0.74, 0.75)	0.54 (0.24, 0.83)	0.03 (0, 0.06)
Sydney	0.71 (0.7, 0.71)	0.54 (0.51, 0.56)	0.04 (0.02, 0.05)	0.98 (0.97, 0.99)	0.71 (0.71, 0.72)	0.42 (0.3, 0.53)	0.02 (0, 0.05)
Taiwan	0.7 (0.7, 0.71)	0.54 (0.53, 0.55)	0.03 (0.02, 0.04)	0.98 (0.98, 0.99)	0.71 (0.71, 0.71)	0.44 (0.31, 0.58)	0.02 (0, 0.04)
Würzburg	0.7 (0.69, 0.7)	0.57 (0.53, 0.6)	0.04 (0.02, 0.05)	0.97 (0.97, 0.98)	0.71 (0.7, 0.71)	0.39 (0.27, 0.5)	0.02 (-0.01, 0.04)

Table A.3: Results for the predict one site out analysis for both the LR and XGB classifiers. *Abbreviations*: area under the curve (AUC) positive predictive value (PPV) negative predictive value (NPV).

Centre	Accuracy	AUC	Sensitivity	Specificity	NPV	PPV	Kappa
XGBoost							
Barcelona	0.73	0.56	0	1	0.73	0	0
Cagliari	0.72	0.6	0	1	0.72	0	0
Geneva	0.77	0.53	0.08	0.98	0.78	0.5	0.08
Halifax	0.55	0.56	0.01	0.99	0.55	0.5	0
Japan	0.76	0.52	0	1	0.76	0	0
Mayo	0.79	0.58	0.09	1	0.78	1	0.13
Paris	0.82	0.5	0.03	1	0.82	1	0.04
Poznan	0.52	0.59	0.02	0.98	0.52	0.5	0
Romania	0.78	0.53	0.06	0.97	0.8	0.4	0.05
San Diego	0.87	0.5	0	0.97	0.89	0	-0.05
Sweden	0.57	0.57	0.04	1	0.56	1	0.05
Sydney	0.79	0.41	0	1	0.79	0	0
Taiwan	0.86	0.35	0	1	0.86	0	0
Würzburg	0.79	0.44	0.03	0.95	0.83	0.12	-0.02
XGBoost							
Barcelona	0.73	0.56	0	1	0.73	0	0
Cagliari	0.71	0.42	0	0.99	0.72	0	-0.01
Geneva	0.74	0.53	0	0.95	0.76	0	-0.06
Halifax	0.55	0.49	0.01	1	0.55	1	0.01
Japan	0.76	0.6	0	1	0.76	0	0
Mayo	0.76	0.51	0	0.99	0.76	0	-0.02
Paris	0.83	0.5	0.08	1	0.83	1	0.12
Poznan	0.49	0.49	0.02	0.94	0.51	0.25	-0.04
Romania	0.78	0.51	0.06	0.97	0.79	0.33	0.04
San Diego	0.87	0.5	0.04	0.96	0.89	0.12	0.01
Sweden	0.55	0.5	0.01	0.99	0.55	0.67	0.01
Sydney	0.78	0.43	0	0.98	0.79	0	-0.03
Taiwan	0.85	0.46	0	0.99	0.86	0	-0.02
Würzburg	0.83	0.54	0.07	0.99	0.84	0.5	0.08

Appendix B

Supplementary Materials: Gene-Wise Feature Selection

Table B.1:

Classifier	Site	Accuracy	AUC	Sensitivity	Specificity	PPV	NPV	Kappa
Feature Selection with LR								
LR	Cagliari	0.63	0.59	0.20	0.80	0.28	0.72	0.00
XGB	Cagliari	0.66	0.52	0.05	0.90	0.18	0.71	-0.06
LR	Geneva	0.77	0.63	0.62	0.82	0.50	0.88	0.40
XGB	Geneva	0.75	0.42	0.00	0.98	0.00	0.77	-0.03
LR	Japan	0.66	0.58	0.23	0.80	0.27	0.76	0.03
XGB	Japan	0.76	0.43	0.00	1.00	0.00	0.76	0.00
LR	Mayo	0.69	0.52	0.23	0.83	0.29	0.78	0.07
XGB	Mayo	0.77	0.55	0.00	1.00	0.00	0.77	0.00
LR	Paris	0.62	0.45	0.18	0.72	0.13	0.80	-0.08
XGB	Paris	0.82	0.47	0.03	0.99	0.50	0.82	0.03
LR	Romania	0.73	0.54	0.19	0.88	0.29	0.80	0.07
XGB	Romania	0.78	0.48	0.03	0.98	0.33	0.79	0.02
LR	San Diego	0.60	0.51	0.30	0.64	0.09	0.88	-0.03
XGB	San Diego	0.74	0.50	0.17	0.81	0.10	0.89	-0.01
LR	Sweden	0.53	0.53	0.07	0.92	0.39	0.54	-0.02
XGB	Sweden	0.55	0.52	0.01	1.00	1.00	0.55	0.01
LR	Sydney	0.78	0.55	0.08	0.96	0.33	0.80	0.05
XGB	Sydney	0.81	0.62	0.08	1.00	1.00	0.81	0.12
LR	Taiwan	0.80	0.61	0.15	0.91	0.22	0.87	0.07
XGB	Taiwan	0.86	0.39	0.00	1.00	0.00	0.86	0.00
Feature Selection with XGB								
LR	Cagliari	0.69	0.54	0.16	0.90	0.39	0.73	0.08
XGB	Cagliari	0.74	0.56	0.20	0.95	0.61	0.75	0.19
LR	Japan	0.77	0.54	0.06	0.99	0.67	0.77	0.08
XGB	Japan	0.75	0.48	0.00	0.99	0.00	0.76	-0.02
LR	Mayo	0.77	0.52	0.00	1.00	0.00	0.77	0.00
XGB	Mayo	0.74	0.57	0.00	0.97	0.00	0.76	-0.04
LR	Paris	0.81	0.53	0.00	0.99	0.00	0.82	-0.01
XGB	Paris	0.79	0.52	0.03	0.95	0.11	0.82	-0.03
LR	San Diego	0.58	0.46	0.30	0.61	0.09	0.88	-0.04
XGB	San Diego	0.68	0.52	0.30	0.73	0.12	0.90	0.02
LR	Sydney	0.79	0.49	0.00	1.00	0.00	0.79	0.00
XGB	Sydney	0.78	0.38	0.00	0.98	0.00	0.79	-0.03
LR	Taiwan	0.84	0.63	0.00	0.97	0.00	0.86	-0.04
XGB	Taiwan	0.86	0.62	0.00	1.00	0.00	0.86	0.00
LR	wuerzburg	0.81	0.51	0.00	0.97	0.00	0.82	-0.04
XGB	wuerzburg	0.79	0.41	0.03	0.95	0.13	0.83	-0.02

Table B.2: Results for the exemplary subject experiment using each combination of feature selection and inference model. *Abbreviations*: area under the curve (AUC) positive predictive value (PPV) negative predictive value (NPV).

Classifier	Group	Accuracy	AUC	Sensitivity	Specificity	PPV	NPV	Kappa
Feature Selection with LR								
LR	Bottom 50%	0.53 (0.47, 0.63)	0.51 (0.48, 0.64)	0.48 (0.12, 0.77)	0.66 (0.40, 0.86)	0.55 (0.50, 0.65)	0.50 (0.46, 0.65)	0.05 (-0.02, 0.27)
LR	Top 50%	0.69 (0.62, 0.83)	0.80 (0.68, 0.91)	0.75 (0.57, 0.86)	0.70 (0.55, 0.88)	0.63 (0.56, 0.86)	0.79 (0.69, 0.88)	0.37 (0.26, 0.66)
XGB	Bottom 50%	0.50 (0.37, 0.63)	0.49 (0.41, 0.68)	0.33 (0.19, 0.75)	0.54 (0.43, 0.80)	0.53 (0.38, 0.65)	0.48 (0.37, 0.65)	0.02 (-0.24, 0.27)
XGB	Top 50%	0.71 (0.62, 0.82)	0.83 (0.70, 0.92)	0.71 (0.71, 0.86)	0.72 (0.55, 0.88)	0.68 (0.56, 0.84)	0.76 (0.70, 0.88)	0.41 (0.26, 0.64)
Feature Selection with XGB								
LR	Bottom 50%	0.62 (0.46, 0.67)	0.61 (0.42, 0.70)	0.71 (0.25, 0.89)	0.50 (0.29, 0.72)	0.62 (0.48, 0.67)	0.61 (0.45, 0.77)	0.23 (-0.06, 0.33)
LR	Top 50%	0.69 (0.56, 0.82)	0.82 (0.72, 0.90)	0.71 (0.42, 0.87)	0.71 (0.56, 0.89)	0.68 (0.50, 0.84)	0.73 (0.60, 0.89)	0.37 (0.09, 0.62)
XGB	Bottom 50%	0.62 (0.46, 0.68)	0.61 (0.43, 0.75)	0.71 (0.36, 0.90)	0.54 (0.29, 0.72)	0.62 (0.48, 0.72)	0.59 (0.44, 0.87)	0.23 (-0.07, 0.37)
XGB	Top 50%	0.81 (0.64, 0.87)	0.83 (0.78, 0.91)	0.75 (0.69, 1.00)	0.71 (0.62, 0.90)	0.72 (0.59, 0.87)	0.81 (0.70, 1.00)	0.62 (0.28, 0.73)

Bibliography

[1] Basma Abdulaimma, Paul Fergus, and Carl Chalmers. "Extracting Epistatic Interactions in Type 2 Diabetes Genome-Wide Data Using Stacked Autoencoder". In: *arXiv preprint arXiv:1808.09517* (2018).

[2] Seth Carbon et al. "AmiGO: online access to ontology and annotation data". In: *Bioinformatics* 25.2 (2008), pp. 288–289.

[3] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM. 2016, pp. 785–794.

[4] Peter JA Cock et al. "Biopython: freely available Python tools for computational molecular biology and bioinformatics". In: *Bioinformatics* 25.11 (2009), pp. 1422–1423.

[5] Sam F Colin et al. "Chronic lithium regulates the expression of adenylate cyclase and Gi-protein alpha subunit in rat cerebral cortex". In: *Proceedings of the National Academy of Sciences* 88.23 (1991), pp. 10634–10637.

[6] 1000 Genomes Project Consortium et al. "A global reference for human genetic variation". In: *Nature* 526.7571 (2015), p. 68.

[7] A Corvin, N Craddock, and P F Sullivan. "Genome-wide association studies : a primer". In: 906000 (2010), pp. 1063–1077. DOI: 10.1017/S0033291709991723.

[8] Noëmie Drancourt et al. "Duration of untreated bipolar disorder: missed opportunities on the long road to optimal treatment". In: *Acta Psychiatrica Scandinavica* 127.2 (2013), pp. 136–144.

[9] Paul Fergus et al. "Utilising deep learning and genome wide association studies for epistatic-driven preterm birth classification in African-American women". In: *IEEE/ACM transactions on computational biology and bioinformatics* (2018).

[10] Manuel AR Ferreira et al. "Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder". In: *Nature genetics* 40.9 (2008), p. 1056.

[11] Paul Grof et al. "Is response to prophylactic lithium a familial trait?" In: *The Journal of clinical psychiatry* 63.10 (2002), pp. 942–947.

[12] Arthur E Hoerl and Robert W Kennard. "Ridge regression: Biased estimation for nonorthogonal problems". In: *Technometrics* 12.1 (1970), pp. 55–67.

[13] Liping Hou et al. "Genetic variants associated with response to lithium treatment in bipolar disorder: A genome-wide association study". In: *The Lancet* 387.10023 (2016), pp. 1085–1093. ISSN: 1474547X. DOI: 10.1016/S0140-6736(16)00143-4.

[14] Omid Kohannim et al. "Predicting temporal lobe volume on MRI from genotypes using L 1-L 2 regularized regression". In: *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2012, pp. 1160–1163.

[15] Malgorzata Maciukiewicz et al. "GWAS-based machine learning approach to predict duloxetine response in major depressive disorder". In: *Journal of psychiatric research* 99 (2018), pp. 62–68.

[16] Mirko Manchia et al. "Assessment of response to lithium maintenance treatment in bipolar disorder: a Consortium on Lithium Genetics (ConLiGen) report". In: *PloS one* 8.6 (2013), e65636.

[17] Mirko Manchia et al. "Genetic risk of suicidal behavior in bipolar spectrum disorder: analysis of 737 pedigrees". In: *Bipolar disorders* 15.5 (2013), pp. 496–506.

[18] Andries T Marees et al. "A tutorial on conducting genome-wide association studies: Quality control and statistical analysis". In: *International journal of methods in psychiatric research* 27.2 (2018), e1608.

[19] Huaiyu Mi et al. "Protocol update for large-scale genome and gene function analysis with the PANTHER classification system (v. 14.0)". In: *Nature protocols* 14.3 (2019), p. 703.

[20] Masahito Miki et al. "Effects of subchronic lithium chloride treatment on G-protein subunits (Golf, Gγ7) and adenylyl cyclase expressed specifically in the rat striatum". In: *European journal of pharmacology* 428.3 (2001), pp. 303–309.

[21] Casimiro A Curbelo Montaez et al. "Deep Learning Classification of Polygenic Obesity using Genome Wide Association Study SNPs". In: *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2018, pp. 1–8.

[22] Thomas W Mühleisen et al. "Genome-wide association study reveals two new risk loci for bipolar disorder". In: *Nature communications* 5 (2014), p. 3339.

[23] Abraham Nunes. *abrahamnunes/newton-raphson: Initial release*. Apr. 2018. DOI: 10.5281/zenodo.1211725. URL: https://doi.org/10.5281/zenodo.1211725.

[24] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[25] Alkes L. Price et al. "Principal components analysis corrects for stratification in genome-wide association studies". In: *Nature Genetics* 38.8 (2006), pp. 904–909. ISSN: 10614036. DOI: 10.1038/ng1847.

[26] Shaun Purcell et al. "PLINK: a tool set for whole-genome association and population-based linkage analyses". In: *The American Journal of Human Genetics* 81.3 (2007), pp. 559–575.

[27] Stephan Ripke et al. "Genome-wide association study identifies five new schizophrenia loci". In: *Nature genetics* 43.10 (2011), p. 969.

[28] Adriana Romero et al. *DIET NETWORKS: THIN PARAMETERS FOR FAT GENOMICS*. Tech. rep. URL: http://www.internationalgenome.org/.

[29] Janusz K Rybakowski, Maria Chlopocka-Wozniak, and Aleksandra Suwalska. "The prophylactic effect of long-term lithium administration in bipolar patients entering treatment in the 1970s and 1980s". In: *Bipolar disorders* 3.2 (2001), pp. 63–67.

[30] Thomas G Schulze et al. "The International Consortium on Lithium Genetics (ConLiGen): an initiative by the NIMH and IGSLI to study the genetic basis of response to lithium treatment". In: *Neuropsychobiology* 62.1 (2010), pp. 72–78.

[31] Eli A Stahl et al. "Genome-wide association study identifies 30 loci associated with bipolar disorder". In: *Nature genetics* 51.5 (2019), p. 793.

[32] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[33] Peter M Visscher et al. "10 years of GWAS discovery: biology, function, and translation". In: *The American Journal of Human Genetics* 101.1 (2017), pp. 5–22.

[34] Bojian Yin et al. "Using the structure of genome data in the design of deep neural networks for predicting amyotrophic lateral sclerosis from genotype". In: *bioRxiv* (2019), p. 533679.

[35] L Trevor Young et al. "Mononuclear leukocyte levels of G proteins in depressed patients with bipolar disorder or major depressive disorder." In: *The American journal of psychiatry* (1994).

[36] Daniel R Zerbino et al. "Ensembl 2018". In: *Nucleic acids research* 46.D1 (2017), pp. D754–D761.

[37] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2 (2005), pp. 301–320.