The Application of Double Machine Learning Onto Genomics Data Associated with Amyotrophic Lateral Sclerosis

by

Crystal Wang

S.B. Electrical Engineering and Computer Science, Massachusetts Institute of Technology (2020)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

Author .....
Department of Electrical Engineering and Computer Science
May 20, 2021

Certified by .....
Ernest Fraenkel
Professor of Biological Engineering
Thesis Supervisor

Accepted by .....
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

The Application of Double Machine Learning Onto Genomics Data Associated with Amyotrophic Lateral Sclerosis

by

Crystal Wang

Submitted to the Department of Electrical Engineering and Computer Science
on May 20, 2021, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Finding causal relationships between a dataset and an observed outcome is especially important when there is potential for meaningful interventions. One such area of focus is a biological setting, where there are many opportunities for diagnosis, prevention, and treatment research. Amyotrophic Lateral Sclerosis (ALS) is a progressive neurodegenerative disease for which which there is no cure and relatively little is known about what causes the disease. Previous work has shown certain genes to be associated with ALS and previous work have used machine learning to try and determine the causal features of ALS. In this thesis we experiment with Double Machine Learning [8] to find causal features of ALS. We apply this method on both synthetic and real datasets that are associated with ALS and explain the advantages and shortcomings of this methodology on genetics data where correlation is present.

Thesis Supervisor: Ernest Fraenkel
Title: Professor of Biological Engineering

Acknowledgments

This thesis could not have been made possible without the support and help of many people.

To Ernest Fraenkel, my thesis supervisor, thank you for your continued guidance throughout this project. I am so appreciative for your help answering my many questions, your patience through long conversations over Zoom and email threads, and especially your insight, which was invaluable. I am truly grateful to have had this opportunity to work with you.

To Rahul Gopalkrishnan, David Alvarez-Melis, Vasilis Syrgkanis, and Miruna Osprecu, thank you for working so closely with me throughout this process. The feedback you have given me have truly helped me learn what it is like to conduct research in a setting where not much is known. Your advice showed me how to think through problems more systematically and how to break down what seems like a large, insurmountable problem into more manageable pieces. I could not have asked for better mentors.

I would like to thank the members of the Fraenkel Lab for welcoming me in with open (but virtual) arms. I would especially like to thank Divya Ramamoorthy, Yogindra Raghav, and Jonathan Li for sharing your knowledge and experience with me.

I have so much gratitude to the friends that I have made both at and outside of MIT. You have become like a second family to me, and spending time with you all always makes me happy.

Finally, thank you to my parents and brother, Jing Li, Wensheng Wang, and Matthew Wang. Your unconditional love and support mean the world to me, and I am so thankful to have you as my family.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

As we are headed towards a more data driven world, the question of how data relates to observed outcomes becomes more prevalent. Given some observed data and outcomes, to gain any sort of interpretability as to why an outcome is happening, we must learn how the data and outcomes relate to one another. In particular, for areas of research where there is possibility of intervention, this problem becomes especially important. We notice, however, that some relationships might be more useful than others, especially when designing interventions. The two types of relationships commonly explored are correlative and causal relationships, where a correlative relationship just indicates that there exists a relationship, but a causal relationship between two entities indicates that one entity directly impacts the other. While correlations are generally easier to detect, causal relations can tell us more about why something is happening rather than just noticing a change.

Finding causal relationships between datasets and outcomes is not a new problem, but it is one that has expanded to many fields. In finance, we might want to know why a certain stock price increases, based off of trading history, in technology, we might want to know why a person clicks on one ad versus another based off their prior behavior, and in biology, we might want to know why a person contracts a certain disease based off of their health history. If these relationships were just correlative instead of causal, it could be useful in helping us notice trends, but would not explain the phenomenon. Not only does finding a causal relationship between the data and

outcomes give us better understanding of why something is happening, it also allows us to then predict what might happen on future, unseen data.

In a biological setting, where there are many opportunities for meaningful interventions, finding robust methods to generalize to unseen data becomes especially important. For the invariably fatal disease Amyotrophic Lateral Sclerosis (ALS), where causes are relatively unknown, finding a causal relationship between features in a person's genome and the disease could open up doors for diagnosis and treatment for the disease.

1.1 Amyotrophic Lateral Sclerosis (ALS)

1.1.1 Background on ALS

ALS is a progressive neurodegenerative disease that affects motor neurons, which are nerve cells that control muscle movement. Motor neurons in the brain and the spinal cord connect to the muscles, so when motor neurons are damaged, the brain cannot send signals to the muscles anymore. Eventually muscles will atrophy, and individuals can lose their ability to speak, eat, move, and even breathe [15].

Around 5-10% of cases of ALS are familial, or inherited from one parent [17]. There are known mutations in genes that cause familial ALS; for example, a mutation in a gene known as C9ORF72 causes around 25-40% of familial cases [17]. On the other hand, a vast majority of cases are sporadic, where it is unclear why a person gets the disease. There is no clear cause of sporadic ALS, but it is theorized that certain factors like age and sex and certain environmental factors like smoking can have an effect on contracting the disease [15]. For sporadic ALS, advancements have found genes such as SOD1, TARDBP, FUS, OPTN, VCP, UBQLN2, C9ORF72 and PFN1 [21] to be causal, but they only explain about 10% of ALS cases [21, 16]. However, research using twin studies suggests that the heritability of sporadic ALS is estimated to be around 61% [2], which motivates the search over genes to find causality.

Currently, without a clear cause of ALS, it is also very hard to diagnose ALS early

on, as for most people the first symptoms include muscle twitches, tight muscles, and muscle weakness, which are subtle. Moreover, there is not a test that is able to diagnose a person with ALS, but rather the diagnosis is based off of a history of symptoms and a series of tests to rule out other diseases [17].

In the United States, around 5000 people get diagnosed with ALS every year [10] and typically patients die within 3-5 years from when symptoms first appear [17]. It is clear that ALS is a serious disease that affects many, so finding the root cause of the disease is a large area of research.

1.1.2 Impact of Causal Feature Discovery on ALS

As we see, there is a clear need for more investigation in many facets regarding ALS. There is a need to learn what factors cause ALS, which can help with the diagnosis of the disease, a need to learn how to prevent ALS if possible, and a need to learn how to treat ALS. These problems have overlap with one another, but it seems like finding the causal features of ALS is a fundamental problem.

To determine the cause of ALS, it is important that we determine *causal* features of ALS and not just features that are associated with ALS. The essence of a causal feature is that the feature is used in the generative process of the data, so changing the feature will have a direct impact on the observed data. Since it is estimated that heritability is high for sporadic ALS, we want to determine what parts of the human genotype have a causal relationship with the phenotype of contracting ALS, and by doing so we will gain a better understanding of who gets ALS, why they do, and how we might prevent it.

Finding genetic markers for ALS is a large area of research, but since there are so many genetic markers in a person's DNA, it would be infeasible to check all of the markers one by one to determine a causal effect. Using statistical methods, if we find a small subset of features that are causal, then we can significantly reduce the scope of biological tests that would have to be run to confirm that these features are in fact, causal. Furthermore, if these statistical techniques work, these methodologies will be general enough to apply to other diseases such as Alzheimer's and Parkinson's, where

relatively little is known about the causes of these diseases.

1.2 Related Work

As mentioned, conducting research on ALS causes and causal feature selection are not new problems. Biologists have done genomic sequencing [32, 25, 5] to find mutations in genetics that are associated with ALS. In data science, there have been many techniques developed to select causal features to represent a dataset [8, 20]. The main contribution in our work is to apply more data science driven approaches to finding causal features of ALS.

1.2.1 Genome Wide Association Studies (GWAS)

In genetics research, a genome-wide association study is used to associate certain gene mutations to a particular disease. The way GWAS are conducted is a large group of people who have the disease and who do not have the disease are sampled. Then, each individual's genome is scanned at certain single nucleotide polymorphisms (SNPs), which are places on the genome where a sufficiently large fraction of the population have a mutation. These SNPs can also be seen as markers on the gene, and a person has about 4-5 million SNPs in their genome [7]. The places where the SNPs are chosen for GWAS are based on experimental constraints, and usually around 500,000-1,000,000 [6] are looked at in a GWAS.

Once every person's genome is scanned in the study, SNPs where there are consistent differences between people who have the disease and people who don't are said to be the SNPs that are associated with the disease. There have been many GWAS conducted for ALS and these studies have found certain genes to be associated with ALS, like KIF5A [19] and *C21orf2* [29]. Buniello et al. [5] compiled a list of 317 SNPs from 39 GWAS found to be associated with ALS.

Hu et al. [12] emphasize the point that causation does not imply correlation and correlation does not imply causation, so there is a need for causal inference methods to be applied onto genomic data. In general with GWAS, it is very hard to distinguish

between causal SNPs and SNPs that are simply associated with the disease because they have a high correlation with the causal SNP. Thus, while there has been significant strides made in finding what SNPs are associated with ALS, there is still a need to determine causal SNPs.

1.2.2 Causal Feature Selection

In machine learning algorithms, a common assumption to make is that the training data is drawn from the same distribution as the testing data; however this assumption is not always valid. For example, a classifier might be trained on ALS patient data from one specific population but if we want a general classifier, the testing data might have patients from different populations. If we want a classifier to be robust against covariate shift, or when the training and testing distributions are not the same, using features that have a close causal relationship with the target variable is more likely to create a reliable classifier. Furthermore, if we uncover a model that is trained on causal features, the model will be more predictive of the effects of interventions. In genetics data, since there can be many causal features and many confounders (variables that affect both the causes and the effects) within the data, we look at causal feature selection algorithms that address the multiple-cause setting and the confounding setting.

Deconfounder

To address the multiple-cause setting, in [31], Wang and Blei propose a framework in which multiple causal inference can actually have fewer constraints than classical causal inference, where there is just a single possible cause. In classical causal inference, all confounders of the data must be identified, measured, and controlled for, so there is a strong assumption that all confounders must be able to be observed. Wang and Blei propose the deconfounder, which is a method to search for causal relationships that uses the weaker assumption that there must not be unobserved single-cause confounders, a variable that affects just one cause and the outcome.

The methodology is first to fit a model that approximates the joint distribution of the potential causal features, which is used to infer the latent variables of those causal features, and these latent variables will be used as substitutes for the confounders. Then, for each individual in the dataset, infer the latent variables based off of the model. Finally, to perform causal inference, the substitutes for the confounders found using the model are augmented to the dataset to use as input data.

The reason this methodology works is if the model in the first step accurately captures the distribution of potential causes, it would mean that the causes are conditionally independent given the latent variables. If somehow these latent variables did not capture all of the multi-cause confounders, then there would be dependence among the causes induced by the uncaptured multi-cause confounder, which is a contradiction. Thus, these latent variables must capture all multiple-cause confounders, and it is why the latent variables can be used as substitutes for the confounders. However, this methodology would break down if there were single-cause confounders, since conditional independence could still be observed if there was an uncaptured single-cause confounder. Figure 1-1 shows a graphical representation of this argument.

Since the only assumption that is made in the deconfounder setting is that there are no single-cause confounders, it seems like this methodology would be able to be used on SNPs data, since if a person has a large number of SNPs, it's unlikely that a confounder has an effect on only one. Wang and Blei use the confounder on genetics data, where they try to find causal SNPs that can predict a person's height. When simulating their data, they artificially group individuals together to introduce confounding effects on the SNPs data, and so the outcome is also dependent on this population bias. Wang and Blei apply their deconfounder method and find that it often outperforms regression, when the unobserved confounder is included. The method heavily relies on the model created in the first step, so checking how well the model used to infer latent variables is crucial.

In fact, one point of failure is when the deconfounder is applied on data that comes from spatially correlated individuals, meaning that individuals' features are close

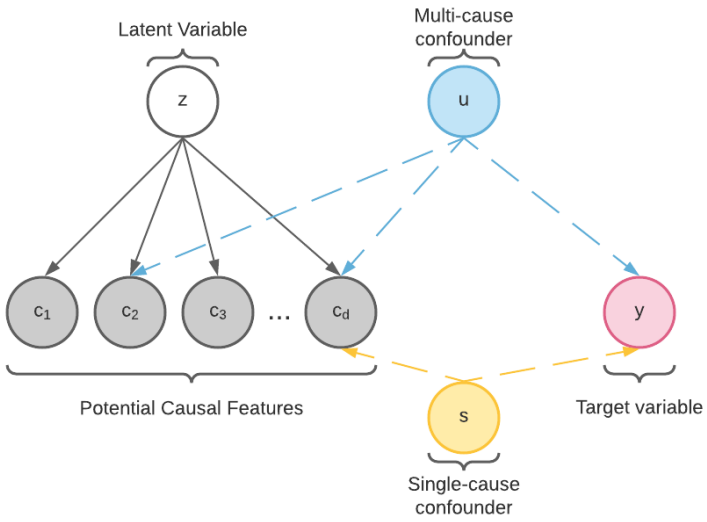

Figure 1-1: The deconfounder argument adapted from Wang and Blei [31]. If a latent variable z is found such that $c_1, c_2, \ldots, c_d$ are conditionally independent, it is impossible for a multi-cause confounder, u, to exist, since that would introduce dependence. However, a single-cause confounder s is still able to exist.

together spatially. One example of spatial correlation would be to sample individuals' features from a unit cube, where no spatial correlation would mean the features are sampled uniformly from within the cube, and a high level of spatial correlation could mean sampling features from close to the corners of the cube. Again, here, the first step is predictive of downstream causal inference failing, since the methods used to model the latent variables do not do so well with the spatial correlation.

In real genomics data, there is the concept of linkage disequilibrium, which refers to correlations between different SNPs in the dataset. Usually, SNPs that are closer together on the chromosome are more correlated with one another, which is an example of spatial correlation. Thus, it is not clear how well the deconfounder method will do on real genomics data, which is plausibly spatially correlated.

Machine learning methods on ALS data

Naturally, the idea of finding causal features for certain diseases has been explored before [31, 30]. As mentioned, Wang and Blei apply their algorithm on GWAS data.

Vasilopoulou et al. [30] provide a review on GWAS that use machine learning approaches that aim to understand the causes of ALS through genomics features. Currently, some challenges there are with applying machine learning methods onto genomics data including making sure the data is of a certain quality, overcoming the sheer amount of features in datasets, and being able to explain and interpret the results.

Feature selection methods are commonly used to reduce the large number of initial features to be used in machine learning methods. Some methods are purely statistical ones and filter features based off significant interactions between features and the target. Other strategies use biological knowledge and only include features that are in regulatory pathways or features that have been found to be associated with ALS through GWAS. One of the most common ways biological knowledge is used is to filter genes based off of ALS-specific knowledge early on. A benefit to this approach is the results are more likely to have biological intepretability, but it also risks introducing bias to the machine learning models to be used.

Machine learning models that have been used for classifications that seem to do well include Gradient Boosting, CNNs, Deep Neural Networks, Logistic Regressions, SVMs, and Random Forest. In particular, deep learning models seem promising in both finding genes that are associated with ALS and patient classification. Unfortunately, these studies have shown little reproducibility, which could be attributed to the different datasets, features, and feature selection algorithms used between these studies but could also be attributed to the more general challenge of finding causal genes amongst a wide pool of possibilities.

This thesis aims to apply different causal feature selection algorithms on genomic data, both real and synthetic. There are many nuances of genomic data such as sparsity, correlation, and high dimensionality that would be interesting to see if these algorithms' assumptions hold and can provide good results. We hope that not only will we find a causal relationship between features and outcomes, but the relationship is interpretable and robust.

Chapter 2

Background

As mentioned, there are many algorithms to perform causal feature selection. We will focus on a technique developed by Chernozhukov et al. [8] and explain other necessary models and definitions that we will use in this thesis.

2.1 Linear Models

2.1.1 Linear Regression with L_1 Regularization (Lasso Regression)

Given a dataset $X \in \mathbb{R}^{n \times d}$ and an outcome $Y \in \mathbb{R}^{n \times 1}$, we wish to find $w \in \mathbb{R}^{d \times 1}$ and $w_0 \in \mathbb{R}$ that minimizes the objective function

$$L_{lasso}(X, Y, w, w_0) := \frac{1}{n} \|Xw + w_0 - Y\|_2^2 + \lambda \sum_{i=1}^{d} |w_i| \tag{2.1}$$

Here, $\lambda \geq 0$ is a hyperparameter that controls the strength of regularization. A larger λ will enforce more sparsity in the found w. Shown in Equation 2.1 is L_1 regularization, but we could also add L_2 regularization which would be $\|w\|^2 = \sum_{i=1}^{d} w_i^2$.

2.1.2 Logistic Regression

Given a dataset $X \in \mathbb{R}^{n \times d}$ and an outcome $Y \in \{0,1\}^{n \times 1}$, we wish to find $w \in \mathbb{R}^{d \times 1}$ and $w_0 \in \mathbb{R}$ that minimizes the objective function

$$L_{CE}(X, Y, w, w_0) := -\frac{1}{n} \sum_{i=1}^{n} \Big(y_i \log\big(\sigma(x_i w + w_0)\big) + (1 - y_i) \log\big(1 - \sigma(x_i w + w_0)\big) \Big) \tag{2.2}$$

Here, $y_i \in \mathbb{R}$ is the i-th index of Y, $x_i \in \mathbb{R}^{1 \times d}$ is the i-th row of X, and σ denotes the sigmoid function, or $\sigma(z) = \dfrac{1}{1 + e^{-z}}$. While not stated in the objective function in equation 2.2, we can add regularization here as well, usually L_1 or L_2.

2.2 Algorithms for Causal Feature Selection

2.2.1 Double/Debiased Machine Learning (DML)

Chernozhukov et al. [8] describe an algorithm that finds an de-biased estimator of a target parameter as follows. Given a set of features in a dataset, we would like to determine the relationship between a policy variable (which is typically causal), denoted as D, in the dataset and the target variable, denoted as Y. This problem is a bit more complicated, since the other features in the dataset, denoted as X, could have a not necessarily linear relationship with the policy variable D and the target variable Y as well. To model these relationships, we have

$$Y = D\theta_0 + g_0(X) + U, \ \mathbb{E}[U|X, D] = 0 \tag{2.3}$$

$$D = m_0(X) + V, \ \mathbb{E}[V|X] = 0 \tag{2.4}$$

where U and V are noise variables. Here, θ_0, m_0, and g_0 are unknown. The main goal is to estimate θ_0, but we see that Y might depend on X in a not necessarily linear way through g_0 and D might depend on X through m_0, also not necessarily linearly.

On a high level, given observations of D, X, and Y, we want to first model the relationship between D and X and we calculate V as the residual. Then, we model

the relationship between Y and X and also calculate the residual $W = Y - g_0(X)$. To estimate θ_0, we regress W onto V, which we see will minimize regularization bias, which comes from using any machine learning approach for modeling. In fact, we will see that fitting models for both g_0 and m_0 is important in minimizing the regularization bias. A naive form of DML would be to only fit for g_0, ignoring the relationship that D and X may have, and regress $Y - g_0(X)$ onto D. But, we will see that this way causes the error between our estimated $\widehat{\theta_0}$ and the true θ_0 to converge very slowly.

To put more DML more formally, our dataset, $\mathcal{A}$, consists of n samples of (X_i, D_i, Y_i), for $i = 1, 2, \ldots, n$. To estimate θ_0, we first split the set of indices in half to get I and I^C such that $I \cup I^C = \{1, 2, \ldots, n\}$. Then separate the dataset $\mathcal{A}$ into two datasets $\mathcal{A}_I := \{X_I, D_I, Y_I\}$ and $\mathcal{A}_{I^C} := \{X_{I^C}, D_{I^C}, Y_{I^C}\}$ using the respective indices. With $\mathcal{A}_I$, use machine learning methods on D_I and X_I to find an estimator $\widehat{m_{0I}}$ fitted on $D_I = \widehat{m_{0I}}(X_I)$. Then, we obtain an estimate for $\widehat{V} = D - \widehat{m_{0I}}(X)$, from equation 2.4. $\widehat{V}$ here represents our policy variable D with the effect of X partialled out. Next, we obtain an estimate $\widehat{g_{0I}}$ using Y_I and X_I by using machine learning models to fit $Y_I = \widehat{g_{0I}}(X_I)$. To get our final estimator, we use the other half of the dataset $\mathcal{A}_{I^C}$ to regress $Y_{I^C} - \widehat{g_{0I}}(X_{I^C})$ onto $\widehat{V}$, but using the model $\widehat{g_{0I}}$ trained on $\mathcal{A}_I$. We see that then, our problem becomes solving least squares over variables Y and D with the effect from X removed.

We obtain the following equation as an estimate of $\widehat{\theta_{0I}}$,

$$\widehat{\theta_{0I}} = \left(\frac{1}{n} \sum_{i \in I^C} \widehat{V}_i D_i \right)^{-1} \frac{1}{n} \sum_{i \in I^C} \widehat{V}_i (Y_i - \widehat{g_{0I}}(X_i)) \tag{2.5}$$

Here, instead of using ordinary least squares to solve for θ_0, θ_0 is obtained using instrumental variable (IV) specifications, where V is our instrumental variable. One form of the instrumental variable linear estimator is $(\hat{V}^T D)^{-1} \hat{V}^T (Y - \widehat{g_{0I}}(X_i))$ [24], which is what we see in equation 2.5.

We repeat the entire process but switch I and I^C to obtain an estimate $\widehat{\theta_{0I^C}}$. Our final estimate of θ_0 is $\widehat{\theta_0} = \frac{1}{2}(\widehat{\theta_{0I}} + \widehat{\theta_{0I^C}})$. The algorithm is detailed in Algorithm 1.

Algorithm 1: Double Machine Learning

Result: An estimator, $\hat{\theta}$, that models the relationship between the policy variable and the outcome.

Input: $X \in \mathbb{R}^{n \times d}$,
$D \in \mathbb{R}^{n \times 1}$,
$Y \in \mathbb{R}^{n \times 1}$

$\hat{\theta} \leftarrow 0$

$P_1, P_2 \leftarrow$ two equally sized partitions of the indices $[1, 2, \ldots, n]$

for $(I, I_c) \in [(P_1, P_2), (P_2, P_1)]$ **do**

- $X_I, D_I, Y_I \leftarrow X[I, :], D[I], Y[I]$
- $X_{I_C}, D_{I_C}, Y_{I_C} \leftarrow X[I_C, :], D[I_C], Y[I_C]$
- $g \leftarrow$ a machine learning model trained on X_I, Y_I
- $m \leftarrow$ a machine learning model trained on X_I, D_I
- $V_{I_C} \leftarrow D_{I_C} - m(X_{I_C})$
- $\theta_I \leftarrow \left(V_{I_C}^T D_{I_C}\right)^{-1} V_{I_C} \left(Y_{I_C} - g(X_{I_C})\right)$
- $\hat{\theta} \mathrel{+}= \theta_I$

end

return $\hat{\theta}/2$

The main reason to first find a model that fits X onto D and then X onto Y is to minimize regularization bias. To gain any sort of generalizability in the models m_0 and g_0, some regularization must be used while fitting these models, but by doing so, bias is also introduced. The naive version of DML only finds the model g_0 and then regresses D onto $Y - g_0(X)$ to find $\widehat{\theta_0}$, but the error $|\widehat{\theta_0} - \theta_0|$ converges slowly due to the regularization bias introduced. Namely, the error $|\widehat{\theta_0} - \theta_0|$ is a function of the error between the models, $|\widehat{g_0}(X) - g_0(X)|$, and this usually does not have mean 0 and diverges quickly.

When we perform the two-step process of fitting models g_0 and m_0 (hence Double Machine Learning), we see that the error term $|\widehat{\theta_0} - \theta_0|$ depends instead on $|\widehat{g_0}(X) - g_0(X)| \cdot |\widehat{m_0}(X) - m_0(X)|$, which is the product of two estimation errors. Now, this error can vanish quickly for many types of data generation processes. Figure 2-1 shows the densities of predicted estimators θ for an ordinary least squares regression, naive DML, and DML over 500 simulations using a dummy dataset. We see that DML is centered around the true θ, and while naive DML is better than ordinary least squares, it is not centered around the true θ, indicating some sort of bias exists

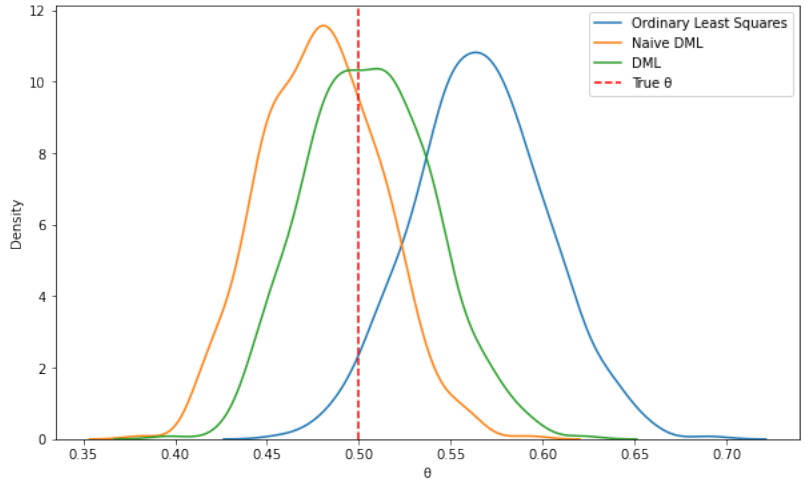

Figure 2-1: A comparison between using ordinary least squares, naive DML, and DML to find the estimator θ. With 500 simulations, plotted are the density distributions of θ for the different methods using a multivariate normal feature set X, a causal feature D that has a Cauchy relationship with X, and an outcome $Y = \theta D + \sin^2(X) + U$ for some normal variable U.

with that estimator.

2.2.2 Causal Feature Selection Using Orthogonal Search

Raj et al. [20] propose an approach for finding direct causal parents of an outcome variable, using only observed data. The setup is given an outcome Y and a set of variables X, where the variables in X can have non-linear relationships between them, we want to find the variables in X that have a direct effect on Y. Since we do not know what the causal features are, we test each feature for causality using DML.

On iteration i, the algorithm uses double machine learning with the i-th variable in X, X_i, as the policy variable and $X \setminus X_i$ as the other variables in the feature set to estimate θ_0, which is the relationship between X_i and Y. A significance test is then performed to see if X_i has an effect on Y, and if it does, then it is considered a direct causal parent of Y.

Chapter 3

Methodology

For our problem, the data that we will look at are an individual's SNPs and the outcome of whether that individual has ALS or not. Our datasets consist of matrices, where each row denotes an individual's SNPs, and each place on the SNP sequence can take on the value of 0, 1, or 2. A value of 0 means that the individual has the reference base-pair allele (the allele found on the reference genome), a value of 1 means that the individual has an alternative allele, and a value of 2 means that the individual has two alternative alleles at that SNP location. Our labels, or phenotype vector, can either be composed of real numbers, if the phenotype is quantitative like height, or the phenotype vector can be categorical if our outcomes are categorical like if a person has a disease or not.

Thus, we can represent our problem using $X \in \{0, 1, 2\}^{n \times d}$, $Y \in \mathbb{R}^{n \times 1}$ where n is the number of individuals in our dataset and d is the number of SNPs each individual has. The goal is to find a vector $w \in \mathbb{R}^{d \times 1}$ that minimizes $\mathcal{L}(Xw, Y)$, where $\mathcal{L}$ is mean squared error loss if Y is a quantitative phenotype and $\mathcal{L}$ is cross entropy loss if Y is categorical, constrained such that index i of w is non-zero iff. the SNP at location i is causal. Hence, we need to not only find a good estimator w, but also find which indices of w should be nonzero.

3.1 Datasets

3.1.1 EpiGEN

Since there are not that many publicly available real datasets that have phenotypes associated with ALS, we will need to generate synthetic data to first test our implementation before applying our implementation to a real dataset. Creating multiple datasets with different settings will allow us to see the pitfalls of our algorithms based off of the different settings. To generate multiple synthetic datasets, we will use use EpiGEN [3]. EpiGEN is a toolkit that allows us to simulate realistic genotype and phenotype matrices based off an inputted risk model. The user has to specify how many individuals the dataset should consist of, what the individuals' ancestry is, how many SNPs each individual should have, what the underlying disease or causal SNPs should be, how rare the disease SNPs should be, and how the disease SNPs interact with each other to produce the phenotype.

The way EpiGEN generates the dataset is as follows. From a precomputed corpora of genotypes that correspond to the inputted individuals' ancestry, EpiGEN subsamples SNPs from this corpora according to the users' preferences of the SNPs. Then, to compute the phenotypes, EpiGEN applies the risk model that the user defines to the SNPs to generate the phenotypes. In our datasets, the risk model we defined assumes a linearly causal relationship between the disease SNPs and the phenotype.

There are a few advantages of using EpiGEN over other synthetic data generation tools. The first is that it simulates epistasis, or interaction between SNPs that have a joint effect on a phenotype but individually do not have a large effect. Moreover, EpiGEN simulates realistic linkage disequilibrium (LD) patterns, and since one facet of interest is how well does DML work on data with correlated features, this is especially important. Finally, one feature of EpiGEN that will be useful in our pipeline is the ability to set the minor allele frequency (MAF) of the true causal SNPs. The minor allele frequency of a SNP is a probability indicating how often the second most common allele occurs in a population. Having control of this variable can give us insight on whether our pipeline will do better if the true disease SNPs are rarer, more

common, or somewhere in between.

Different settings of EpiGEN

The three main ways we generate datasets are by varying the MAF ranges of the disease SNPs, choosing disease SNPs that come from the same LD block, and choosing disease SNPs that come from different LD blocks.

For all of these datasets, we sample 10,000 individuals, 10,000 SNPs, and 10 causal or disease SNPs, where the relationship between the causal SNPs and the phenotype is linear. To simulate the covariate shift, we sample the training datasets and testing datasets from different populations. The training set comes from the HapMap 3 [13] population 'ASW', which consists of people of African ancestry in Southwest, USA and the testing set comes from the HapMap 3 population 'CEU', which consists of Utah residents with European ancestry.

Varying the MAF ranges of disease SNPs

We created ten datasets where we vary the MAF ranges of the disease SNPs by increments of 0.1, i.e. datasets where the disease SNPs' MAF ranges vary from 0-0.1, from 0.1-0.2, all the way to 0.9-1. We call these datasets `MAF_0_01`, `MAF_01_02`, ... `MAF_09_1` respectively. A dataset with a higher MAF range will consist of disease SNPs that are more prone to mutations. Here, the non-disease SNPs stay the same throughout each of these settings, so that we are able to compare between different MAF ranges.

We vary the MAF in our experiments because the MAF of a SNP indicates how rare or common the variant is, and SNPs with a wide range of MAF ranges have been show to be associated with ALS [9]. We want to observe if there are any differences in how well our pipeline works over different MAF ranges so that if there are any failures, we understand the limitations when applying this pipeline to the real dataset.

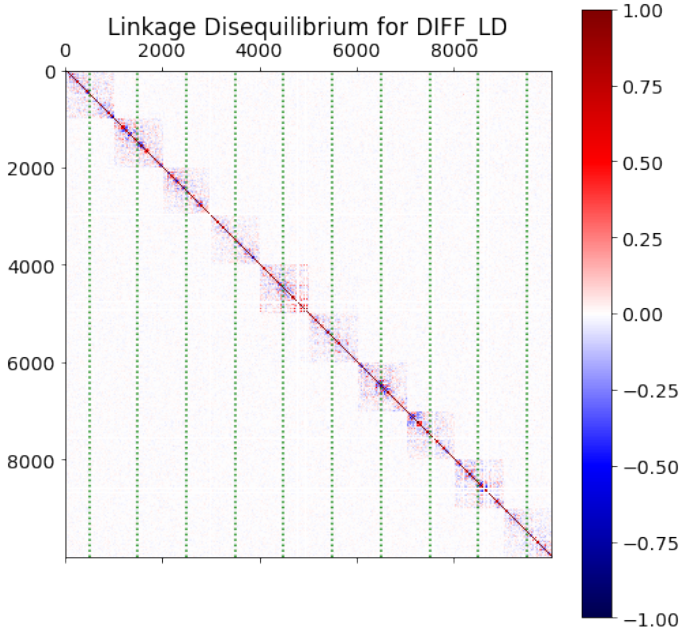

Figure 3-1: An example of the the LD block structure for the `DIFF_LD` dataset. The green lines denote where the disease SNPs lie in the blocks. As shown, there are 10 distinct LD blocks shown on the diagonal with one disease SNP sampled from each block.

Sampling disease SNPs that come from different LD blocks

Here, we sample 10 LD blocks and then from each LD block we sample a disease SNP for a total of 10 disease SNPs from 10 different LD blocks. We call this dataset `DIFF_LD`. An example of the LD patterns for `DIFF_LD` is shown in Figure 3-1. Here, we see 10 distinct blocks, indicating high correlation between features within each block, and the disease SNPs are located in the middle of each block.

In previous work, a point of failure for other algorithms was due to the linkage disequilibrium structure between SNPs [31]. This added correlation could also present a challenge to DML, because we know that linkage disequilibrium in real datasets can present confounding effects [1].

Sampling disease SNPs that come from same LD blocks

Here, we sample one LD block and choose 10 disease SNPs from that LD block. We call this dataset `SAME_LD`. An example of the LD pattern for `SAME_LD` is shown in

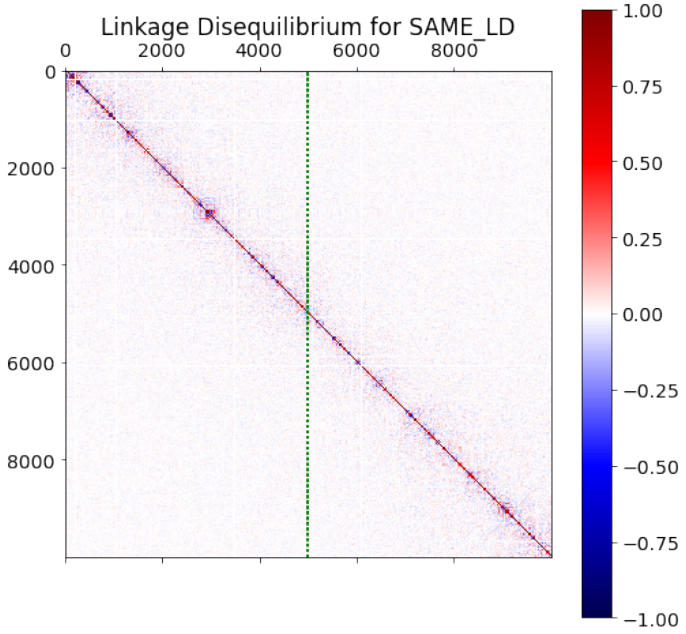

Figure 3-2: An example of the LD block structure for the `SAME_LD` dataset. The green lines denote where the disease SNPs lie in the block; here they are all in the middle of the block.

Figure 3-2. Here, we have only one block of correlation that spans the entire feature set, where the disease SNPs are all sampled from the middle of the block.

3.1.2 New York Genome Center ALS Dataset

The real dataset comes from the New York Genome Center (NYGC), a member of the Answer ALS Research Consortium. The dataset consists of 5,890 individuals, where 1,676 are cases, or they have ALS, and 4,214 are control, or they do not have ALS.

Preprocessing the real data

We had access to 5,890 patients' genomics data, which were all stored as Genome Variant Call Format (gVCF) files. Each gVCF file contains a list of SNPs each individual has, and each SNP had fields for its chromosome, position on the chromosome, ID if available, reference allele, and alternative allele. Additionally, we had access to metadata that told us what the control samples and case samples were. Here, we list the steps of preprocessing that was done to get the data into the SNP matrix, X, and

phenotype vector Y.

1. To construct the phenotype, or Y, vector, we looked at the metadata and mapped the samples that correspond to a control to 0 and mapped the samples that correspond to a case to 1. Here, the most important part is making sure that the index of each sample is consistent between the phenotype vector and the SNP matrix.

2. We performed joint genotyping [28] to merge the samples together. To keep with the Broad's GATK best practices [26], we used Sentieon's GVCFtyper which corresponds to GATK's GenotypeGVCFs [27]. Joint genotyping is advantageous because it validates the quality of data from the sequencing step by reducing the number of false positive variants and recovering variants in low-coverage sequencing areas by comparing those places to all other samples in the dataset [4].

3. Once all of the gVCF files were merged, to convert the merged files into SNPs matrices that only contain 0, 1, or 2, we used the tool vcftools (https://github.com/vcftools/vcftools, version v0.1.16) that takes in as input the merged gVCF file and outputs a SNP matrix where the individuals are the rows and features are the columns with associated metadata saying what position on the chromosome each feature is from. One quirk about vcftools is that the first column in the outputted matrix is the index corresponding to what individual the features are for, so we just drop that column.

3.2 Uncovering Correlated Features

Overall, since this problem tackles feature selection, we do two passes of feature selection, as seen in Figure 3-3. Using our SNP and phenotype matrices as inputs, we use lasso regression to do a first pass at feature selection, outputting only the features that are correlated to our phenotype. Then, we restrict our feature set to only the features that are outputted by lasso. It's very important that the true disease SNPs

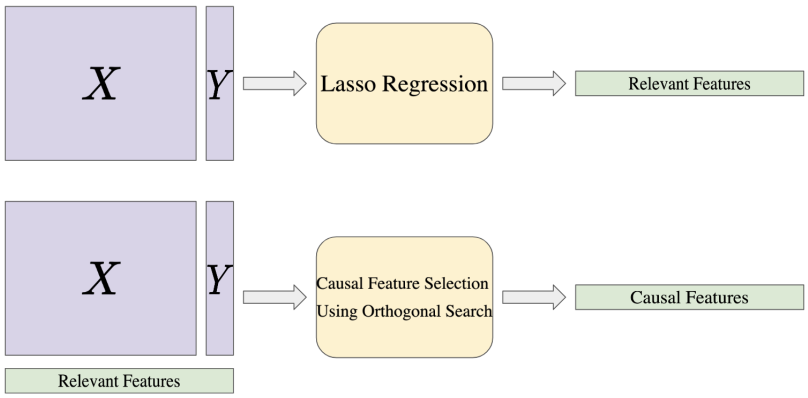

Figure 3-3: The two-step experimental pipeline, which involves two rounds of feature selection

do not get filtered out on this step, because we only test these features for causality in the second step.

3.2.1 Proposed Regularizations

Since regularization in this pass of feature selection plays a big role in what features end up selected, we want to choose a regularization method that will balance not choosing too many features to be inputted into the second step while making sure that the true causal SNPs are still selected. These are the four possibilities of regularization we will consider:

1. L_1 Regularization: We use cross validation to find the best L_1 regularization strength and output the coefficients that minimize cross validation loss

2. L_2 Regularization: We do the same thing as L_1 Regularization but with L_2 Regularization.

3. "Adaptive L_1" Regularization: We use cross validation to find the best L_1 regularization strength. Instead of using the best regularization strength though, we find the smallest regularization strength that still gives a cross validation error of within $\frac{1}{\sqrt{n}}$ the minimum cross validation loss. This means that we will likely get a larger set of coefficients than just using L_1. Figure 3-4 shows an example of the chosen regularization strengths. We see that for most cases the

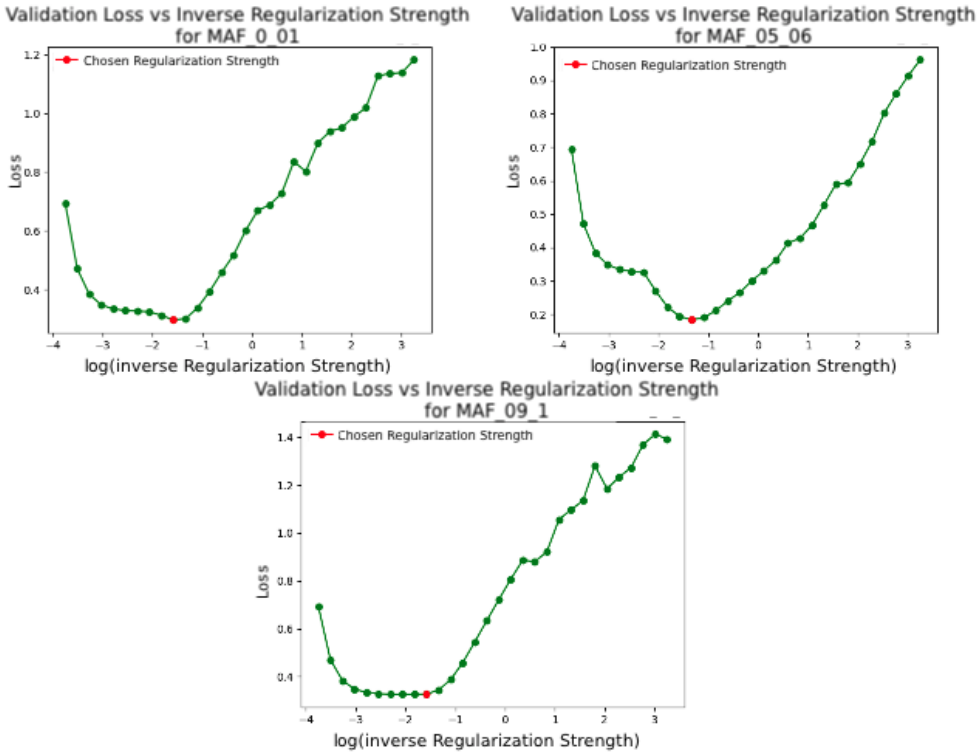

Figure 3-4: Examples of chosen regularization strength for the "Adaptive L_1" Regularization setting.

chosen regularization is exactly the one that minimizes out of sample loss, but for `MAF_09_1`, we see that it chooses a smaller regularization that still achieves an acceptable out of sample loss.

4. "Adaptive L_2" Regularization: We do the same thing as "adaptive L_1" except use L_2 regularization instead.

3.3 From Correlated Features to Causal Features

After we have the features from the lasso regression, we then perform causal selection using orthogonal search only on these features to determine if any of these features have a causal effect on the phenotype. We're modifying the algorithm slightly, because instead of searching over all d SNPs in the dataset X, we only search over a subset of features (specifically the ones outputted from the lasso classifier), so the number

of iterations reduces. This greatly helps with computation time, but with the added risk that lasso regression misses a true disease SNP in the first feature selection pass. Algorithm 2 details this entire causal feature selection algorithm.

Algorithm 2: The overall two-step feature selection algorithm to find causal features.

Result: A set of causal features to the outcome, Y
Input: $X \in \{0,1,2\}^{n \times d}$,
$Y \in \{0,1\}^{n \times 1}$,
$L_p \in \{L_1, L_2, \text{“Adaptive” } L_1, \text{“Adaptive” } L_2\}$
$S \leftarrow \{\}$
$w, w_0 \leftarrow \arg\min_{w,w_0} L_{CE}(X, Y, w, w_0) + L_p$, where L_{CE} comes from Equation 2.2
$\hat{w} \leftarrow \{i \mid w_i \neq 0\}$
for $i \in \hat{w}$ **do**
  $D \leftarrow X_i \in \{0,1,2\}^{n \times 1}$
  $W \leftarrow X \setminus X_i \in \{0,1,2\}^{n \times (d-1)}$
  $\theta \leftarrow$ DML(W, D, Y) where DML comes from Algorithm 1
  if θ *significant* **then**
    $S \leftarrow S \cup i$
  end
end
return S

Chapter 4

Results

In this chapter, we analyze the features selected by our pipeline in both the Logistic Regression step and the DML step. In the Logistic Regression step, we want to reduce the feature space significantly while keeping the true disease SNPs in our feature pool, so that the computation time of DML reduces but we still test the true disease SNPs for causality.

Then, to analyze the SNPs outputted by DML, we want to achieve high precision, which measures how many of the outputted SNPs are truly causal and we want to achieve high recall, which measures how many of the true disease SNPs were actually found. Of course, to measure precision and recall, we need to know the true disease SNPs which is possible for the EpiGEN datasets but not for the NYGC dataset. So, we also look at other standard evaluation metrics such as cross entropy error and the area under the ROC Curve (ROC-AUC) score to determine how well a model trained only on the significant features performs.

Furthermore, since the presence of correlation in the feature set is a potential issue for DML, we analyze the effect correlation has on DML.

4.1 EpiGEN

4.1.1 Analysis of Logistic Regression Using Different Regularizations

For the first round of feature selection, we fit a Logistic Regression to our training SNPs matrix and phenotype vector and experiment with different types of regularization. Here, we examine each regularization choice and explain whether it would be a good choice or not.

L_1 regularization

L_1 regularization does very well for most of the datasets–particularly the datasets that have disease SNPs in the non-extreme MAF ranges, as seen in Table 4.1. When we look at the error and ROC-AUC metrics, we see that our model learns and generalizes pretty well for each dataset, except for the `MAF_09_1` dataset. With the exception of `MAF_09_1`, out of sample ROC-AUCs are all pretty high (>0.8), suggesting good generalization. When we look at the number of features found, we see that L_1 does well in keeping the number of found SNPs small, while maintaining a good recall. For most datasets, we are able to retrieve all of the disease SNPs meaning that they will not be lost as a cause of this round of feature selection. Unfortunately, L_1 regularization is only able to select one SNP for the `MAF_09_1` dataset and it is not a causal SNP. Clearly, using L_1 regularization is a point of failure for this dataset. This could be due to the fact that since mutations in the disease SNPs are much more likely, there isn't that much meaningful signal so the Logistic Regression just pushes all of the weights to 0. With the exception of `MAF_09_1`, L_1 would be a good choice of regularization since it reduces the feature space while maintaining good recall.

L_2 regularization

We see that L_2 regularization is able to achieve perfect recall in all datasets, as shown in Table 4.2. Here, we see that the training accuracy and ROC-AUC are very high,

	MAF Datasets											
	(0, 0.1)	(0.1, 0.2)	(0.2, 0.3)	(0.3, 0.4)	(0.4, 0.5)	(0.5, 0.6)	(0.6, 0.7)	(0.7, 0.8)	(0.8, 0.9)	(0.9, 1)	DIFF_LD	SAME_LD
Train Error	0.27	0.18	0.15	0.14	0.13	0.12	0.13	0.13	0.12	0.32	0.18	0.094
Train ROC-AUC	0.83	0.95	0.97	0.97	0.98	0.98	0.98	0.99	0.99	0.58	0.94	0.99
Validation Error	0.30	0.23	0.21	0.20	0.17	0.18	0.19	0.19	0.19	0.32	0.21	0.13
Validation ROC-AUC	0.71	0.87	0.90	0.91	0.94	0.92	0.93	0.92	0.93	0.54	0.90	0.96
Test Error	0.35	0.24	0.21	0.21	0.27	0.29	0.49	0.21	0.22	0.33	0.40	0.33
Test ROC-AUC	0.81	0.88	0.90	0.91	0.91	0.92	0.92	0.92	0.91	0.51	0.86	0.89
Number of SNPs found	159	532	457	414	446	440	516	534	550	1	275	339
Disease SNPs found	6	10	10	10	10	10	10	10	10	0	8	8
Precision	0.038	0.019	0.022	0.024	0.022	0.022	0.019	0.012	0.018	0	0.029	0.024
Recall	0.6	1	1	1	1	1	1	1	1	0	1	0.8

Table 4.1: Evaluation metrics for fitting a Logistic Regression model using L_1 regularization on the different EpiGEN datasets. Here, the (0, 0.1) column under MAF Datasets represents `MAF_0_01`, for example.

	MAF Datasets											
	(0, 0.1)	(0.1, 0.2)	(0.2, 0.3)	(0.3, 0.4)	(0.4, 0.5)	(0.5, 0.6)	(0.6, 0.7)	(0.7, 0.8)	(0.8, 0.9)	(0.9, 1)	DIFF_LD	SAME_LD
Train Error	0.28	0.22	0.17	0.12	0.12	0.13	0.13	0.13	0.12	0.28	0.16	0.090
Train ROC-AUC	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.97	1.00
Validation Error	0.32	0.32	0.32	0.30	0.29	0.29	0.29	0.29	0.29	0.33	0.25	0.17
Validation ROC-AUC	0.54	0.59	0.63	0.70	0.75	0.76	0.75	0.74	0.76	0.51	0.84	0.95
Test Error	0.32	0.33	0.32	0.31	0.32	0.38	0.30	0.30	0.32	0.33	0.45	0.35
Test ROC-AUC	0.55	0.58	0.65	0.69	0.71	0.75	0.75	0.74	0.74	0.49	0.79	0.81
Number of SNPs found	9826	9828	9829	9826	9828	9825	9823	9827	9824	9828	9827	9846
Disease SNPs found	10	10	10	10	10	10	10	10	10	10	9	10
Precision	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Recall	1	1	1	1	1	1	1	1	1	1	0.9	1

Table 4.2: Evaluation metrics of fitting a Logistic Regression model using L_2 regularization on the different EpiGEN datasets. Here, the (0, 0.1) column under MAF Datasets represents `MAF_0_01`, for example.

but out of sample accuracy and ROC-AUC are lower, suggesting overfitting. The out of sample ROC-AUCs are much worse ($<$0.8) in this table compared to using L_1 in Table 4.1. Even though we prioritize the recall in the first step of feature selection, it is still important to look at the error and ROC-AUC metrics, since that will give an indication of how relevant these features actually are. When we then look at the number of features found, precision, and recall, it corroborates the theory that we are overfitting here, since around 98% of the SNPs are used in the model. So not only do we get too many features to use for the second step in our pipeline, the features that are selected might be overfitted to our training set and not truly causal. This is not really a surprise, since L_2 is not generally used for feature selection. While L_2 maintains perfect recall, it does not reduce the feature space significantly, making it a suboptimal choice for regularization.

	MAF Datasets											
	(0, 0.1)	(0.1, 0.2)	(0.2, 0.3)	(0.3, 0.4)	(0.4, 0.5)	(0.5, 0.6)	(0.6, 0.7)	(0.7, 0.8)	(0.8, 0.9)	(0.9, 1)	DIFF_LD	SAME_LD
Train Error	0.27	0.18	0.15	0.14	0.13	0.12	0.13	0.13	0.12	0.32	0.15	0.093
Train ROC-AUC	0.83	0.95	0.97	0.97	0.98	0.98	0.98	0.99	0.99	0.67	0.96	0.99
Validation Error	0.30	0.23	0.21	0.20	0.17	0.18	0.19	0.19	0.19	0.33	0.19	0.14
Validation ROC-AUC	0.71	0.87	0.90	0.91	0.94	0.92	0.93	0.92	0.93	0.51	0.93	0.96
Test Error	0.35	0.24	0.21	0.21	0.27	0.29	0.49	0.21	0.22	0.33	0.35	0.30
Test ROC-AUC	0.81	0.88	0.90	0.91	0.91	0.92	0.92	0.92	0.91	0.9	0.91	0.87
Number of SNPs found	159	532	457	414	446	440	516	534	550	78	307	377
Disease SNPs found	6	10	10	10	10	10	10	10	10	10	10	8
Precision	0.038	0.019	0.022	0.024	0.022	0.022	0.019	0.012	0.018	0.12	0.033	0.024
Recall	0.6	1	1	1	1	1	1	1	1	1	1	0.8

Table 4.3: Evaluation metrics of fitting a Logistic Regression model using "adaptive L_1" regularization on the different EpiGEN datasets. Here, the (0, 0.1) column under MAF Datasets represents `MAF_0_01`, for example.

"Adaptive L_1" regularization

Then, we move on to look at the results in Table 4.3, which are the metrics from using "Adaptive L_1" as regularization in the first step. Looking at the errors and ROC-AUCs, training and out of sample performance are both good (>0.8), suggesting that the features found here are not just relevant to the training dataset. When we look at the number of features found using "adaptive" L_1, it is mostly around 300-500 features, which is a significant decrease from the original 10,000 SNPs in the dataset. In fact, for most datasets, using L_1 regularization and "adaptive L_1" regularization achieve the same results, with maybe a few more coefficients found in some datasets when using "adaptive" L_1. However, we see that for the `MAF_09_1` dataset, there is a significant improvement here, since using by using "adaptive L_1" regularization, we are able to retrieve back all of the causal SNPs in this round of feature selection. The good recall coupled with the fact that we are able to get good performance with a small number of features make "adaptive" L_1 a promising choice to use as regularization.

"Adaptive L_2" regularization

Compared to L_2 regularization, "adaptive L_2" regularization achieves very similar results, seen in Table 4.4. The out of sample performance of the model is low and the number of features selected is still too high; so while there is near-perfect recall,

	MAF Datasets											
	(0, 0.1)	(0.1, 0.2)	(0.2, 0.3)	(0.3, 0.4)	(0.4, 0.5)	(0.5, 0.6)	(0.6, 0.7)	(0.7, 0.8)	(0.8, 0.9)	(0.9, 1)	DIFF_LD	SAME_LD
Train Error	0.28	0.25	0.22	0.17	0.12	0.13	0.13	0.13	0.12	0.28	0.18	0.11
Train ROC-AUC	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.96	0.99
Validation Error	0.32	0.32	0.32	0.30	0.29	0.29	0.29	0.29	0.29	0.33	0.25	0.17
Validation ROC-AUC	0.54	0.59	0.63	0.70	0.75	0.76	0.75	0.74	0.76	0.51	0.84	0.95
Test Error	0.32	0.33	0.32	0.31	0.32	0.38	0.30	0.30	0.32	0.33	0.42	0.34
Test ROC-AUC	0.55	0.58	0.65	0.69	0.71	0.75	0.75	0.74	0.74	0.49	0.79	0.80
Number of SNPs found	9826	9828	9829	9826	9828	9825	9823	9827	9824	9828	9827	9846
Disease SNPs found	10	10	10	10	10	10	10	10	10	10	9	10
Precision	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Recall	1	1	1	1	1	1	1	1	1	1	0.9	1

Table 4.4: Evaluation metrics of fitting a Logistic Regression model using "adaptive L_2" regularization on the different EpiGEN datasets. Here, the (0, 0.1) column under MAF Datasets represents `MAF_0_01`, for example.

using L_2 or "adaptive L_2" as a regularization step will be computationally infeasible and not worthwhile since the performance is also not great.

Thus, we choose "adaptive L_1" as our method of regularization, since it is able to significantly reduce the feature space while maintaining good recall of our true disease SNPs.

4.1.2 Analysis of DML Results

	MAF Datasets											
	(0, 0.1)	(0.1, 0.2)	(0.2, 0.3)	(0.3, 0.4)	(0.4, 0.5)	(0.5, 0.6)	(0.6, 0.7)	(0.7, 0.8)	(0.8, 0.9)	(0.9, 1)	DIFF_LD	SAME_LD
Train Error	0.28	0.22	0.18	0.17	0.18	0.15	0.16	0.16	0.18	0.33	0.29	0.13
Train ROC-AUC	0.75	0.88	0.93	0.93	0.92	0.94	0.93	0.93	0.93	0.90	0.90	0.96
Validation Error	0.29	0.22	0.20	0.19	0.18	0.17	0.17	0.16	0.19	0.33	0.29	0.13
Validation ROC-AUC	0.71	0.88	0.91	0.92	0.93	0.94	0.94	0.94	0.93	0.49	0.73	0.96
Test Error	0.35	0.31	0.20	0.19	0.22	0.24	0.69	0.34	0.21	0.33	0.39	0.25
Test ROC-AUC	0.81	0.89	0.91	0.92	0.91	0.93	0.93	0.93	0.90	0.49	0.68	0.84
Number of SNPs found	5	11	10	10	10	12	11	11	13	1	7	6
Disease SNPs found	5	10	10	10	9	10	10	10	9	0	2	6
Precision	1	0.91	1	1	0.9	0.83	0.91	0.91	0.69	0	0.29	1
Recall	0.5	1	1	1	0.9	1	1	1	0.9	0	0.2	0.6

Table 4.5: Evaluation metrics of running causal orthogonal search using an "adaptive L_1" regularization in the first step. Here, the (0, 0.1) column under MAF Datasets represents `MAF_0_01`, for example.

After the first pass of feature selection using a logistic regression with "adaptive L_1" regularization, we now run causal feature selection via orthogonal search on the selected features from the first step, using a significance test with a p-value of 1e-5. Again, we look at metrics such as error, ROC-AUC, number of features found,

precision, and recall. We look at performance metrics like error and ROC-AUC to gain a sense of how well the chosen features can predict the outcome and we look at number of features found, precision, and recall to gain a sense of how well this pipeline works for causal feature selection. Table 4.5 shows the results of the pipeline on the synthetic datasets. For the performance metrics like error and ROC-AUC, most scenarios do well in both training and out of sample. In fact, the performance is very comparable to the performance of the Logistic Regression using L_1 or "adaptive L_1" regularizations as in Tables 4.1 and 4.3, even though the feature set that the model is trained on is much smaller. This suggests that the features we find after DML are sufficient for the decision.

To further support this thought, we notice that for most scenarios, namely when the MAF ranges are non-extreme, this pipeline does extremely well in preserving good recall. However, in extreme MAF ranges like in the dataset `MAF_0_01` and `MAF_09_1`, this pipeline fails in finding many of the true disease SNPs that are causal. This is probably due to the MAF range being too extreme, meaning that there is not really a differentiation in these features between case and control outcomes. Thus, it might be hard to get any signal here. Because we were able to select more disease SNPs in the first round of feature selection, the failure here suggests that this is a point of failure in the causal feature selection step.

Moreover, looking at both the `DIFF_LD` and `SAME_LD` dataset, few of the true disease SNPs were found to be significant, suggesting another point of failure. We hypothesize that the causal orthogonal feature selection step is hurt by the correlation. When compared to Logistic Regression using any regularization method, DML performs worse for `DIFF_LD` and around the same for `SAME_LD`.

How the confidence of DML affects the results of `DIFF_LD` and `SAME_LD`

To investigate this failure further, we range the significance test's p-value to see how confident the algorithm is in picking out causal features. Figures 4-1 and 4-2 show the same evaluation metrics we have been looking at with for the `DIFF_LD` and `SAME_LD` datasets when using p-values in [1e-7, 1e-5, 1e-2, 1e-1]. To reiterate, we look at cross

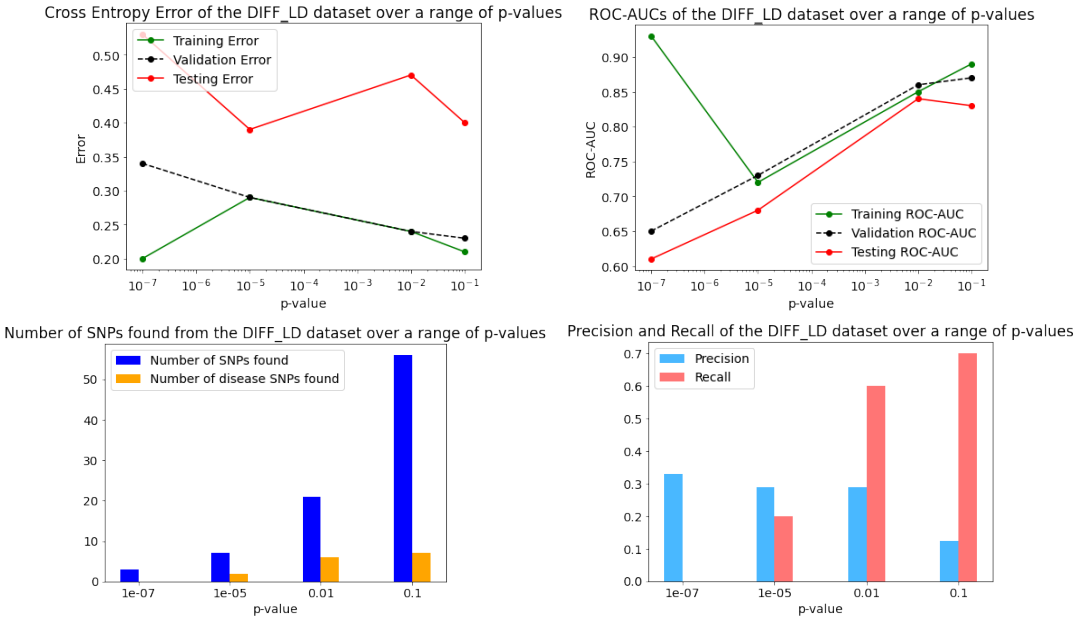

Figure 4-1: Evaluation metrics for `DIFF_LD` over a range of p-values. The cross entropy error and ROC-AUC score using significant features are on the top. The number of SNPs found, number of disease SNPs found, precision, and recall are on the bottom. See Table A.1 for more details.

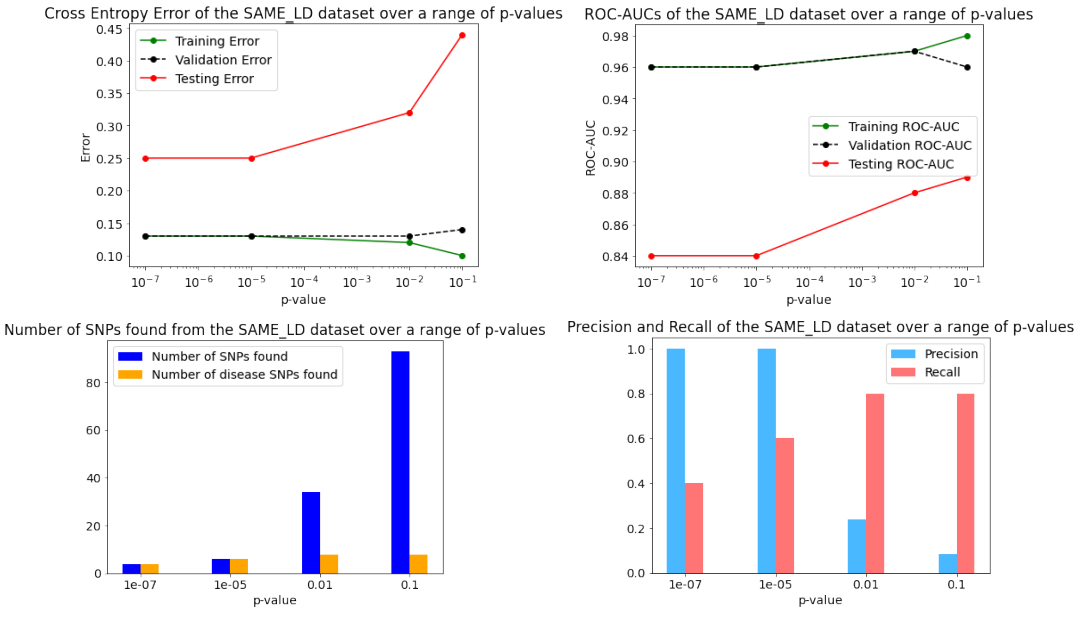

Figure 4-2: Evaluation metrics for `SAME_LD` over a range of p-values. The cross entropy error and ROC-AUC score using significant features are on the top. The number of SNPs found, number of disease SNPs found, precision, and recall are on the bottom. See Table A.2 for more details.

entropy loss and ROC-AUC of the fitted Logistic Regression model on the selected DML features, the number of selected SNPs, the number of disease SNPs selected, precision, and recall. For `DIFF_LD`, we see that the out of sample cross entropy loss and ROC-AUC increases as we increase the p-value, suggesting a better fit as we are becoming less confident. This could be attributed to the fact that we do not see any causal features selected in the smallest p-value of 1e-7 and we see an increase of true disease SNPs selected as we increase the p-value. For `SAME_LD`, we see that out of sample ROC-AUC also increases as we increase the p-value, which could also be attributed to selecting more true disease SNPs as we increase the p-value. For both datasets, we see that as we increase the p-value, corresponding to a less confident selection, we are able to pick out more of the true causal features. However, with the increase in p-value, we also end up being more imprecise and pick out SNPs that are not truly causal.

Analysis of the correlation structure of significant features of `SAME_LD`

For the `SAME_LD` dataset, we are able to achieve perfect precision for p-values smaller than 1e-2, but once we relax the significance test by increasing the p-value, we see that we get a mix of true disease SNPs and non-disease SNPs that were selected to be causal. This means that the algorithm is having trouble distinguishing between causal and non-causal features here, but at least these are low confidence predictions.

Furthermore, for the `SAME_LD` dataset, Figures 4-3 and 4-4 show the correlation matrices of the features that DML picks out to be causal on the top row and the correlation matrices of the features that DML picks out to be causal plus the true disease SNPs that weren't found added in on the bottom row. The green dotted lines indicate the indices of the disease SNPs that DML was able to select and the red dotted lines indicate the indices of the disease SNPs that DML was not able to select. Figure 4-3 shows the correlation matrices of the significant features using a p-value of 1e-5, and even though all the disease SNPs are correlated by construction, we see that in particular, the disease SNPs that were found by DML are more correlated to one another than the disease SNPs that were not found by DML. In Figure 4-

4, plotted are the correlation matrices of the selected causal features when using a significance test with a p-value of 1e-1. We see that the causal SNPs that DML finds are very correlated, particularly centered around where all of the true disease SNPs lie (circled). Since we were not seeing non-disease SNPs get selected when the p-value was lower, in this regime, DML does a good job of picking out what is truly causal versus what is just correlated with a causal feature. We also see, though, that there are other blocks of high correlation that aren't centered around any of the true disease SNPs (circled), and these blocks of high correlation are more pronounced than the correlation structure in the underlying dataset. So, we are seeing arbitrary SNPs being selected here. This suggests that regardless of whether a selected feature is causal or not, once DML deems a feature to be causal, it's also likely to find the features correlated with the selected feature to be causal. For this regime though, we see the correlation among non-causal features when we relax the statistical significance test heavily.

Analysis of the correlation structure of significant features of `DIFF_LD`

On the other hand, in the `DIFF_LD` dataset, performance is a bit worse because we are confidently predicting both true disease SNPs and non-disease SNPs to be causal. Even when we set the confidence to be very high (p-value=1e-7), we see that DML picks out non-disease SNPs to be causal, as in Figure 4-1; and in fact, no true disease SNPs are found at this confidence level. Figures 4-5 and 4-6 show the correlation matrices of the features deemed causal by DML using p-values of 1e-5 and 1e-1 respectively. They also show the correlation matrices of the significant features deemed causal plus the true disease SNPs that were not found to be causal by DML. For the test using a p-value of 1e-5 in Figure 4-5, we see that the features that DML does pick out in the top row do not show strong correlation with one another. However, if we observe the bottom row and look at the correlation between the selected causal features and the true disease SNPs that were *not* found, we see there is more correlation there. This implies that DML cannot differentiate between a causal feature and what is only correlated with a causal feature, and falsely picks

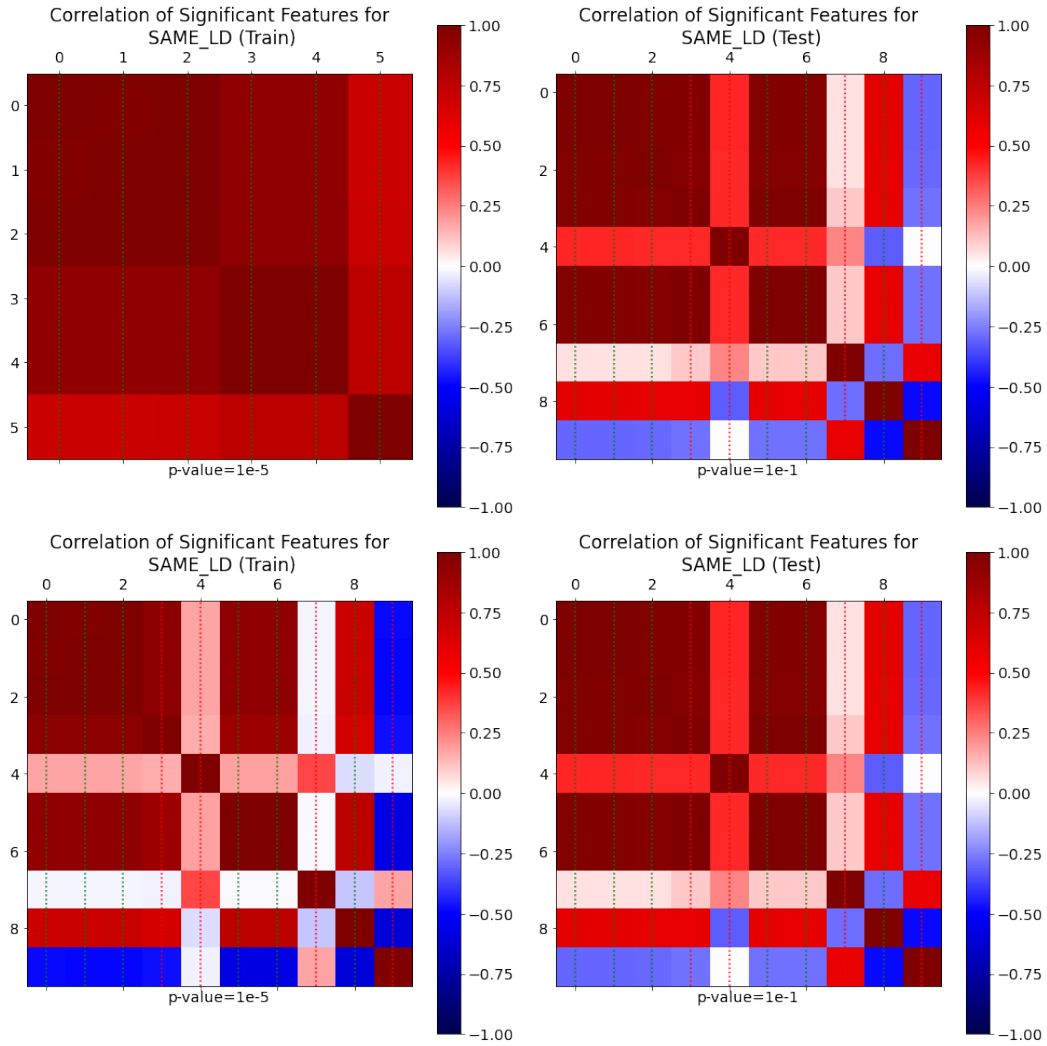

Figure 4-3: The first row denotes correlation plots over the significant features that DML deems causal using a p-value of 1e-1 for the `SAME_LD` dataset for both training and testing. The green lines denote where the true disease SNPs lie. The second row shows the correlation plots over significant features that DML deems causal using a p-value of 1e-5 and the disease SNPs that DML did not find added in. The green lines denote the disease SNPs that DML found, and the red lines denote the disease SNPs that DML did not find to be causal that we added in for analysis. Circled are the places where we see more correlation between the significant features.

the correlated features to be causal. More insidiously, DML still chooses features that are not causal with a high confidence.

In Figure 4-6, where the features are selected with a p-value of 1e-1, we see that there is much more correlation in both rows, meaning that the found causal features are correlated with each other and the found causal features are more correlated with

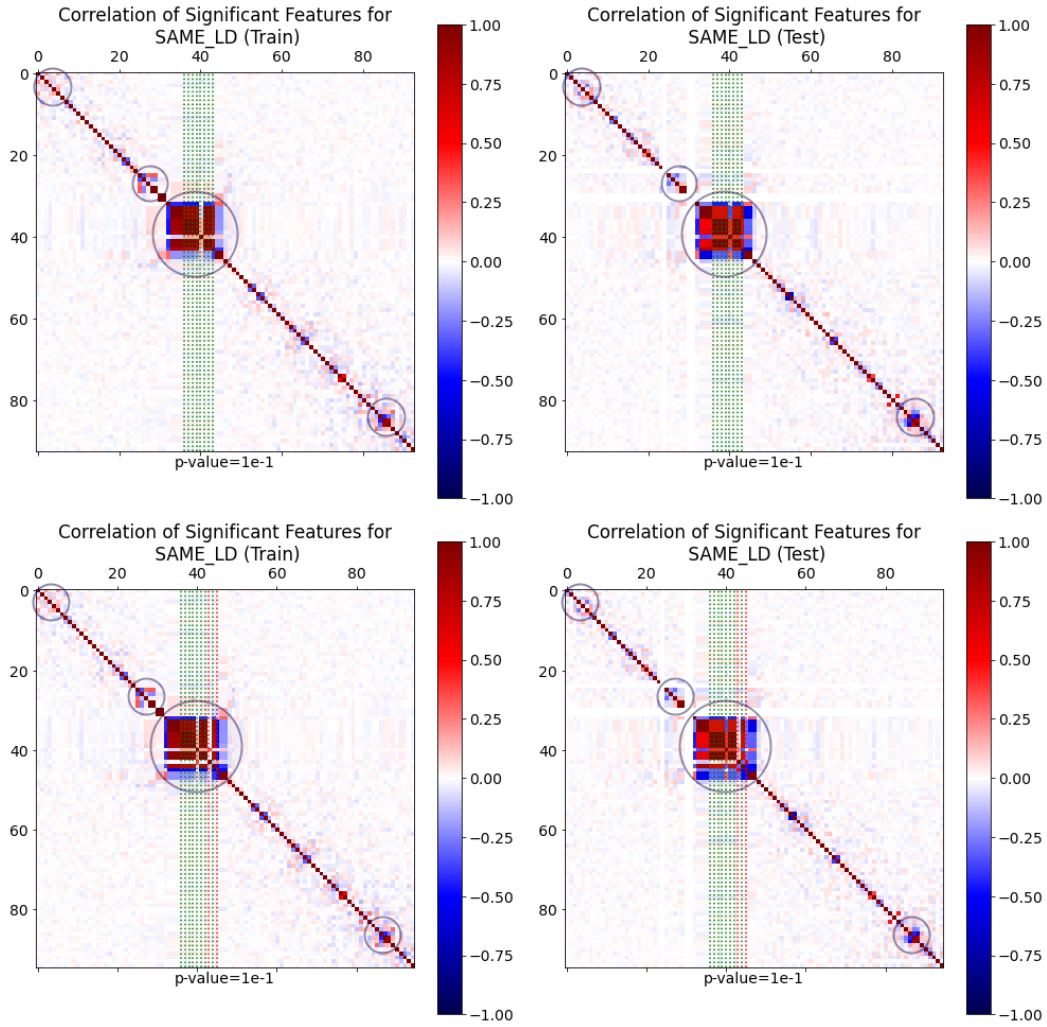

Figure 4-4: The first row shows the correlation matrices of the significant features that DML deems causal using a p-value of 1e-1 for the `SAME_LD` dataset for both training and testing. The green lines denote where the true disease SNPs lie. Circled are the correlation structures between the significant features that could be an issue. The second row shows the correlation plots over significant features that DML deems causal using a p-value of 1e-1 and the disease SNPs that DML did not find added in. The green lines denote the disease SNPs that DML found, and the red lines denote the disease SNPs that DML did not find to be causal that we added in for analysis.

the true disease SNPs. We see that as the significance test becomes more lenient, DML is more likely to select SNPs that are simply correlated with a true disease SNP as causal. Here though, we see that the method's imprecision is due to most of the selected features having some correlation with a true disease SNP, unlike in the `SAME_LD` dataset, where arbitrary features without any correlation to the true

disease SNPs were being picked. But, the correlation structure between the selected causal features in Figure 4-6 supports the finding that once DML considers one feature causal, it's likely to consider other correlated features as causal too.

This could be an issue, since the real dataset is likely to have linkage disequilibrium blocks, where SNPs are more correlated with one another. So, if DML outputs SNPs that are correlated with one another, since there is no ground truth model in the real dataset, it would be hard to tell which SNPs are truly causal.

In summary, on EpiGEN, DML performs exceptionally well in the datasets where we do not see much correlation; see Table 4.5. We are able to maintain high precision and recall for those settings as well as low cross entropy error and a high ROC-AUC score. When the features are correlated, DML selects features that are correlated to the true causal features that were selected and features that are correlated to the true causal features that were not selected; see Figures 4-3, 4-4, 4-5, and 4-6.

4.2 NYGC Dataset

For this project, we were limited in the amount of computational resources we were able to obtain (see Appendix B), so running the pipeline on the entire human genome with all chromosomes concatenated together would be infeasible. For this reason, we focused on areas of the chromosome where we know are associated with ALS. We had access[1] to a table of compiled genes and SNPs on the genes where previous studies have shown associations to ALS, which was generated by Matthew Harms (MD, Associate Professor of Neurology at Columbia). Genes in this pathogenic variant table include SOD1, PFN1, TARDBP, OPTN, VCP, and UBQLN2. Since there is no ground truth model of determining whether a SNP is causal or not, we break down the analysis into looking at certain genes where we have some biological knowledge for evaluation and looking at all of the genes where we use statistical methods for evaluation.

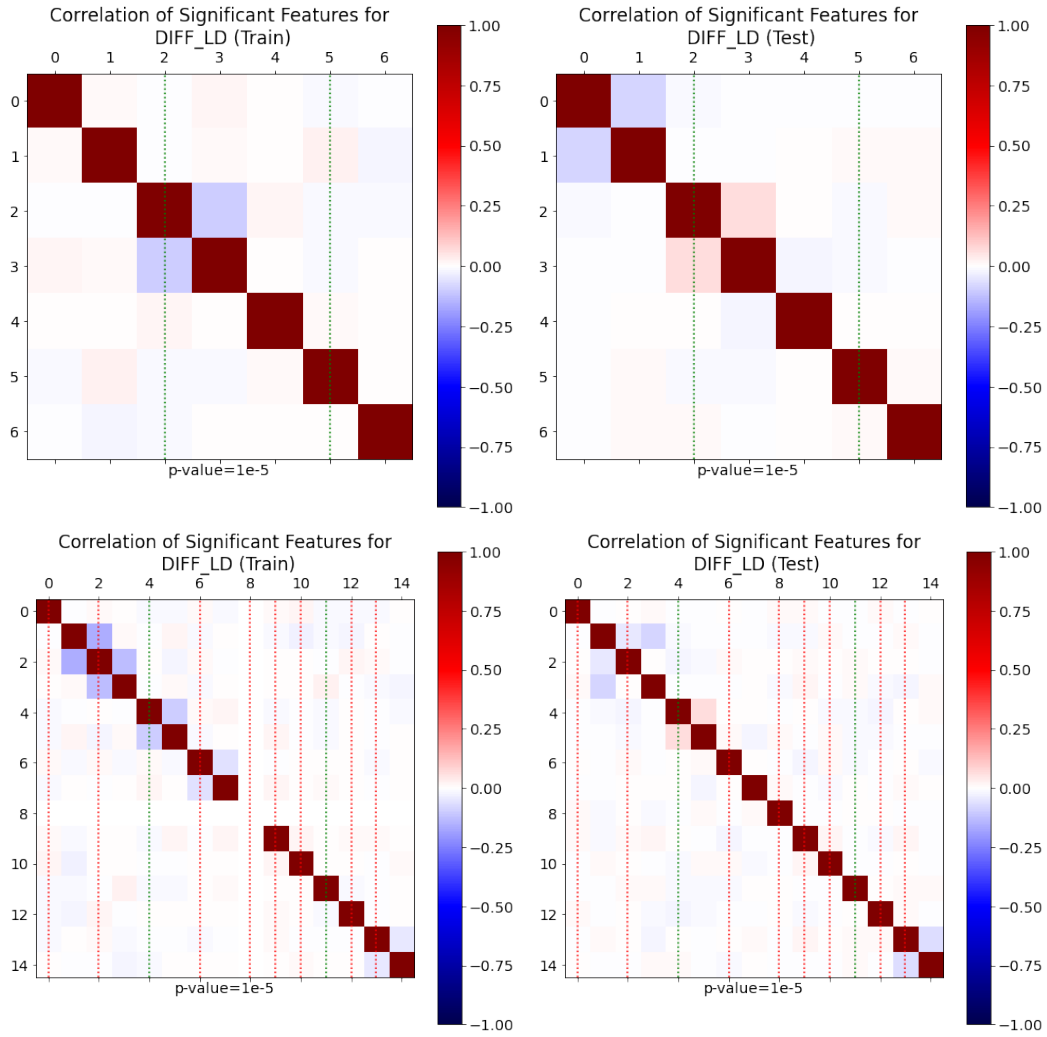

Figure 4-5: The first row denotes correlation plots over the significant features that DML deems causal using a p-value of 1e-5 for the `DIFF_LD` dataset for both training and testing. The green lines denote where the true disease SNPs lie. The second row shows the correlation plots over significant features that DML deems causal using a p-value of 1e-5 and the disease SNPs that DML did not find added in. The green lines denote the disease SNPs that DML found, and the red lines denote the disease SNPs that DML did not find to be causal.

4.2.1 Analysis of SOD1 Gene

We first chose to run the pipeline on the SOD1 gene, since SOD1 has been widely studied and shown to have mutations that are associated with ALS [23, 25]. Within the dataset, we were able to locate 224 SNPs on the SOD1 gene for each individual.

[1] via personal communication

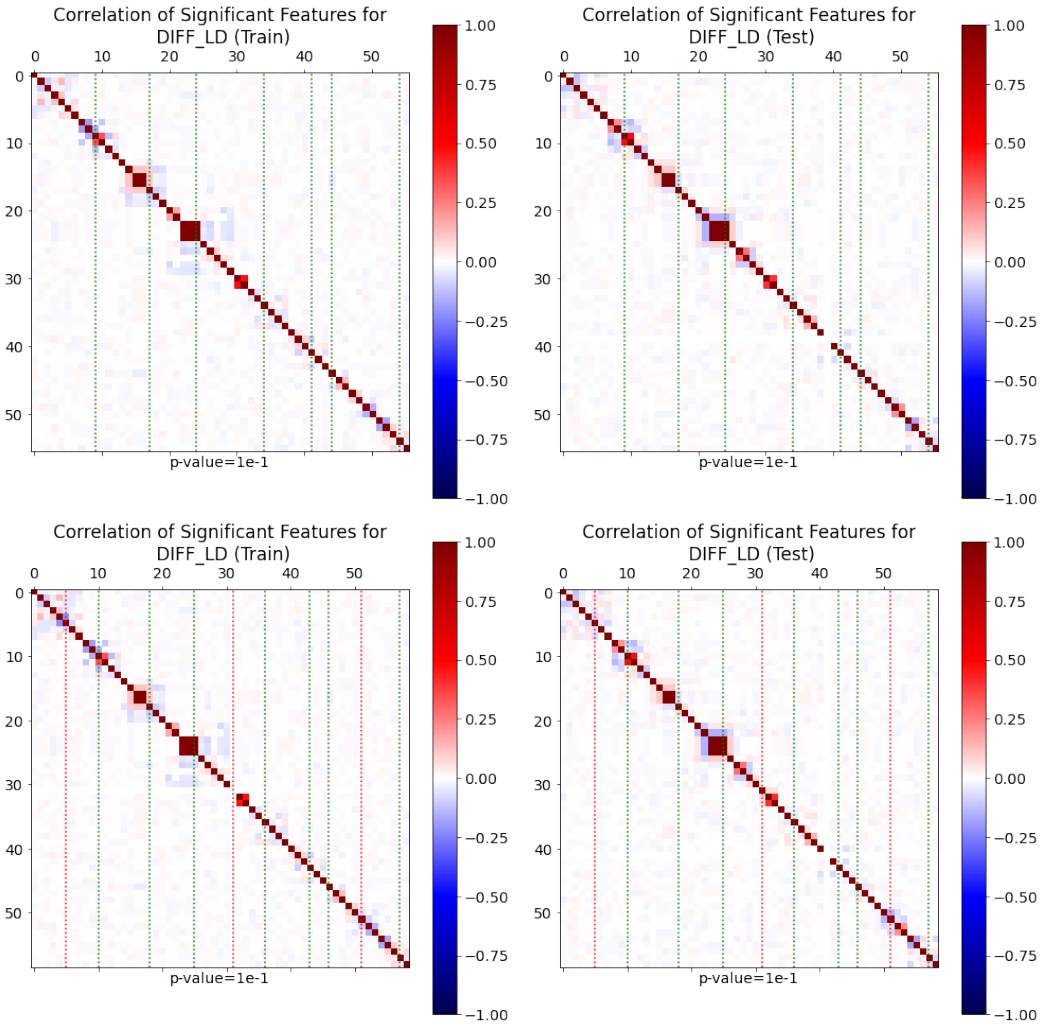

Figure 4-6: The first row denotes correlation plots over the significant features that DML deems causal using a p-value of 1e-1 for the `DIFF_LD` dataset for both training and testing. The green lines denote where the true disease SNPs lie. The second row shows the correlation plots over significant features that DML deems causal using a p-value of 1e-1 and the disease SNPs that DML did not find added in. The green lines denote the disease SNPs that DML found, and the red lines denote the disease SNPs that DML did not find to be causal.

In this set of experiments, we run our pipeline to retrieve the causal SNPs and then fit a Logistic Regression to those features and then compare those metrics to fitting a Logistic Regression model and Random Forest model on the entire SOD1 dataset. Because the dataset is so imbalanced (1,676 cases and 4,214 controls), we use ROC-AUC as a metric.

Table 4.6 shows evaluation metrics for the different models, including training and

	Model			
	Logistic Regression (L_1 Regularization)	Random Forest	Causal Orthogonal Search (DML) (L_1 Regularization in first step)	Causal Orthogonal Search (DML) (L_2 Regularization in first step)
Training ROC-AUC	0.67	0.65	0.66	0.661
Testing ROC-AUC	0.64	0.62	0.62	0.63
Number of Features	52	24	33	42

Table 4.6: Training and testing ROC-AUC scores for baseline models and Causal Orthogonal Search using L_1 and L_2 regularization in the first step for the SOD1 gene. Also included is the number of features that each model deems relevant to the prediction problem. For Logistic Regression, it is the number of non-zero coefficients and for Causal Orthogonal Search, it is the number of causal features.

testing ROC-AUC and the number of features selected from each method. When comparing the ROC-AUCs between the Logistic Regression and Random Forest model using all of the features, the linear model performs slightly better, indicating that we don't gain much more by using a non-linear model. Then, when comparing the performance between the Logistic Regression on all of the features and Logistic Regression on the selected causal features in both regularization settings, the out of sample performance is slightly better when using all of the SNPs. However, both versions of DML, using L_1 and L_2 regularization in the first step, were able to pick fewer features while maintaining a relatively close ROC-AUC to the model trained on all of the features, suggesting that the SNPs that DML found capture substantial predictive power.

After the coefficients from DML were found, we were able to use the found SNPs' chromosome and positions to compare to the table of compiled SNPs that have associations to ALS. Many of the SNPs found were intronic variants, meaning SNPs found on non-coding region, which unfortunately, the pathogenic variant table does not contain many of. We got one match to the table, the SNP with rsid rs121912442, which has an amino acid variant of Ala5Val and cDNA variant of 14C>T. This is in line with other studies which hypothesize this variant to be pathogenic [11, 33].

4.2.2 Analysis of PFN1 Gene

Then, we looked at the PFN1 gene, since according to Li et al. [14], there are many eQTLs [18] on chromosome 17, which PFN1 is also on. In our dataset, there were 64

	Model			
	Logistic Regression (L_1 Regularization)	Random Forest	Causal Orthogonal Search (DML) (L_1 Regularization in first step)	Causal Orthogonal Search (DML) (L_2 Regularization in first step)
Training ROC-AUC	0.65	0.65	0.62	0.62
Testing ROC-AUC	0.65	0.64	0.60	0.60
Number of Features	57	19	10	9

Table 4.7: Training and testing ROC-AUC scores for baseline models and Causal Orthogonal Search using L_1 and L_2 regularization in the first step for the PFN1 gene. Also included is the number of features that each model deems relevant to the prediction problem. For Logistic Regression, it is the number of non-zero coefficients and for Causal Orthogonal Search, it is the number of causal features.

SNPs that are on the PFN1 gene. We run the same comparison to the baselines as we did for SOD1, and those results are shown in Table 4.7. Again, the out of sample ROC-AUC is similar between a Logistic Regression model fit on all of the features and a Random Forest model fit on all of the features, showing us that a non-linear model is not necessary for this prediction problem. Also similar to the SOD1 analysis, we see that DML is able to significantly reduce the feature space, but sacrifices some performance when a linear model is trained on the DML feature space. In this case, the out of sample performance decreases more than in the SOD1 case, suggesting that these selected SNPs do not capture as much predicted power as the selected SNPs on the SOD1 gene did. Also, interestingly, between using L_1 and L_2 regularization in the first step, the performance is the same but DML using L_1 regularization finds one more causal SNP than DML using L_2, and all other selected SNPs are the same.

Of all the eQTLs in [14], three of them are on the PFN1 gene and DML using L_1 regularization was able to recover all three of them, while DML using L_2 regularization was able to recover two out of the three. The three eQTLs on the PFN1 gene are rs79843668, rs238243, and rs1859433 [14], and DML using L_2 was unable to retrieve rs1859433. However, since the performance is the same between the two methodologies, it is unclear whether the extra eQTL found by DML using L_1 regularization is just not causal or if the SNPs found by DML are correlated to that eQTL and capture the effects of that eQTL.

To investigate this, we looked at the correlation matrices of the found significant features, and we were specifically interested in the correlation between the eQTLs

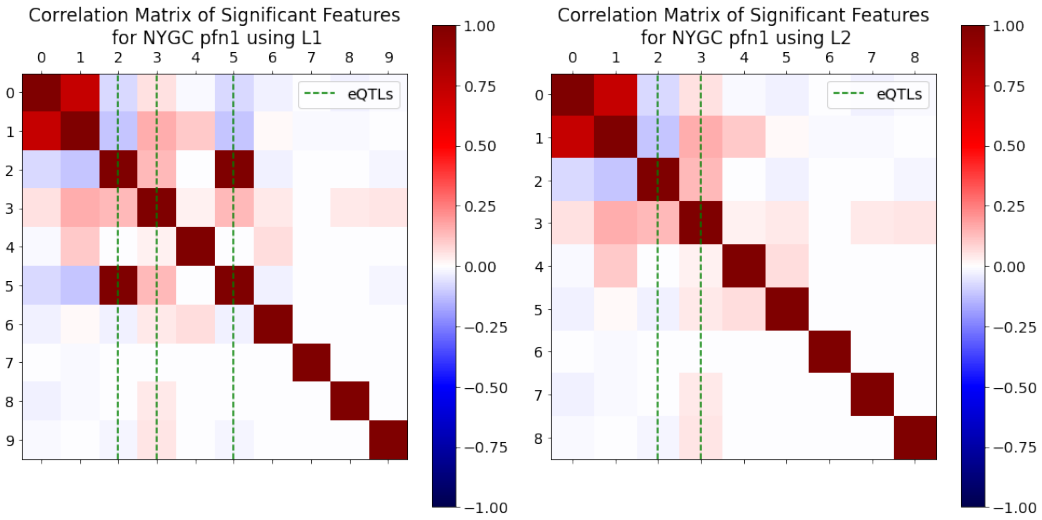

Figure 4-7: The correlation matrices of significant features for PFN1 using L_1 regularization as a first step (left) and L_2 regularization as a first step (right). The green lines denote the eQTL matches we found. The rightmost eQTL in the left figure is the extra eQTL that DML using L_1 regularization found. All other significant features are the same.

and the other SNPs, seen in Figure 4-7. We immediately notice the strong correlation between the first eQTL and the third eQTL in the left figure, where the significant SNPs were found using L_1 as a first step. We also see that the third eQTL is correlated to other significant features as well, which explains why adding this eQTL in does not affect the performance of the final Logistic Regression. We observed the same effect that was happening when running DML on the synthetic datasets: DML tends to pick out features that are correlated with one another. So while we were able to retrieve SNPs that are thought to be associated with ALS, we need to validate whether these SNPs are truly causal.

4.2.3 Analysis of Genes Associated with ALS

Finally, we ran the pipeline on all 12 genes from the table, which include VCP, TARDBP, OPTN, CCNF, SOD1, TBK1, ALS2, CHCHD10, PFN1, UBQLN2, VAPB, and FUS. We concatenated all the SNPs on these genes for a feature set size of 6,968 for each individual. Again, we compare the results a Logistic Regression model trained on the significant features found from DML to a Logistic Regression model trained

	Model			
	Logistic Regression (L_1 Regularization)	Random Forest	Causal Orthogonal Search (DML) (L_1 Regularization in first step)	Causal Orthogonal Search (DML) (L_2 Regularization in first step)
Training ROC-AUC	0.85	0.84	0.75	0.80
Testing ROC-AUC	0.78	0.70	0.72	0.76
Number of Features	287	225	31	264

Table 4.8: Training and testing ROC-AUC scores for baseline models and Causal Orthogonal Search using L_1 and L_2 regularization in the first step for all the genes in the pathogenic variant table.

on all features and a Random Forest Classifier model trained on all features, as seen in Table 4.8. Immediately, we can notice that the ROC-AUCs across all columns are higher than just using SOD1 or PFN1 by themselves, meaning that these added genes have a strong effect on the classifiers. Similar to what we were seeing before on the SOD1 and PFN1 datasets, the performances of the Logistic Regression on all of the features and the Random Forest Classifier on all of the features are similar, again suggesting that a non-linear model does not help much here.

The out of sample ROC-AUC of DML with L_1 regularization is slightly lower than the out of sample ROC-AUC from the overall Logistic Regression, but, even with fewer features, does better than the SOD1 and PFN1 Logistic Regression models using all of the features. This is expected, since this dataset consists of more genes that are associated with ALS, so it's likely that these added features are important. DML with L_1 regularization was able to significantly reduce the number of features, and we were able to retrieve the same SOD1 SNP as in the previous section as well as a SNP from TBK1, rs201970436, which is also included in the table of pathogenic variants that we have. Again, we saw that many of the other SNPs that were selected by DML were intronic variants, and when we looked at metrics such as the CADD score, which looks at how disastrous mutations of those SNPs are in general [22], or conservation scores of these SNPs, which tells us how unchanged a SNP was through evolution, nothing really stood out.

The out of sample ROC-AUC of DML with L_2 regularization is slightly lower than that of Logistic Regression trained on all features and slightly higher than that of DML with L_1 regularization, probably because the number of features that DML

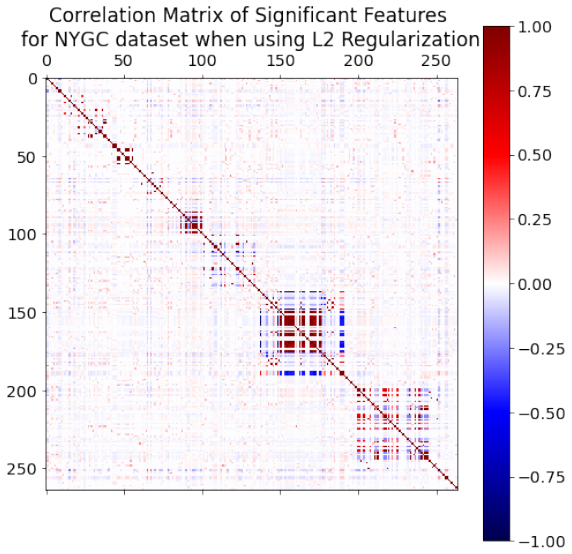

Figure 4-8: Correlation Matrix of the significant features found by DML when using L_2 Regularization in the first step. This selection is done from all of the genes in the pathogenic variant table.

with L_2 regularization finds is in between those models. We see that the feature space is not reduced by that much here, and the amount of features found is comparable to what Logistic Regression with L_1 regularization finds, without the DML step. When we look at Figure 4-8, we see strong block patterns, indicating high levels of correlation between the selected SNPs. This is a pattern we've seen many times already, both in the EpiGEN dataset and the NYGC dataset. When correlation between features is present in the dataset, DML tends to select significant features that are correlated with one another. This problem could have been exacerbated when using L_2 in the first step, because the correlated features might not have been filtered out there. But since we do not know what are the true causal SNPs and what are just correlated with the causal SNPs, we cannot reduce the correlation too much in the first step.

Out of the 264 SNPs found from DML using L_2 regularization, we retrieve the same variants in the table as we did in DML using L_1 in addition to rs138901914 from FUS, but this variant is thought to be benign. The other SNPs we found were mostly intronic variants and not in the table.

Another issue we faced here was not that many SNPs that were in our pathogenic variant table were found in the NYGC dataset. We set up a mapping from the variants

	Model				
	Logistic Regression (L_1 Regularization)	Logistic Regression (L_2 Regularization)	Random Forest	Causal Orthogonal Search (DML) (L_1 Regularization in first step)	Causal Orthogonal Search (DML) (L_2 Regularization in first step)
Training ROC-AUC	0.61	0.61	0.63	0.55	0.57
Testing ROC-AUC	0.60	0.60	0.61	0.55	0.56
Number of Features	11	33	15	2	5

Table 4.9: Training and testing ROC-AUC scores for baseline models (with the added Logistic Regression using L_2 regularization model) and Causal Orthogonal Search using L_1 and L_2 in the first step for all the SNPs in the pathogenic variant table.

on the table to the chromosome and positions of the SNPs, but only found 37 SNPs in our dataset. Table 4.9 show us the same evaluation metrics we have been using for these 37 SNPs. Noticeably, the ROC-AUCs across the models are lower than that from the SOD1, PFN1, and concatenation of all the genes datasets that we saw. This suggests that the SNPs that we do have in the NYGC dataset that also intersect with the pathogenic variant table do not contain enough predictive power.

We also added the metrics for using Logistic Regression with L_2 regularization to do a comparison of the models when using a similar amount of features for the prediction problem. In particular, we compare column 3 from Table 4.8 which are the metrics for a Logistic Regression using 31 SNPs from all of the ALS-associated genes to column 2 from Table 4.9 which are the metrics for a Logistic Regression using 33 SNPs from the pathogenic variant table. We notice that the performance using 31 features from all the genes does significantly better, which also suggests that perhaps we are missing SNPs from the pathogenic variant table in our dataset to account for the worse performance in Table 4.9 or that the intronic variants found from using all the genes play a big role in this prediction problem. This could also mean that there is more predictive power in the other parts of genes that are not the 37 SNPs from the pathogenic variant table. Since the pathogenic variants table do not include that many introns, there is a need to study the effects of these intronic variants and their causal nature in further detail.

To summarize, we applied DML on the genes SOD1, PFN1, and a concatenation of ALS-associated genes from the NYGC dataset. We noticed that between a Logistic Regression and Random Forest classifier on these datasets, Logistic Regression outperformed slightly, indicating that a non-linear model does not add substantial

predictive power. We also noticed that DML usually was able to significantly reduce the feature space while maintaining slightly lower performance than the Logistic Regression using all the features; see Tables 4.6, 4.7, 4.8, and 4.9. When compared to EpiGEN, we noticed the same pattern of DML selecting features with high correlation; see Figures 4-7 and 4-8.

Chapter 5

Discussion

In this thesis, we applied a new framework of selecting causal features from genetics data. We applied Causal Feature Selection Via Orthogonal Search [20] on EpiGEN generated datasets and the New York Genome Center ALS dataset. We noticed that when we induced correlation between features in the EpiGEN datasets to simulate linkage disequilibrium, DML picked significant coefficients that were very correlated with each other, correlated to true disease SNPs that were also picked by DML, and correlated to true disease SNPs that DML missed. This shows that when causal features and non causal features are correlated, instead of picking the causal features, DML is susceptible to picking the features that are simply correlated with the causal features and miss the truly causal ones.

When we analyzed the results of DML on the NYGC dataset, we noticed similar patterns. The significant features found were also quite correlated with one another, especially when looking at the PFN1 gene and the SNPs from all ALS-associated genes when using L_2 regularization in the first step. Thus, to analyze if our SNPs were truly causal, we compared results to a pathogenic variants table generated by Matthew Harms and a table of eQTLS generated by Li et al. [14]. We were able to retrieve SNPs rs121912442 from SOD1, rrs20197043 from TBK1, and rs138901914 from FUS that intersected with the pathogenic variants table as well as three eQTLs on PFN1 which were rs79843668, rs238243, and rs1859433. Most of the SNPs we retrieved from DML were intronic variants, and we did notice that the presence of

these intronic variants increased the predictive power of the model. Because there have not been too many studies done on intronic variant effects on ALS and we noticed that DML tends to pick out correlated features, more work would have to be done here to determine whether these SNPs are truly causal.

The main setback in our analysis was the inability to confidentally determine if a found SNP was causal or not. In the future, it would be very valuable to see how this pipeline works on other datasets associated with other diseases where the causes are more clear, such as Type 1 Diabetes or certain cancers. From there, we could observe potential different pitfalls or advantages of the DML algorithm that we might not have seen here.

Furthermore, since we tested our pipeline on genes known to be associated with ALS, a natural next step would be to test our pipeline on parts of the human genome that aren't known to be associated with ALS. However, since there is less that is known about those genes, it would be important to first validate the pipeline well.

The prospect of using purely statistical methods on genomics data to classify ALS is very exciting. If validated, then this method is generalizable enough to apply on any genomic dataset that consists of the SNPs of each individual and the corresponding phenotypes. This method worked well in selecting a small subset of SNPs while still maintaining decent predictive power. Selecting a small amount of features that can explain the underlying distribution of the data is a very important and relevant problem, especially in a biological setting. Finding and understanding the causal features of such a debilitating disease like ALS can open up avenues for researching diagnoses, preventions, and treatment of ALS.

Appendix A

Tables

	p-value			
	1e-07	1e-05	1e-02	1e-01
Train Error	0.20	0.29	0.24	0.21
Train ROC-AUC	0.93	0.72	0.85	0.89
Validation Error	0.34	0.29	0.24	0.23
Validation ROC-AUC	0.65	0.73	0.86	0.87
Test Error	0.53	0.39	0.47	0.40
Test ROC-AUC	0.61	0.68	0.84	0.83
Number of SNPs found	3	7	21	56
Disease SNPs found	0	2	6	7
Precision	0.33	0.29	0.29	0.125
Recall	0	0.2	0.6	0.7

Table A.1: Evaluation metrics for a range of confidence scores for the `DIFF_LD` dataset

	p-value			
	1e-07	1e-05	1e-02	1e-01
Train Error	0.13	0.13	0.12	0.10
Train ROC-AUC	0.96	0.96	0.97	0.98
Validation Error	0.13	0.13	0.13	0.14
Validation ROC-AUC	0.96	0.96	0.97	0.96
Test Error	0.25	0.25	0.32	0.44
Test ROC-AUC	0.84	0.84	0.88	0.89
Number of SNPs found	4	6	34	93
Disease SNPs found	4	6	8	8
Precision	1	1	0.24	0.086
Recall	0.4	0.6	0.8	0.8

Table A.2: Evaluation metrics over a range of confidence scores for the `SAME_LD` dataset

Appendix B

Computational Resources

B.1 Engaging Cluster

To run the experiments on the synthetic datasets, we used the Engaging Cluster, which uses a slurm job scheduler to allocate resources to the cluster users. We were generally able to obtain CPU resources with 15 tasks per node and 10 GB of memory for 12 hours at a time relatively easily. Running the pipeline end to end using the synthetic dataset took on average 5-6 hours for each dataset setting.

B.2 C3DDB

For the real dataset, we needed a cluster that could handle more computational intensive requests, so we used the C3DDB (Commonwealth Computational Cloud for Data Driven Biology) cluster. Here is where we stored all the processed datafiles from the NYGC. We were able to obtain CPU resources with 16 tasks per node and 200GB of memory for 5 days at a time relatively easily. Since we experimented with different feature sets on the real data, running the pipeline using around 300 features into the DML step took around an hour and running the pipeline using around 5000 features into the DML step took around 90 hours.

Appendix C

List of Found SNPs

Here we compile the SNPs that DML found to be causal using both L_1 and L_2 regularization as a first step. A * denotes they were found in the pathogenic variants table or they are one of the eQTLs in [14].

C.1 DML Selected SNPs using L_1 Regularization in the first step with a p-value of 1e-3

rs121912442*, rs201970436*, rs200729007, rs41258154, rs55655838, rs185084517, rs78703800, rs1984061, rs676302, rs6602633, rs61933196, rs73120375, rs9788149, rs148489734, rs73530283, rs112957921, rs775807888, rs16956408, rs79843668, rs720426, rs530122552, rs79873474, rs73616272, rs6650814, rs4816405, rs131442

C.2 DML Selected SNPs using L_2 Regularization in the first step with a p-value of 1e-3

rs121912442*, rs201970436*, rs138901914*, rs200729007, rs142544531, rs144628752, rs72870033, rs545713318, rs533364416, rs144401164, rs563793948, rs386628445, rs111510738, rs143585777, rs2273348, rs192716110, rs184723358, rs78026355, rs148597168, rs146134526, rs183880610, rs118129726, rs190718349, rs186087319, rs535733859, rs7603563, rs79546324,

rs41258154, rs7589068, rs140257386, rs561327572, rs6752508, rs10200479, rs74422443, rs372823004, rs201295943, rs150282277, rs76163254, rs190341431, rs533950860, rs78461894, rs55655838, rs113524715, rs111659619, rs77709457, rs113016388, rs182849420, rs183092733, rs189282198, rs7601642, rs6743712, rs74666822, rs12470499, rs796962683, rs79078118, rs12469940, rs12473742, rs112292399, rs185084517, rs12470229, rs11888145, rs115585280, rs541223303, rs142151681, rs150288841, rs6435102, rs7560439, rs78703800, rs190369242, rs541588367, rs147940110, rs6435104, rs6435105, rs1030027, rs79825965, rs371286964, rs75261995, rs563385193, rs768675930, rs374391034, rs1984059, rs1984060, rs1984061, rs139707968, rs548638257, rs113230475, rs373785808, rs149380309, rs538458489, rs146658966, rs143997916, rs572085032, rs553743801, rs552494483, rs371677514, rs183638268, rs79913169, rs114119115, rs181586435, rs185948264, rs7905921, rs10752285, rs372211102, rs190945137, rs57057378, rs57199113, rs549441201, rs189419702, rs151065414, rs568470843, rs140438390, rs676302, rs80263050, rs553891747, rs386741045, rs386741045, rs10752287, rs184536495, rs185730053, rs61933191, rs117271805, rs144620061, rs56198126, rs139000280, rs200587014, rs185338419, rs73120375, rs10128757, rs140728178, rs10878175, rs76423079, rs117936898, rs568048370, rs538208188, rs561656725, rs61933202, rs4075094, rs78253940, rs558710329, rs7487627, rs79528635, rs148082871, rs78427520, rs73313896, rs544117930, rs6581571, rs148489734, rs13331598, rs530962466, rs144172354, rs76549515, rs4589553, rs556367468, rs182878285, rs28550541, rs543598536, rs544042257, rs12926827, rs28647184, rs4786283, rs571588465, rs4371158, rs4544244, rs12932220, rs139018313, rs192518497, rs12922325, rs569382499, rs929867, rs143668562, rs141242311, rs61732970, rs2735393, rs775807888, rs201544187, rs150529460, rs554281103, rs546409934, rs140177490, rs6026232, rs141683494, rs559895553, rs6015263, rs544757877, rs552819618, rs185072455, rs73179863, rs138312362, rs149641189, rs73296813, rs720426, rs2064383, rs555406860, rs530122552, rs552404182, rs6026252, rs78550036, rs6100057, rs532751236, rs2284883, rs540377639, rs185202276, rs6026255, rs386815632, rs2268920, rs542488313, rs530852146, rs149659282, rs13043088, rs146845627, rs79873474, rs184645205, rs7265867, rs2268916, rs565219336, rs2143611, rs6026271, rs77854718, rs193290516, rs2300760, rs12625788, rs2234490, rs573360646, rs118055109, rs189009703, rs115932469, rs6650814, rs4816405, rs17880795, rs17880227, rs549984048, rs17880490, rs1041740, rs4817420, rs114886035, rs131442, rs192883799,

rs186638654

Bibliography

[1] Brahim Aissani. Confounding by linkage disequilibrium. *Journal of Human Genetics*, 59(2):110–115, Feb 2014. ISSN 1435-232X. doi: 10.1038/jhg.2013.130. URL https://doi.org/10.1038/jhg.2013.130.

[2] A. Al-Chalabi, F. Fang, M. F. Hanby, P. N. Leigh, C. E. Shaw, W. Ye, and F. Rijsdijk. An estimate of amyotrophic lateral sclerosis heritability using twin data. *Journal of neurology, neurosurgery, and psychiatry*, 81(12):1324–1326, Dec 2010. ISSN 1468-330X. doi: 10.1136/jnnp.2010.207464. URL https://pubmed.ncbi.nlm.nih.gov/20861059. 20861059[pmid].

[3] David B Blumenthal, Lorenzo Viola, Markus List, Jan Baumbach, Paolo Tieri, and Tim Kacprowski. EpiGEN: an epistasis simulation pipeline. *Bioinformatics*, 04 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa245. URL https://doi.org/10.1093/bioinformatics/btaa245. btaa245.

[4] Jean-Simon Brouard, Flavio Schenkel, Andrew Marete, and Nathalie Bissonnette. The gatk joint genotyping workflow is appropriate for calling variants in rna-seq experiments. *Journal of Animal Science and Biotechnology*, 10(1):44, Jun 2019. ISSN 2049-1891. doi: 10.1186/s40104-019-0359-0. URL https://doi.org/10.1186/s40104-019-0359-0.

[5] A Buniello, JAL MacArthur, M Cerezo, LW Harris, J Hayhurst, C Malangone, A McMahon, J Morales, E Mountjoy, E Sollis, D Suveges, O Vrousgou, PL Whetzel, R Amode, JA Guillen, HS Riat, SJ Trevanion, P Hall, H Junkins, P Flicek, T Burdett, LA Hindorff, F Cunningham, and Parkinson H. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47, 2019. URL https://www.ebi.ac.uk/gwas/efotraits/EFO_0000253.

[6] William S. Bush and Jason H. Moore. Chapter 11: Genome-wide association studies. *PLOS Computational Biology*, 8(12):e1002822, Dec 2012. doi: 10.1371/journal.pcbi.1002822. URL https://doi.org/10.1371/journal.pcbi.1002822.

[7] C. Børsting and N. Morling. Single-nucleotide polymorphisms. In Jay A. Siegel, Pekka J. Saukko, and Max M. Houck, editors, *Encyclopedia of Forensic Sciences (Second Edition)*, pages 233 – 238. Academic Press, Waltham,

second edition edition, 2013. ISBN 978-0-12-382166-9. doi: 10.1016/B978-0-12-382165-2.00042-8. URL http://www.sciencedirect.com/science/article/pii/B9780123821652000428.

[8] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL https://doi.org/10.1111/ectj.12097.

[9] Annelot M. Dekker, Frank P. Diekstra, Sara L. Pulit, Gijs H. P. Tazelaar, Rick A. van der Spek, Wouter van Rheenen, Kristel R. van Eijk, Andrea Calvo, Maura Brunetti, Philip Van Damme, Wim Robberecht, Orla Hardiman, Russell McLaughlin, Adriano Chiò, Michael Sendtner, Albert C. Ludolph, Jochen H. Weishaupt, Jesus S. Mora Pardina, Leonard H. van den Berg, and Jan H. Veldink. Exome array analysis of rare and low frequency variants in amyotrophic lateral sclerosis. *Scientific Reports*, 9(1):5931, Apr 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-42091-3. URL https://doi.org/10.1038/s41598-019-42091-3.

[10] Agency for Toxic Substances and Disease Registry. Amyotrophic lateral sclerosis. April 2017. URL https://www.cdc.gov/als/WhatisALS.html.

[11] Cecilia Garcia, Jose Manuel Vidal-Taboada, Enrique Syriani, Maria Salvado, Miguel Morales, and Josep Gamez. Haplotype analysis of the first a4v-sod1 spanish family: Two separate founders or a single common founder? *Frontiers in Genetics*, 10:1109, 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.01109. URL https://www.frontiersin.org/article/10.3389/fgene.2019.01109.

[12] P Hu, R Jiao, L Jin, and M Xiong. Application of causal inference to genomic analysis:advances in methodology. *Frontiers in Genetics*, 9, 2018. doi: 10.3389/fgene.2018.00238.

[13] Sanger Institute. Hapmap 3. URL https://www.sanger.ac.uk/resources/downloads/human/hapmap3.html.

[14] Jonathan Li, Jie Wu, Ryan G. Lim, AnswerALS Consortium, Ernest Fraenkel, and Leslie Thompson. Functional genomics reveals gene regulatory networks and novel loci underlying als. Manuscript in preparation.

[15] Mayo Clinic Staff. Amyotrophic lateral sclerosis (als). August 2019. URL https://www.mayoclinic.org/diseases-conditions/amyotrophic-lateral-sclerosis/symptoms-causes/syc-20354022.

[16] Rita Mejzini, Loren L. Flynn, Ianthe L. Pitout, Sue Fletcher, Steve D. Wilton, and P. Anthony Akkari. Als genetics, mechanisms, and therapeutics: Where are we now? *Frontiers in neuroscience*, 13:1310–1310, Dec 2019. ISSN 1662-4548. doi: 10.3389/fnins.2019.01310. URL https://pubmed.ncbi.nlm.nih.gov/31866818. 31866818[pmid].

[17] National Institute on Neurological Disorders and Stroke. Amyotrophic lateral sclerosis (als) fact sheet. *NIH Publication*, 16(916), 2013. URL https://www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/Fact-Sheets/Amyotrophic-Lateral-Sclerosis-ALS-Fact-Sheet.

[18] Alexandra C. Nica and Emmanouil T. Dermitzakis. Expression quantitative trait loci: present and future. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 368(1620):20120362–20120362, May 2013. ISSN 1471-2970. doi: 10.1098/rstb.2012.0362. URL https://pubmed.ncbi.nlm.nih.gov/23650636. 23650636[pmid].

[19] Aude Nicolas et al. Genome-wide analyses identify kif5a as a novel als gene. *Neuron*, Mar 2018. doi: 10.1016/j.neuron.2018.02.027. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5867896/.

[20] Anant Raj, Stefan Bauer, Ashkan Soleymani, Michel Besserve, and Bernhard Schölkopf. Causal Feature Selection via Orthogonal Search. *arXiv e-prints*, art. arXiv:2007.02938, July 2020.

[21] Alan E. Renton, Adriano Chiò, and Bryan J. Traynor. State of play in amyotrophic lateral sclerosis genetics. *Nature neuroscience*, 17(1):17–23, Jan 2014. ISSN 1546-1726. doi: 10.1038/nn.3584. URL https://pubmed.ncbi.nlm.nih.gov/24369373. 24369373[pmid].

[22] Philipp Rentzsch, Daniela Witten, Gregory M Cooper, Jay Shendure, and Martin Kircher. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, 47(D1):D886–D894, 10 2018. ISSN 0305-1048. doi: 10.1093/nar/gky1016. URL https://doi.org/10.1093/nar/gky1016.

[23] Y. Sheng, M. Chattopadhyay, J. Whitelegge, and J. S. Valentine. SOD1 aggregation and ALS: role of metallation states and disulfide status. *Curr Top Med Chem*, 12(22):2560–2572, 2012.

[24] J. H. Stock. *Instrumental Variables in Statistics and Econometrics*, pages 205–209. Mar 2015. doi: 10.1016/B978-0-08-097086-8.42037-4. URL https://doi.org/10.1016/B978-0-08-097086-8.42037-4.

[25] Matthis Synofzik, Dario Ronchi, Isil Keskin, Ayse N. Basak, Christian Wilhelm, Claudio Gobbi, Anna Birve, Saskia Biskup, Chiara Zecca, Rubén Fernández-Santiago, Toomas Kaugesaar, Ludger Schöls, Stefan L. Marklund, and Peter M. Andersen. Mutant superoxide dismutase-1 indistinguishable from wild-type causes ALS. *Human Molecular Genetics*, 21(16):3568–3574, 05 2012. ISSN 0964-6906. doi: 10.1093/hmg/dds188. URL https://doi.org/10.1093/hmg/dds188.

[26] GATK Team. About the gatk best practices, Sep 2020. URL https://gatk.broadinstitute.org/hc/en-us/articles/360035894711-About-the-GATK-Best-Practices.

[27] GATK Team. Genotypegvcfs, May 2020. URL https://gatk.broadinstitute.org/hc/en-us/articles/360037057852-GenotypeGVCFs.

[28] GATK Team. The logic of joint calling for germline short variants, Apr 2021. URL https://gatk.broadinstitute.org/hc/en-us/articles/360035890431-The-logic-of-joint-calling-for-germline-short-variants.

[29] W van Rheenen et al. Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nature Genetics*, 48:1043–1048, 2016. doi: 10.1038/ng.3622.

[30] Christina Vasilopoulou, Andrew P. Morris, George Giannakopoulos, Stephanie Duguez, and William Duddy. What can machine learning approaches in genomics tell us about the molecular basis of amyotrophic lateral sclerosis? *Journal of personalized medicine*, 10, november 2020. doi: 10.3390/jpm10040247.

[31] Yixin Wang and David M. Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019. doi: 10.1080/01621459.2019.1686987. URL https://doi.org/10.1080/01621459.2019.1686987.

[32] Chi-Hong Wu, Claudia Fallini, et al. Mutations in the profilin 1 gene cause familial amyotrophic lateral sclerosis. *Nature*, 488, Feb 2013. doi: 10.1038/nature11280.

[33] Yeomin Yun and Yoon Ha. Crispr/cas9-mediated gene correction to understand als. *International journal of molecular sciences*, 21(11):3801, May 2020. ISSN 1422-0067. doi: 10.3390/ijms21113801. URL https://pubmed.ncbi.nlm.nih.gov/32471232. 32471232[pmid].