Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

On the Use of Machine Learning Methods for Genomic Prediction

Ciaran Michael Kelly

Supervisor: Russell McLaughlin

Department of Genetics & Microbiology

Trinity College Dublin

This thesis is submitted for the degree of

Doctor of Philosophy

Smurfit Institute of Genetics

January 2023

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work. I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement. I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

Ciaran Michael Kelly

January 2023

Summary

This thesis is concerned with exploring the use of machine learning in comparison to traditional linear methods for the genomic prediction of complex traits. Special attention is paid to the issue of confounding in large-scale population studies using genetic data and how best to ensure that the models developed from such data remain bias-free. The thesis explores several quantitative traits in *Arabidopsis thaliana* as well as human disease genetics using an amyotrophic lateral sclerosis patient database.

The first chapter of this thesis is a general introduction to the fields of quantitative genetics, genomic prediction, and machine learning. It covers both the history behind some of the concepts in the field as well as current best practices and techniques. The remaining chapters also carry smaller, more specific introductions to their respective topics.

The second chapter explores the use of genomic prediction for the prediction of complex traits in *Arabidopsis thaliana*. It finds that machine learning methods are sometimes able to statistically out-compete the traditional linear methods but that the overall gain in performance is generally modest. Performance was found to be dependent on the specific trait under consideration as well as feature selection method and input feature-set size. Best practices are discussed from this data and considerations on the interpretation of the results are discussed.

The third chapter focuses on a large dataset of cases and controls in a form of human motor neuron disease known as amyotrophic lateral sclerosis. It also investigates the ability of machine learning methods to improve upon standard practices in the field of human genomic prediction. The results in this chapter show that improvements upon the baseline model are

not guaranteed and any additional discriminatory power is again generally modest. Similarly to the second chapter, performance was found to be dependent on feature selection method and input feature-set size. The limitations of using amyotrophic lateral sclerosis data are also discussed.

The fourth chapter is concerned with the development of novel techniques for reducing bias in non-linear genomic prediction models. It discusses several new approaches to incorporating adversarial components into neural networks and how confounding may be measured and reduced during model learning. This chapter finds that the distance correlation between prediction and confounding variables may be a useful metric to track bias in model building and that certain adversarial network designs can help to mitigate this bias.

The final chapter gives a general discussion of the results from the previous chapters, as well as their implications, and directions for future work. Overall, this thesis supports the continued use of machine learning methods in the field of genomic prediction, especially as sample sizes and computational power grows. However, the results presented here also point towards moderate expectations of future performance gains. This thesis also cautions against the use of machine learning models using genomic data without serious attention being paid to the issue of confounder handling in model development. Some techniques to address this problem are proposed, and further work on this area is considered especially important to prioritize over the coming decade as both machine learning and genomic prediction models continue to be implemented clinically.

“Let us keep looking, in spite of everything. Let us keep searching. It is indeed the best method of finding, and perhaps thanks to our efforts, the verdict we will give such a patient tomorrow will not be the same we must give this man today.”

Jean-Martin Charcot (1889)

Acknowledgements

First and foremost I wish to thank my supervisor Russell McLaughlin. This PhD began over a coffee, where we excitedly discussed our views on complex disease genetics and the idea of employing machine learning to explore genetic architecture. There have been many ups and downs in my research since then, but you were always ready to encourage my good ideas, poke deserved holes in my less fruitful ones, and always keep me on the right track. Thank you for all the work you have put into supervising this project and I am especially grateful for the grace you showed me when I hit a COVID lockdown wall.

Secondly, I wish my thank all my lab-mates for the last four years. If nothing else, I come away from this experience with true friends. Mark Doherty, thank you for helping me with all my beginner questions in bioinformatics and for putting up with all of my day-to-day grievances. Ross Byrne, it is cliché to say but I probably could never have completed this project without all the help you've given me; from the smallest Linux question to the biggest questions on the latest genomic statistical techniques. I only hope I didn't annoy you *too* much. Jenny Hengeveld, our work didn't overlap as much but that didn't stop you from helping out wherever you could, including the occasional pick-me-up coffee. Your friendship and support were invaluable. Laura O'Briain, the pandemic certainly cut short our lab time together but your positive energy was always felt, even through a computer screen.

To Dan Bradley, Lara Cassidy and everyone in their teams, thank you for all the help, the discussions and the productive atmosphere you both have created in our little group of labs. Truthfully, the Smurfit Institute is filled with such wonderful people it would be impossible to

acknowledge everyone but I would also like to specifically thank Karsten Hokamp, Matthew Campbell, Aoife McLysaght and Brenda Campbell.

Conor Rossi, Dáire Gannon, Joseph Beegan and Thomas Dineen, you were my PhD brothers-in-arms and I can't wait to celebrate with you all. Laura Anderson and Dr. Abbie O'Brien, I am so proud of what you both have managed to achieve over the last four years.

To Laura Whelan, Dearbhaile Casey, Kristina Borcea and Joseph O'Callaghan, I am so lucky to have you all to rely on, to make me laugh and to just relax and enjoy each other's company. Laura, thank you for your support and advice over the last few years. All of our discussions hopefully kept both our research sharper and more interesting than it would have been otherwise. Every once in a while, a piece of chocolate would appear on my desk or at home and it's the small things that get you through the day. Joseph, thank you for keeping me distracted from frustrating scientific issues by replacing them with even more frustrating political ones. Kristina, I could always rely on you to be straight with me to either pick up the pace, or just let go a little, depending on the circumstances. I would be both figuratively and literally lost without you. Dearbhaile, from helping me in the beginning with small coding questions, right to the end when you swapped presentation slots with me as I finished my thesis, I can't thank you enough.

Ba mhaith liom mo bhuíochas a ghabháil le mo theaghlach fosta. Thug sibh uchtach dom an tionscnamh seo a chriochnú, 's bhí sibh i gconaí ann chun béim a chur ar mo shláinte agus chun sos beag a ghlacadh thall agus abhus. I will forever be indebted to the sacrifices my parents made in order for me to pursue my career in science to the best of my abilities. There is no way to truly pay you back but I will try my best.

This thesis could not have been done without the help of all the ALS patients who kindly shared their data with the hope of one day finding a cure for this devastating disease. We are not there yet, but there must come a day when the collective efforts of all the patients, health staff and researchers who fight ALS will change the prognosis considerably.

I also wish to thank all of my collaborators at Project MinE who have done fantastic work on ALS and whose data greatly facilitated my research. I owe a huge debt to my funders at Science Foundation Ireland who have been central to so much great science conducted on this island.

Finally, I would like to thank my partner Daniel Fernandes Moreira Neto who supported me every step of the way. You were able to lift me through all of my bad days and I owe you so much. I know it wasn't easy, especially during my final year, but I appreciated it more than you know. I promise to repay all my absences back in double. This road has led to many things, with perhaps many more to come, but best of all to you. Te amo muito.

Table of contents

List of figures

List of tables

Abbreviations

V_A Additive Variance. 5, 8

V_D Dominance Variance. 5, 7

V_E Enviornmental Variance. 4

V_G Genetic Variance. 4, 5

V_I Epistatic Variance. 5

V_P Phenotypic Variance. 4, 5

β Effect Size Estimate. 20, 76

β_0 Regression Constant/Y-intercept. 10

δ Change (Differentiation). 78, 80

λ Lambda Tuning Parameter. 21, 77–80, 82, 85, 87, 89, 98–100

$\mathscr{L}$ Loss. 20, 26, 78, 80

θ Neural Network Weights. 78, 80

ρ Pearson's Correlation Coefficient. 36–38, 44, 58, 78, 81, 82, 84–97, 100

$\boldsymbol{M}$ Number of Markers. 12, 21

N Number of Individuals. 9, 21

i *i*th Individual. 10, 21

j *j*th Marker. 12, 21

ALS Amyotrophic Lateral Sclerosis. 53–55, 57–59, 68–71, 81, 82, 84, 93–100, 104–108, 134, 135

API Application Programming Interface. 82

AUC Area Under the Curve. 14, 15, 56, 65–67, 69, 82, 97

AUROC Area Under the Receiver Operating Characteristic. 14

Bagging Bootstrap Aggregation. 24

BLUP Best Linear Unbiased Prediction. 11

bp Base Pair. 37, 38, 81, 82

BR-Net Bias Resilient Neural Network. 78, 82, 85, 89, 93, 95, 97, 98

CAD Coronary Artery Disease. 53, 105

cm Centimeters. 37

CNN Convolutional Neural Network. 27, 64–66

DANN Domain-Adversarial Neural Network. 76–80

DC Distance Correlation. 78, 79, 82–100, 106–108

DisCo Distance Correlation. xviii, 79, 82, 83, 85, 87–89, 92, 94, 96, 98–100, 107

DNA Deoxyribonucleic Acid. 8, 9, 33, 55

ESS Explained Sum of Squares. 13, 35

FE Feature extractor. 78

FNN Feed-Forward Neural Network. 24, 27, 84, 85, 88, 90, 93, 95, 97, 100

FTD Frontotemporal dementia. 54

g Grams. 37

gBLUP Genomic Best Linear Unbiased Prediction. 11, 30, 37–40, 42–44, 46, 104

GCTA Genome-Wide Complex Trait Analysis. 11, 37–39, 81

GREML Genome-based Restricted Maximum Likelihood. 37

GRM Genetic/Genomic Relationship Matrix. 11, 17, 37–39, 48, 81, 106, 135

GWAS Genome-Wide Association Study. 9–12, 16–18, 20, 38, 41–43, 45–47, 52–54, 57–59, 61–65, 67, 74, 81, 82, 105, 106

H^2 Broad-Sense Heritability. 5, 34

h^2 Narrow-Sense Heritability. 5, 7, 34, 40, 104

IBD Identity By Descent. 12

kb Kilobase. 37, 38, 58, 81, 82

LASSO Least Absolute Shrinkage and Selection Operator. 19, 21, 30, 37, 44, 58, 59, 64–66

LD Linkage Disequilibrium. 9, 12, 18, 27, 37, 38, 56, 58, 81, 82

LMM Linear Mixed Model. 11, 17

LSVM Linear Support Vector Machine. 19, 63, 66, 67

MAE Mean Absolute Error. 39, 82, 83

ML Machine Learning. 19, 21, 30, 36, 55, 57, 75, 101, 104, 105

MLMA Mixed Linear Model Association. 11, 17, 38, 48, 106

MLR Multiple Linear Regression. 21

MND Motor Neurone Disease. 53

PANN Pivotal Adversarial Neural Network. 79, 82, 83, 85, 89, 91, 94, 96, 98

PC Principal Component. 17, 18, 48, 61–65, 67, 74, 75, 78, 81, 82, 84, 86–88, 91–95, 99, 106

PCA Principal Component Analysis. 16, 17

PRS Polygenic Risk Score. 12, 13, 18, 19, 52, 53, 57–59, 62, 65, 68, 71, 74, 105, 106

QC Quality Control. 37, 57

QTL Quantitative Trait Locus. 11, 47

R^2 Coefficient of Determination. 13, 19, 35, 36, 40–42, 45

RBF Radial Basis Function. 22

RF Random Forest. 23, 63, 67

RNA Ribonucleic Acid. 55

SNP Single Nucleotide Polymorphism. 9–13, 16, 17, 36–43, 45–47, 51, 54, 58–65, 67–70, 81, 82, 105, 108, 135

SSE Sum of Squared Estimate of Errors. 13, 20, 21, 35

SVM Support Vector Machine. 22, 58, 65

TSS Total Sum of Squares. 13, 35

Y Phenotype Value. 9, 78, 79, 83

Z Confounder/Nuisance Variable(s). 78, 79, 83

Chapter 1

General Introduction

1.1 Complex Traits

1.1.1 Background

One of the fundamental questions in genetics has always been how an individual's genotypes relate to their various phenotypes. For Mendelian traits, this pathway is normally well-understood, for example how variants in the starch synthesis gene affect seed shape in peas, or the ways in which missense mutations in the *CFTR* gene lead to abnormal mucosal thickness and cystic fibrosis disorder in humans (Thomas 2014). However, for the vast majority of traits and diseases, these genotype-to-phenotype maps are much more complex and much less understood. They can be affected by thousands of genes, interacting with one another and their environment, largely in a non-deterministic and probabilistic manner. These traits, whose variation is affected by both genetic and non-genetic influences, are known as complex traits and disorders (Falconer and Mackay 1995; Walsh and Lynch 1998).

The field of genomic complex trait analysis is thus concerned with quantifying and understanding the genetic component of any given phenotype of interest. This involves

describing its genetic architecture[1], elucidating the cellular and higher-order pathways through which the phenotype manifests, and ultimately how best to predict the trait value of interest (or risk of disease) given a particular genotype. This understanding not only contributes to furthering basic biological knowledge, but can also be an important tool in improving the outcomes of drug discovery and agricultural breeding programmes. Accurate genomic risk estimates of human disease can help in personalized and preventative medicine, and are now beginning to move into the clinic (Lewis and Green 2021; Torkamani et al. 2018).

In the early twentieth century, there was academic debate about how best to reconcile the results of Mendel's discrete inheritance laws with the observations of quantitative (i.e. continuous) traits that Darwin had shown to also be affected by evolution through natural selection. Ultimately, seminal work by Ronald Fisher, Sewall Wright, and John Haldane synthesized these two apparently conflicting understandings of genetics through statistical means (Fisher 1919; Provine 1971). The "infinitesimal" model that was developed involved discrete genetic particles, each undergoing independent Mendelian segregation, and each with a relatively modest effect on the trait value. The smooth continuous variation measured in the ultimate phenotype is as a result of an underlying normal distribution of these various small effects (see Fig. 1.1). It was shown that there was no inherent conflict between Mendel's inheritance laws and Darwin's theory of evolution through natural selection. In fact, these genetic particles were the exact raw material that natural selection acted upon.

[1]The genetic architecture of a trait describes the effect size and allele frequency distributions of said trait, as well as its polygenicity and patterns of interactions with environmental and other genetic factors (Mackay 2001).

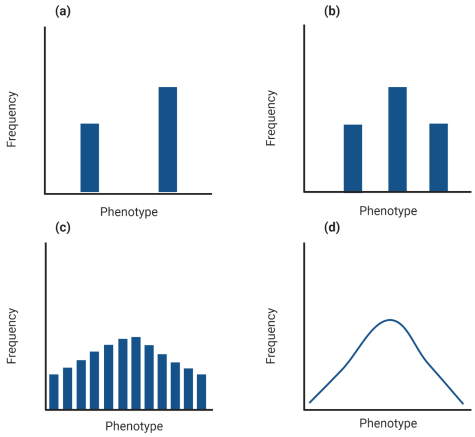


Fig. 1.1 **Infinitesimal Model**[2]**:** Figure shows how discrete units of inheritance can combine and result in a continuous distribution of phenotype values. Phenotype values and their frequency are shown at: **(a)** A diallelic locus. **(b)** Two diallelic loci. **(c)** Six diallelic loci. **(d)** A great number of diallelic loci, resulting in a smooth continuous distribution of phenotype values. *Recreated and adapted from Ridley (2004).*

By and large, this is still how the additive genetic components behind complex traits are modelled today, with variation per trait in the number of loci involved, their independence (i.e. through linkage), effect sizes, and overall distribution shape.

1.1.2 Non-additive Effects

The most basic infinitesimal model assumed for simplicity that there were no interactions between genetic factors and that each summed additively. However, non-additive effects were known to exist from experiment, and early work in statistical genetics was very much concerned with the consequences such effects could have on the variation observed in a trait (Walsh and Lynch 1998). For example, under additivity at a single diallelic locus, the expected value of the heterozygote is half the average difference in phenotype between the

[2]Created using BioRender.com

two homozygous states in the population. There may, however, be a dominance factor at play that shifts this heterozygote value above or below the mean difference.

In addition to dominance effects, the observation of non-additive interactions at different chromosomal locations was given the term *epistasis*. It should be noted that there is some variation regarding the definition of epistasis and its meanings are not always interchangeable[3] (Cordell 2002). Broadly, epistasis can be understood as any interaction between two or more genes, however, this thesis is largely concerned with *statistical epistasis* which is defined as any deviation from a purely additive model of effects between genes.

1.1.3 Heritability

In order to understand the variation underlying a phenotypic trait (V_P), it is possible to decompose the overall variation into components reflecting the contributions of genetic (V_G) and environmental effects (V_E) as can be seen in Equation 1.1 (Falconer and Mackay 1995; Walsh and Lynch 1998). The environmental effects include residual effects that are stochastic in origin, as well as maternal effects, among all other environmental exposures.[4]

$$V_P = V_G + V_E \tag{1.1}$$

[3]The term as coined by William Bateson describes the phenomenon of an allele masking the effect of another at a separate locus (Bateson 1909). This was observed when conducting breeding experiments with both plants and animals. This specific *Batesonian epistasis* was an important step in understanding the inheritance of quantitative traits. The definition of *biological epistasis* can be described as any "situation in which the qualitative nature of the mechanism of action of a factor is affected by the presence or absence of the other" i.e. protein-protein interactions in the cell (Cordell 2002; Siemiatycki and Thomas 1981). This is important when trying to understand the genotype to phenotype maps that underlie a trait of interest, as well as in targeted drug development.

[4]This framework can be expanded to include *genotype* $\times$ *environment* interactions but this has been omitted for clarity. The exact complications that arise from such considerations are largely beyond the scope of this thesis but some are addressed in Chapters 4 and 5.

An important population parameter one can estimate from variance decomposition is the heritability, which may be defined in both a *broad* and *narrow* sense. The broad-sense heritability (H^2) of a trait is defined as the ratio between the genetic variation in a trait and the total phenotypic variation (V_G/V_P). This number represents the absolute upper bound for explained variance using prediction methods that utilize genetic information as their sole discriminator.

It is possible to further decompose the broad-sense heritability into additive and non-additive components (see Equation 1.2). The ratio between the strictly additive genetic component of trait variation (V_A) and the phenotypic variance (V_A/V_P) is termed the narrow-sense heritability (h^2) and is useful in estimating response to selection efforts and the resemblance between relatives. The breeding value of an individual, important in agricultural selection, can be estimated by multiplying a trait's h^2 by the difference between the individual's phenotype and the population mean. Additional pedigree and genomic information can also be used in the calculation of a breeding value.

The non-additive variances include dominance (V_D) and epistatic variance (V_I) which are discussed further in Section 1.1.4.

$$V_G = V_A + V_D + V_I \tag{1.2}$$

The quantitative genetics framework developed by Fisher has been expanded to also account for binary/dichotomous traits, such as some important human diseases, with the development of the liability-threshold model (Falconer 1965; Wright 1934). An underlying normally distributed liability is presumed to exist within a population but only those individuals who cross a certain value threshold manifest in having the disease (see Fig. 1.2). Thus, many tools that are appropriate for analyzing normally distributed variables can also be used, with some modification, in complex disease genetics.

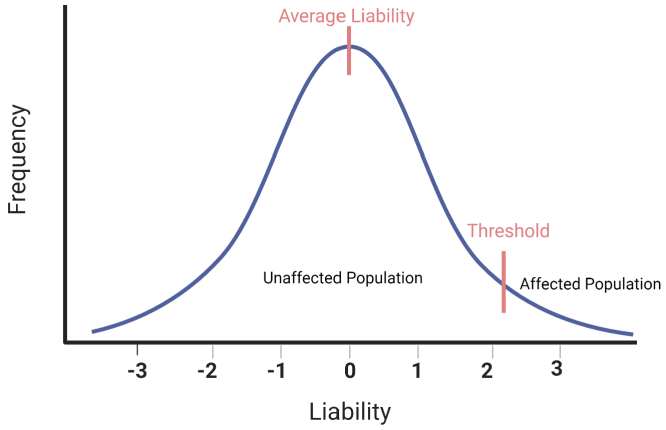


Fig. 1.2 **Liability-Threshold Model**[2]**:** When liability exceeds a certain risk threshold; the binary trait manifests. Below the threshold, all members of the population remain unaffected. An individual's total liability can be modelled as a function of both genetic and non-genetic factors.

As this underlying risk distribution is again a combination of genetic, environmental and stochastic factors, the variation may be decomposed into its various components and the heritability estimated.

Traditionally, heritabilities have been estimated from detailed pedigrees and, particularly in humans, twin studies. Under a few assumptions, including an equal shared environment, the difference in phenotype correlation between monozygotic and dizygotic twins can be attributed to the higher fraction of genomic sharing among monozygotic twins, and thus used to estimate the overall heritability (Verweij et al. 2012)[5]. The heritabilities for many human traits (behavioral, physical, and biochemical) and diseases have been estimated over the decades. Although high heritabilities are frequently observed for both rare and common human disorders, the amount of variance explained by established effect alleles for most traits thus far remains significantly lower than that of the aforementioned theoretical maximum.

[5]For a detailed consideration of heritability estimation in plants, see Section 2.1.2

This is known as the 'missing heritability' problem (Manolio et al. 2009; Torkamani et al. 2018).

It is important to recognize that the estimated heritability value is population-specific and not fixed in time or space, as it is a ratio of two variances. Both the numerator and denominator can change and affect its measured value. For example, the heritability of human height may vary depending on the country in which the measurements took place; there may be varying nutrient availability during childhood or differences in the allele frequencies of causal variants that increase height (Visscher et al. 2008).

1.1.4 Variance Component Analysis

It must be noted that the additive variance captured in h^2 does not only refer to the effects of additive gene action at the molecular level. Importantly, non-additive biological effects can be partly subsumed into the additive variance (Falconer and Mackay 1995). This is true of both dominance and epistatic effects, and is dependent on allele frequencies in the studied population. The implications of this are that a purely additive genomic prediction model is able to exploit some non-additive interactions in its prediction.

Nevertheless, dominance variance is often found to be a large fraction of the total genetic variance in various organisms; an average V_D figure of 38% was found from a thorough literature review (Charmantier et al. 2014). Epistatic variance is particularly challenging to estimate in non-clonal populations and an agreed estimate is lacking (Hemani et al. 2013). The importance of non-additive genetic variance is particularly debated in humans, and it is common to solely focus on additive variance and additive prediction methods (Hill et al. 2008; Mäki-Tanila and Hill 2014; Moore 2003; Zhu et al. 2015).

An underappreciated point when discussing the relative importance of genetic variances is that the standard parameterization into additive, dominance and epistasis variances is not the only framework by which to decompose a phenotype (Dai et al. 2020; Huang and

Mackay 2016). The classical partition (see Equation 1.2) assumes a particular genetic architecture by first maximizing additive genetic variance before calculating dominance and then epistatic variance (Carey 2003). One can get a misleading picture of the actual genetic architecture of a trait with such an approach. Alternate partitions whereby epistatic variances are first maximized might be more informative of underlying gene action depending on the actual genotype-phenotype map. This point has important implications for how an optimal prediction model might be constructed for an individual's phenotype[6]. It has been shown that is not necessarily the case that the classical partitioning of variance into additive and then non-additive effects is always the ideal framework for approaching a given genomic prediction task (Morgante et al. 2018; Ober et al. 2015). The optimal prediction method depends on the actual genetic architecture of the trait, even when the magnitude of the non-additive variance is significantly less than the additive variance. It therefore cannot be assumed *a priori* that linear models will be the most favourable prediction approach for any given trait of interest.

1.2 Modern Quantitative Genetics

The field of quantitative genetics has grown immensely in the more than one hundred years since its inception, aided by a greater understanding of the molecular basis of inheritance, DNA-sequencing technologies and novel statistical methods for genomic mapping, heritability estimation, variant association and phenotype prediction.

[6]This is in contrast with response to selection and designing generational animal breeding programmes, where the traditional additive variance (V_A) is extremely important. Individual-level prediction within a generation (the focus of this thesis) is a separate task where considering the non-additive variance may play a more important role (see Huang and Mackay (2016)). The utility of the classical partitioning in the genomic selection of plants has also been questioned (see Section 2.1.2).

The modern development of techniques and tools in statistical genomics has allowed for an ever more clear picture of the genetic architecture of various traits and diseases, although much more remains to be uncovered.

1.2.1 Genetic Data

Although whole-genome and whole-exome sequencing data are becoming more and more amenable to the types of large-scale analyses conducted across populations in complex trait genetics, this thesis is concerned with single nucleotide polymorphism (SNP) data. When considering variation across a population, it is useful and cost-effective to ignore loci with no variation (99% of the human genome), and so SNP-chips make use of DNA micro-arrays to sequence specific base-pairs with known variability (LaFramboise 2009). The number of SNPs captured by such an array can order in the millions. SNP data can even exploit large-scale linkage disequilibrium (LD) across the genome and capture large genomic segments that occur in linkage blocks by imputing the missing information probabilistically (Li et al. 2009).

1.2.2 Association Techniques

The most widely used method for measuring a SNP's effect on a given phenotype is the Genome-Wide Association Study (GWAS) (Altshuler et al. 2008; Visscher et al. 2017). In its most basic form, the phenotypes (Y) of each N individuals are regressed against the population's minor allele counts (X_i) for each diallelic SNP. Generally, independence between SNPs and additivity among the allele counts[7] is assumed. From a linear regression, effect sizes (β_1) and error estimates can be calculated (see Equation 1.3) and the SNPs then ranked

[7]The allele counts will be 0,1 or 2 for diploid organisms.

by their p-value [8].

$$Y_i = \beta_0 + \beta_1 X_i \tag{1.3}$$

For binary traits, logistic regression can be implemented and the odds ratios calculated for each SNP. The Manhattan plot is a useful way to visualize the results of a GWAS, with the skyscraper-like peaks representing genomic regions of high association with the phenotype (see Fig. 1.3).

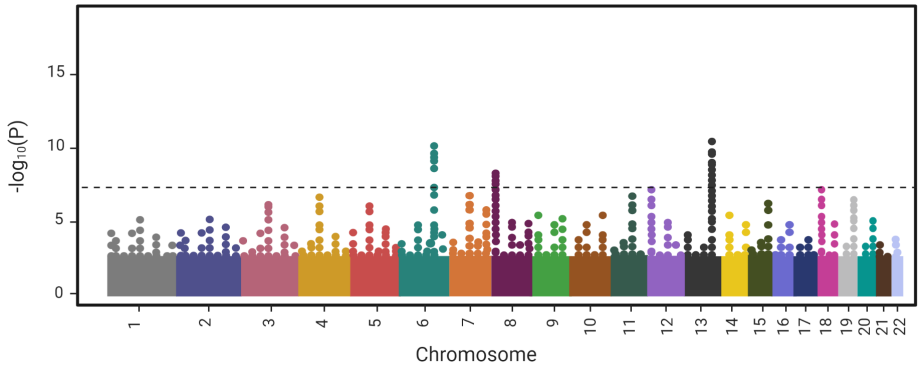


Fig. 1.3 **Manhattan Plot Example[2]:** The dashed line represents a significance threshold of 10^{-8}. Significant peaks on chromosomes 6, 8 and 13 are clearly visible.

As millions of SNPs are tested independently in a single GWAS, the problem of multiple testing arises (Johnson et al. 2010). One simple and common method to deal with this issue is to use the conservative Bonferroni correction on p-values in order to label an association as significant. Permutation-based and other methods have also been used to calculate more appropriate significance thresholds. There remains debate as to the best way to calculate such a threshold (Chen et al. 2021).

One can extend the simple linear GWAS protocol to detect epistatic interactions between SNPs or to model dominance/recessive allele effects. An exhaustive search of all pairwise

[8] β_0 is simply the regression constant and the y-axis intercept. In a GWAS, it is the mean phenotypic value of those individuals homozygous for the major allele.

interactions between millions of SNPs is very computationally intensive and exacerbates the multiple testing problem inherent to GWAS (Wei et al. 2014). To reduce the search space, a hypothesis-driven approach can be taken where only genic SNPs or those in regions known to be related to the phenotype are examined. When moving beyond pairwise comparisons to higher-order comparisons, the problems associated with detecting epistasis compound.

One way to mitigate the problem of multiple testing is to fit all SNPs simultaneously using a linear mixed model (LMM) (Hayes 2013). This procedure can be referred to as Mixed Linear Model Association (MLMA). This approach is powerful in the common situation when there are many more markers than individuals in the reference population. These models also have the advantage of automatically controlling for the unwanted effects of population structure and cryptic relatedness, a problem explored in detail in Section 1.3 (Yang et al. 2014). Genomic analysis software now commonly implement LMM methods for association in humans, such as PLINK and GCTA (Purcell et al. 2007; Yang et al. 2011). One of the most widely used mixed models approaches in non-human genetics is the genomic best linear unbiased prediction (gBLUP). The gBLUP method makes use of the LMM, where SNPs are modelled as fixed effects, to evaluate all possible QTLs in the genome, with the assumption that effect sizes are small and normally distributed.

gBLUP was introduced as an extension of the traditional best linear unbiased prediction (BLUP) through the incorporation of a genomic relationship matrix (GRM) rather than the previous reliance on pedigree information in accounting for covariance between relatives in the population (Habier et al. 2007; VanRaden 2008). The GRM consists of the realized proportion of shared genome for each pair of individuals within the matrix. gBLUP is generally more accurate than BLUP as it relies on the actual relatedness between individuals (even distantly so) rather than their assumed relationship. It is a standard method of choice in animal and plant association/prediction tasks, while also being applicable to human data (Yang et al. 2010).

1.2.3 Genomic Heritability Estimation

Mixed models have also been implemented on population-scale SNP datasets to estimate heritabilities (Evans et al. 2018; Yang et al. 2010). Although these models can often result in lower heritability estimations than as with twin studies, by and large, the substantial role of genetics in the variation of many complex traits has been upheld (Visscher et al. 2008). Importantly, they have the ability to decompose the results into additive (allowing for the calculation of the narrow-sense heritability) and non-additive genetic variances (Zhu et al. 2015). Other common methods to calculate heritability using genomic data include LD-score regression and examining the extent of identity-by-descent (IBD) sharing between individuals instead of those inferred by pedigree (Bulik-Sullivan et al. 2015; Visscher et al. 2006).

1.2.4 Genomic Prediction

With the summary statistics from a GWAS, one can create a score from an individual's SNP profile in order to predict their phenotype. In humans, this is generally done by using a polygenic risk score (PRS) as seen in Equation 1.4 (Choi et al. 2020). For an individual, this is simply a summation of all M included SNP effects (β), each multiplied first by their corresponding genotype dosage[9]. The main parameter of interest that varies when constructing the score is the p-value threshold for inclusion of SNPs.

$$PRS = \sum_{j=1}^{M} \beta_j \times Dosage_j \tag{1.4}$$

Generally, the thresholding is preceded by a pruning step whereby only a subset of SNPs free from high LD with one another are chosen in order to include only independent effects.

[9]Dosages are 0, 1 or 2 for humans and other diploid organisms.

Another method for choosing which SNPs to include is termed clumping. This involves choosing the top SNP in each linkage block, and is generally preferable to pruning as it takes *p*-value into account before thresholding.

Metrics

When evaluating the success of a PRS for a quantitative trait, generally the coefficient of determination (R^2) is the metric of choice. The R^2 measures the amount of variance in the phenotype that is explainable by the score. The upper-bound for this score would be the narrow-sense heritability although in practice it is often much lower (Manolio et al. 2009). The calculation of R^2 can be seen in Equation 1.6. The sum of squared estimate of errors (SSE) and the total sum of squares (TSS), defined in Equation 1.5, can be understood as the unexplained variation and total variation respectively. Equivalently, it is the explained sum of squares (ESS) as a fraction of the TSS.

$$TSS = ESS + SSE \tag{1.5}$$

$$R^2 = 1 - \frac{SSE}{TSS} = \frac{ESS}{TSS} \tag{1.6}$$

For binary traits, the concept of variance explained is not simply defined and so pseudo-R^2 metrics are used, such as Nagelkerke's R^2. It is advisable to adjust this metric for prevalence in the case of a disorder that is less common in the overall population than the collected dataset (as is common in order to increase power in the analysis). A few variants of pseudo-R^2 measures have been developed with some debate as to the most appropriate (Choi et al. 2020; Lee et al. 2012).

Another useful metric in disease genetics is the area under the curve (AUC)[10]. The curve is drawn by plotting the true positive rate in classification against the false positive rate (see Fig. 1.4).

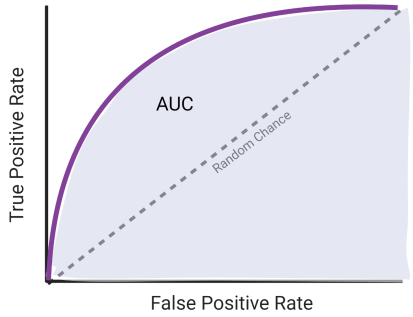


Fig. 1.4 **AUC Plot for Binary Classification[2]:** The dashed line represents an AUC of 0.5 which is equivalent to random chance classification between a single case and single control. The area under the unbroken line represents an AUC of ~ 0.8 which is a much better discriminator.

The area under such a curve describes the probability of correctly classifying a randomly selected pair of one case and one control. To be clinically useful, a minimum AUC value of 0.75 is recommended (Janssens et al. 2007). The AUC can also be affected by disease prevalence in the sample, and so adjustment may be necessary (see Section 3.1.1) (Cook 2007).

An important but commonly overlooked performance metric when analyzing binary outcomes such as disease is a model's calibration (Cook 2007; Van Calster et al. 2019). When consulting the AUC, which is a single number, one does not get information on individual risk estimates in the sample. These risk estimates may be systematically under- or over-inflated leading to possible mistreatment of patients. Therefore, it is important to measure the calibration of a model which is defined as the "agreement between the estimated

[10]Technically, this is specifically referring to the area under the receiver operating characteristic (AUROC), but the term AUC as shorthand is most commonly seen in the genetics literature.

and observed number of events". Ideally, one should observe disease manifestation in 75% of the cohort of patients who are predicted to be at 75% risk of disease over a given interval. Calibration, therefore, can be observed in a reliability diagram which plots the observed proportion of cases against the predicted risk in a sample (see Fig. 1.5). A model with a lower AUC but better calibration than another model could be more clinically useful, and so both metrics are important in binary classification tasks (Steyerberg 2019). Poor calibration can be rectified through methods such as Platt scaling which transform the output into improved probability estimates (Platt et al. 1999). The procedure is fairly straightforward and simply involves fitting a logistic regression model on the original output to result in more well-calibrated prediction probabilities.

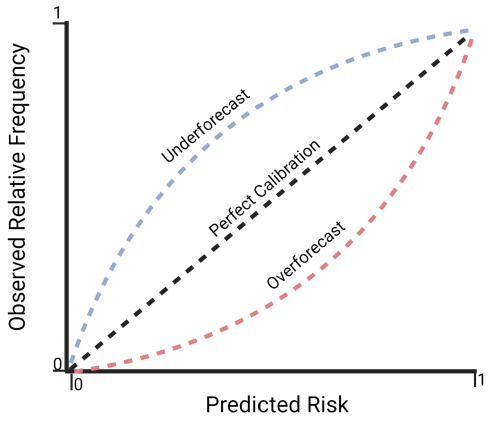


Fig. 1.5 **Calibration Reliability Diagram**[2]: The straight dashed line represents perfect calibration between the predicted risk and the observed number of cases at that risk level. The curve above this line is systematically under-forecasting the risk level for individuals, while the curve below is over-forecasting. The calibration curve of a predictive model may also be S-shaped whereby it under- and over-forecasts at different levels.

1.3 Population Structure and Confounding

The associations detected between genetic variants and phenotypes may not arise solely out of direct biological effects, but may instead be confounded by non-genetic effects correlated with ancestry differences within the population (Barton et al. 2019). The development of

statistical methods to detect and adjust for the confounding effects of population structure has been one of the main priorities of the complex trait field in the GWAS era.

A canonical example of this phenomenon is a hypothetical GWAS conducted on chopstick use in a mixed population (Barton et al. 2019; Lander and Schork 1994). Although there may exist actual variants (related to a trait such as dexterity) that could affect chopstick use, the most significant SNPs identified in such a GWAS would most likely be those that reflect East Asian ancestries. Such spurious associations reflect cultural differences in the use of chopsticks, and yield no genuine insights into the actual trait biology. This principle holds true across virtually all complex traits and diseases in human populations, as we cannot ignore the myriad ways in which culture, society, and the environment play a role across ancestries. This could apply even on relatively small geographical scales such as sub-regions within countries.

Even in a hypothetical setting in which environmental differences were negligible, confounding by shared genetics (i.e. cryptic relatedness) within a population would also exist (Vilhjálmsson and Nordborg 2013). The correlation between causal loci and the shared genetic background of distantly related individuals in the sample population would lead to many false positive associations with ancestry-related SNPs. For these reasons, one must be careful when associating genetic variables to phenotypes to account for any confounding variables.

1.3.1 Principal Component Analysis

In order to account for confounding, it is common to include covariates during a GWAS regression that account for factors (sometimes termed nuisance variables) such as sex, age and population structure in the sample. The most popular approach taken by geneticists with regard to generating covariates that tag population structure is called principal component analysis (PCA) (Price et al. 2006). The procedure involves a dimensionality reduction step

that calculates a set of vectors called principal components (PCs) from a higher-dimensional array. These PCs are uncorrelated with one another whilst explaining the maximum possible amount of variation in the data. The vectors sequentially explain less and less variation meaning that a relatively small number of PCs can capture much of the variation in a very high-dimensional space.

It has been found that the first PCs of genomic SNP-matrices tend to describe broad ancestry patterns in the population, and so can be included as covariates in a GWAS in order to offset confounding effects of population structure (Novembre et al. 2008). There remains some debate about the effectiveness of this approach, as well as how best to determine the exact number of PCs to include[11], but PCA remains standard practice for genetic association and prediction tasks (Abegaz et al. 2019).

1.3.2 Mixed Model Approaches

As mentioned in section 1.2.2, the use of a LMM can also account for confounding in a sample. Through the use of a GRM as a random effect, that captures the realized proportion of shared genome for each pair of individuals within the matrix, the model can account for population structure and cryptic relatedness during SNP association. These two techniques are not mutually exclusive and one can include PCs as covariates during an MLMA.

1.3.3 Genomic Prediction

Population structure must be corrected for not only at the association stage but also during any prediction using the most significant SNPs (Choi et al. 2020). In human polygenic risk scoring, if the target population for prediction is ancestrally similar to the base population

[11] For more details on the considerations of a PCA procedure see Section 4.1.1.

from which the GWAS was performed, correction with PC covariates is generally seen as sufficient. However, if the populations are too genetically dissimilar, wide variation in prediction accuracy can occur. This arises due to confounding, but is also due to large-scale differences in LD structure across human populations. This reality creates a problem for the potential utility of PRS, as the vast majority of GWAS have been conducted on those with predominantly European ancestries and may not be applicable to those of differing backgrounds (Martin et al. 2017). This disparity in representation has been slightly improving over the last decade and the task of creating diverse genomic risk assessment tools is becoming more widely undertaken (Lewis and Green 2021; Márquez-Luna et al. 2017; Mills and Rahal 2020).

Although it is largely the aim of genetic association techniques to detect purely causal variants, when dealing with medical risk prediction there is some debate as to whether or not it is strictly necessary to ensure the predictors are entirely free from confounding (Abdellaoui and Verweij 2021; Barton et al. 2019; Kaplan and Fullerton 2022; Wray et al. 2013). Population structure (acting largely as a proxy for non-genetic effects), despite not being causal, may nonetheless be important in identifying those at most risk of a certain disease and those who may best be put forward for early-screening. Although an individual's score may be derived from a combination of genetic and non-genetic sources, the risk could remain very much present. There is no universal agreement among academics and clinicians regarding this point, but it is worth noting here.

1.3.4 Validation

The gold-standard approach in humans to validate that a genomic risk score is not confounded is through the use of a within-family (sibling) study (Choi et al. 2020; Lello et al. 2020). Environmental variation, and the effects both caused and tagged by population structure (see Section 1.3), are limited in this design, and so any variance explained by the score should be

as the result of a true causal effect. Typically, the R^2 revealed in these designs is significantly lower than those seen in population-wide data. The availability of such datasets is quite limited, although they are growing in number as their importance in PRS validation has become clear.

1.4 Machine Learning

Machine learning (ML) refers to the use of a set of algorithms that are trained to detect and exploit complex patterns in data that can then be used to make predictions on previously unseen test datasets (Friedman et al. 2009; Géron 2019; James et al. 2021). For genomic prediction, the training dataset generally consists of individual-level genotype inputs and phenotype value outputs. As the training individuals' trait values are labelled, this is referred to as a supervised prediction problem.

Although more simple models may provide causal insights into how input variables relate to the output, the appeal of some machine learning families lies in their greater flexibility to model complexity, such as that which may arise from the genetic architecture of a trait. The exact genotype-phenotype map may remain obscured, however.

It is important to note that not all machine learning methods are non-linear, such as regularised regression models (Ridge and LASSO) or linear support vector machines (LSVMs) that are discussed below. However, these approaches may, in some instances, still be more suitable for the genetic architecture underlying certain traits (e.g. via effective regularisation and feature selection) and the resulting models may provide a more appropriate framework to capture some non-additive biological action as explained in Section 1.1.4.

When training a machine learning model, many hyperparameters of the chosen model type may need to be optimized, depending on the specific machine learning method in question. This can be achieved by searching through various combinations of these hyperparameters, with time and computational power usually being limiting factors. As a result of these

complex optimization techniques, the final models are often not readily interpretable because they can involve complicated high-dimensional non-linearities.

When training a machine learning algorithm there is a risk that the model learns certain idiosyncrasies and patterns in the training data that do not apply to the outside test data. This problem is known as overfitting. For this reason, a penalty feature is often included in machine learning algorithms that penalize complexity (variance) in the final model. This penalization is referred to as regularization.

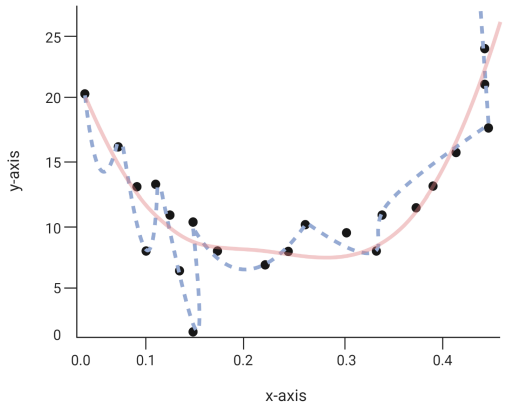


Fig. 1.6 **Overfitting Example**[2]**:** The blue dotted line represents a model that is overfitted to the data, as it tries to closely match every point. It will probably not generalize well to new data. The red unbroken line will probably be a better predictor, as it has less variance.

A brief explanation of the concepts behind the various machine learning methods used in this thesis is given below:

1.4.1 Modified Regression - Shrinkage Methods

To understand some basic machine learning algorithms, we can extend the linear regression model used in GWAS (see Section 1.2.2) except with multiple variables being fit simultaneously (see Equation 1.7). Similar to Equation 1.3, one is interested in finding the optimum parameter values (β) for the regression by minimizing the loss ($\mathscr{L}$) function through the ordinary least squares approach. The loss in this multivariate problem is the SSE seen in Equation 1.6

and which is also the basis for other ML techniques. This multiple linear regression (MLR) itself is considered to be a machine learning algorithm although it is strictly linear in its construction and there are no hyperparameters to be tuned (Bishop 2006; Friedman et al. 2009).

.

$$\mathscr{L}_{MLR} = SSE = \sum_{i=1}^{N}(Y_i - \beta_0 - \sum_{j=1}^{M}\beta_j x_{ij})^2 \tag{1.7}$$

Extensions to linear regression that put a penalty on the number or the magnitude of coefficients are Ridge regression and Least Absolute Shrinkage and Selection Operator (LASSO) regression respectively (Friedman et al. 2009). By shrinking some coefficients towards zero, a ridge regression model tends to become more sparse and has the possible advantage of reducing overfitting and model complexity (see Equation 1.8). A tuning hyperparameter lambda (λ) needs to be learned that controls the strength of regularization appropriate for the data at hand. This technique is also known as L2-regularization regression.

$$\mathscr{L}_{Ridge} = SSE + \lambda \sum_{j=1}^{M}\beta_j^2 \tag{1.8}$$

LASSO regression is similar except it takes the absolute value of the coefficient into account rather than the square (see Equation 1.9). This allows for the shrinkage of some coefficients to absolute zero (essentially performing feature selection and removing them from the model). This technique is also known as L1-regularization regression.

$$\mathscr{L}_{LASSO} = SSE + \lambda \sum_{j=1}^{M}|\beta_j| \tag{1.9}$$

Both these machine learning methods have been used extensively for genomic prediction with success (Ogutu et al. 2012).

1.4.2 Support Vector Machines

Support Vector Machines (SVMs) are another popular machine learning method (Géron 2019; James et al. 2021). In the simplest example, and considering a two-dimensional classification problem, a straight-line decision boundary[12] can be drawn the maximizes the distance between two classes (see Fig. 1.7). The distance between the two hyperplanes on either side of the decision boundary, on which the closest points of each class lie, defines the maximal margin of the SVM. The marginal cases themselves are called the support vectors. A hyperparameter can be tuned on how strict the model should be in penalizing any instances of margin violations past the hyperplanes. For classes that are not easily linearly separable, even after allowing for some margin violations, a kernel transformation can be applied that projects the data into a higher-dimensional feature space. This projected data may be more easily separable by a decision boundary that maximizes the distance between classes (see Fig. 1.8). This transformation is performed at a relatively low computational cost through what is known as the "kernel trick". When the decision boundary in the high-dimensional space is back-transformed, a non-linear curve is obtained that can separate the classes. Both a polynomial and Radial Basis Function (RBF) kernel are investigated here.

SVMs for regression are similar in their construction except that, unlike in classification, the goal is to fit as many instances within the margin as possible. The resultant decision boundary then gives the line of best fit from which predictions are made. As before, the kernel trick can be applied and a non-linear decision boundary formed.

[12]This decision boundary is also a hyperplane, and may be referred to as such in some sources, but should not be confused with the hyperplanes on which the support vectors lie. For this reason, it will simply be referred to as the decision boundary here.

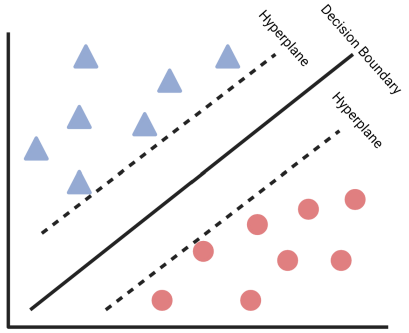


Fig. 1.7 **Support Vector Machine**[2]**:** The decision boundary maximizes the margin between the two hyperplanes on which the closest points (support vectors) of each class instance lie.

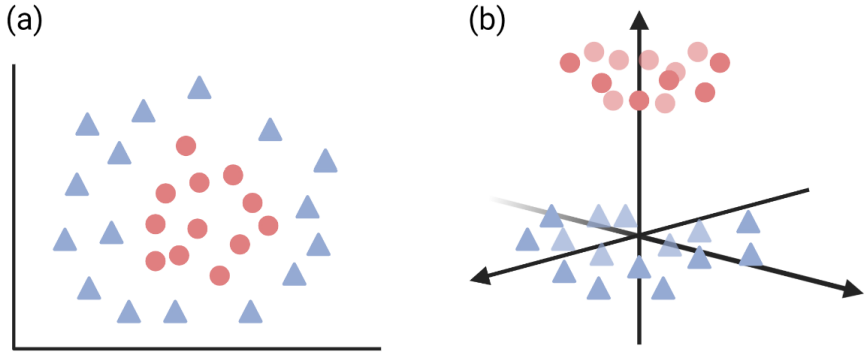


Fig. 1.8 **Kernel Trick Example**[2]**: (a)** The two classes are not linearly separable in this case. **(b)** The data from (a) has been projected to a higher-dimensional space through a kernel transformation. A decision boundary in the form of a plane can now be drawn to separate the classes. When this boundary is back-transformed to (a) it will create a non-linear (circular in this case) boundary between the classes.

1.4.3 Random Forests

The random forest (RF) is another popular learning algorithm (Ho 1995). The basis of a random forest is the decision tree. These simple trees learn hierarchical rules by which to recursively split and classify data (see Fig. 1.9). Each tree is comprised of internal decision nodes and terminates with predictive leaves. In order to improve their performance and generalizability, random forests were designed through an ensemble method known as

bootstrap aggregation (Bagging). Bagging involves taking random subsets of the data, with replacement, to construct multiple decision trees. The most common classification of the multiple trees is then returned as the final prediction. This Bagging procedure, using subsets of the inputs and features, aids in reducing the overall variance of the final prediction.

Random forests can also be used for regression tasks except the instances in the final node of the bottom tree are averaged, and this value is returned as the final prediction. Hyperparameters that must be trained in random forests include the number of trees and their depth.

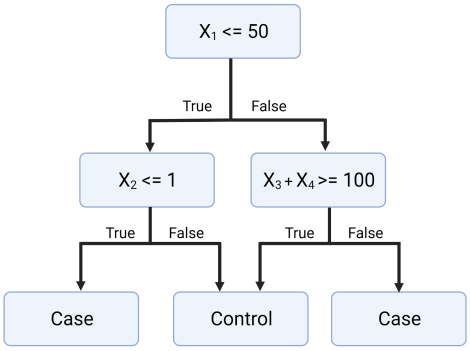


Fig. 1.9 **Decision Tree[2]:** Starting from the top, the data is split sequentially by specified rules at each node. The terminal leaf nodes then make a classification of the data. Decision trees are the basis of random forests.

1.4.4 Neural Networks

The basic structure of a standard feed-forward neural network (FNN) consists of layers of nodes connected by edges (see Fig. 1.11) (Géron 2019; James et al. 2021). The first layer of nodes consists solely of the input data samples. A single node connected by all its edges is known as a perceptron and is the basis of the network (see Fig. 1.10). These nodes are connected to the second layer of nodes through edges containing weights that apply a

non-linear activation function to the sum of the weighted inputs (plus a bias term[13]) from the previous layer. This process is repeated for all nodes across all layers of the network. The final output layer uses a specialized activation function to generate a single prediction. This output can be changed depending on the task being a classification or regression one.

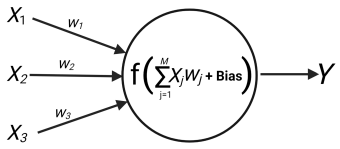


Fig. 1.10 **Perceptron[2]:** The inputs are fed into a single node where they are summed after being multiplied by specific weights. A bias term is then added. A chosen function is applied to the resultant number, which is then outputted from the node. In a neural network, this output can form part of the input to another layer of nodes. It may also be the final classifying function at the end of the network i.e. a binary output can be obtained through the use of the sigmoid function.

[13]The bias is simply a constant vector that is added to the sum of the inputs multiplied by their weights. It may prevent the activation function from reaching a threshold value and thus sending a significant output to the next layer. The bias can therefore reduce variance in the model and prevent overfitting. The bias term may also have weights attached that are updated during back-propagation. For clarity, these bias weights are left out of figure 1.10.

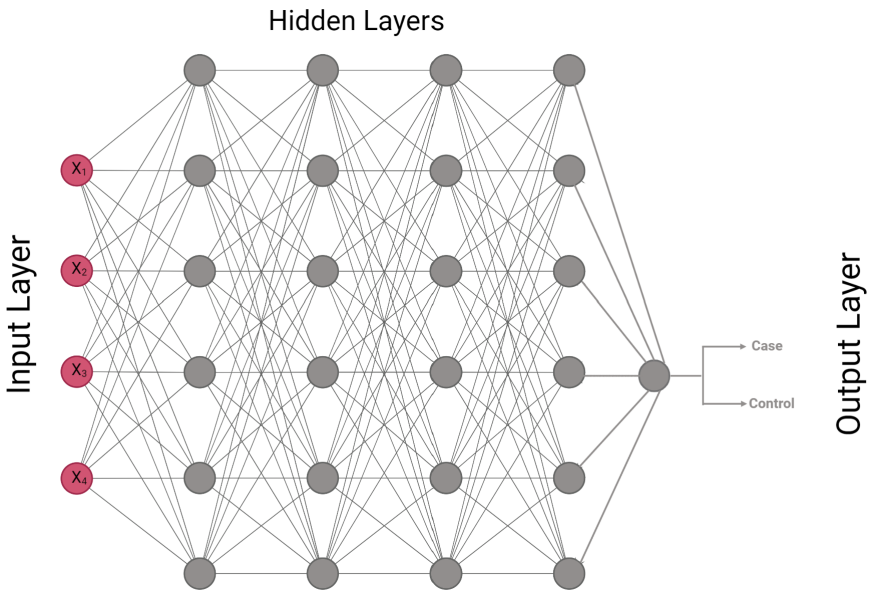


Fig. 1.11 **Neural Network Example**[2]: Each input variable is connected to all nodes in the first hidden layer which take the form of Fig. 1.10. The output of each node is fully-connected to the next hidden layer and the process repeated. The output layer receives input from the final hidden layer and makes a prediction. In this example, the sigmoid function makes a binary classification. The loss between the prediction and training labels is then back-propagated through the network in order to adjust its weights.

A method termed back-propagation feeds the loss ($\mathscr{L}$) from the predicted output back through the layers of the network and adjusts the initial weight values in order to minimize the loss over several iterations (each iteration is known as an epoch). Many loss metrics may be suitable for the problem at hand. To ensure the loss minimization and back-propagation process is efficient, a range of different optimizer methods are available. As neural networks tend to create complex models, regularization of the weight coefficients is important. Another important regularization method termed "Dropout" randomly disconnects a certain fraction of edges from their nodes in a layer. This can help greatly in avoiding overfitting to the training data.

Neural networks found early success in image classification and it was found that as the input data was manipulated through the layers, complex high-level features were extracted and were used for the final prediction. In general, the use of neural networks with more than

three layers can be referred to as “deep learning”. Even more complex network architectures exist such as those of a convolutional neural network (CNN) which is especially suited to spatial input. Given the LD structure of the genome it is possible that patterns exist within the linear, spatial sequence that can be exploited by CNNs.

CNNs can be understood as a variant of the FNN except that a convolution step is undertaken first, followed by pooling which is essentially a smoothing procedure. The convolution step differs from the traditional FNN in that in order to emphasise more restricted connectivity in the data, a local feature extraction is performed. In this case, the network may not be fully-connected at all nodes but instead windows are chosen along the feature space with which to analyse patterns.

1.4.5 Hyperparameter Optimization

As stated previously, machine learning algorithms may have many different hyperparameters that need to be simultaneously optimized during training. Depending on the task in question, there may be hyperparameter value windows with which performance is known to be optimal, however, a machine learning framework generally begins with a rather broad range of possible values to evaluate. If both the number of hyperparameters and the possible values such hyperparameters take is low, then a full-grid search of all possible combinations can be undertaken. However, this is often infeasible due to time-constraints and so random combinations can be drawn and tested. With enough iterations of this procedure a general picture of the optimal value of each hyperparameter may emerge. This approach is known as random grid-searching. A further method for hyperparameter optimization, using a Bayesian approach, is discussed in Section 3.1.4.

1.4.6 Prediction and Evaluation

In order to estimate predictive performance on outside datasets, a hold-out set or a cross-validation strategy is usually applied to machine learning procedures (Friedman et al. 2009). *k*-fold cross-validation is generally the preferred method of choice, and involves splitting the data *k* times, each time resulting in an independent training and validation set of individuals. This way, the model is trained and evaluated *k* times, and gives a more accurate estimate of the prediction capability on independent test data. Overfitting is avoided, as would happen in the naive approach whereby a model is trained and then evaluated on the entire dataset. An additional benefit is that all of the data are utilized for training purposes, unlike the hold-out method, resulting in greater power.

In the case where hyperparameters are also being selected, a nested cross-validation procedure is required in order to avoid information leak (and thus an overly-optimistic result) from the validation sets (Krstajic et al. 2014; Vabalas et al. 2019; Varma and Simon 2006). The nested approach involves further performing an *n*-fold cross-validation on each of the *k* training sets, resulting in inner training and validation sets alongside the original outer sets (see Fig. 1.12 for a stepwise explanation of the procedure). The best hyperparameter combinations are chosen within each inner *n* split and then applied independently on each outer training set before being evaluated on the final *k* outer validation sets.

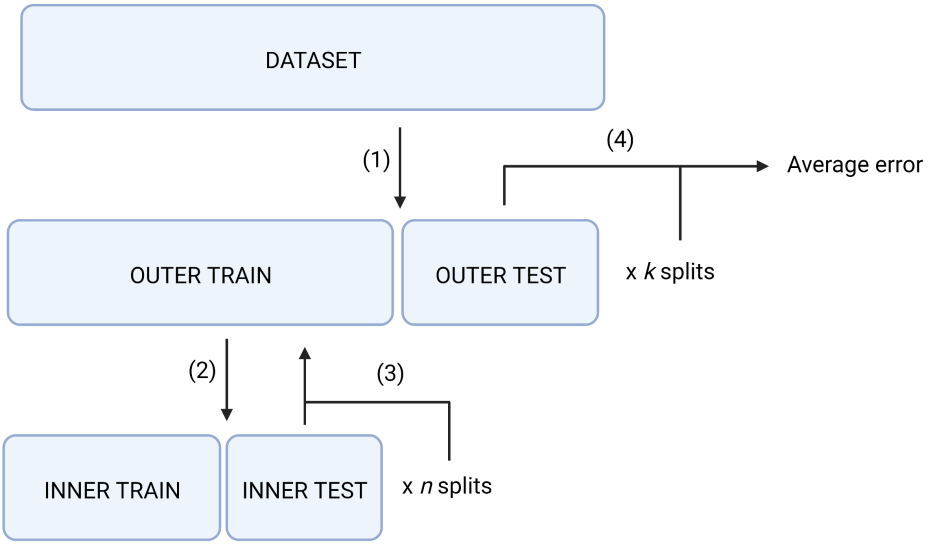


Fig. 1.12 **Nested Cross-Validation Procedure**[2]**: (1)** Split entire dataset k times into outer train and test sets. **(2)** Split each k outer train set into n inner train and test sets. **(3)** By training on each inner train set, find the overall best hyperparameter combinations across the n test sets and then apply this to train the corresponding outer train set. **(4)** Evaluate final k trained models on their corresponding test sets and average the errors.

In the non-nested approach, one would end up choosing the best hyperparameter values from repeated testing on the k validation sets, resulting in an optimistic estimate of the actual predictive performance. When the cross-validation is nested, the hyperparameter values are only configured within the inner n loops and so the error when applied to the outer validation sets should result in a more realistic estimate of the performance on an independent dataset.

This is not commonly implemented in the genomic prediction literature, despite the risks that simple cross-validation entails (Bracher-Smith et al. 2021). Thus, in this thesis, a fully nested cross-validation hyperparameter search was chosen in the absence of available independent datasets. The resulting estimated performance should be relatively free from bias introduced by both hyperparameter tuning and model training.

1.4.7 Feature Selection

An important aspect of any predictive model is the number of predictors included in the feature space (Hua et al. 2009). An efficient feature selection strategy can reduce the computational burden of training and remove unnecessary predictors from the model. Feature selection may be performed either before training, often by first associating the input variables to the output, or during model training, as with methods such as random forests or LASSO regression which remove variables internally. It is extremely important that any feature selection step must only be performed using information from the training set. Otherwise, bias will be introduced from validation sets and result in an over-optimistic performance measurement.

1.4.8 Model Comparison

Cross-validation may be used to select the best hyperparameters for each machine learning algorithm but it is also important to construct a valid comparison *between* different machine learning approaches. There are a few procedures one can implement to accomplish this task, but one that includes a measure of statistical significance between methods would be generally preferable[14]. In this thesis, the ML models are compared with respect to baseline polygenic risk scoring in humans or gBLUP in plants and so the multiple-comparison Dunnett test was chosen as the most appropriate statistical test to compare between models (Dunnett 1955; Japkowicz and Shah 2011; Maxwell et al. 2018). This test allows for an increase in statistical power in comparison to other multiple comparison methods when dealing with the

[14]In my review of previous machine learning comparisons in genomic prediction, many did not explicitly state any statistical test in comparing between models (Bellot et al. 2018; Cope et al. 2021; Gola et al. 2020; Lello et al. 2018; Liu et al. 2019). Only one reported a formal statistical test but did not explicitly adjust for multiple comparisons (Ma et al. 2018).

relatively small sample sizes given by k-fold cross-validation. Otherwise, having to conduct all pairwise comparisons between models would result in a less powerful test, although method ranking can still be achieved using the Dunnett test through p-values (Sheskin 2000).

1.5 General Motivation and Research Outline

This thesis is concerned with exploring the potential of using machine learning methods for the task of genomic prediction. It is clear that non-additivity and epistasis may be key components of many genetic architectures and that linear models may be limited in their ability to exploit this. A substantial amount of heritability for many important traits remains to be explained using existing methods, and it is possible that the ability of machine learning algorithms to model complexity may aid in accounting for more of the variation that we see in these traits. This potential is explored in Chapter 2 in the case of quantitative traits using *Arabidopsis thaliana* and in the case of human disorders in Chapter 3. Both chapters contain specific reviews of prior work on genomic prediction and machine learning using plant and human data respectively. Furthermore, although the confounding effects of population structure have long been widely acknowledged in the human genetics literature, methods to account for this at the nascent intersection of genomic prediction and machine learning remain under-explored. Chapter 4 of this thesis is thus concerned with the development of new techniques to mitigate this profound problem. Finally, a general discussion and conclusion are given in Chapter 5.

Chapter 2

Genomic Prediction in *Arabidopsis Thaliana*

2.1 Introduction

As discussed in Section 1.2.4, genomic prediction refers to the task of estimating an individual's phenotype based on known genotypic information using DNA markers. Genomic selection, a technique important to the agricultural industry, refers to the use of such genomic information to prioritize the selection of mating individuals through the estimation of breeding values.

Genomic prediction and genomic selection have been important tools in agricultural programmes and have seen success in improving overall yields of many important crops. These increased yields from plant food production are imperative in the coming decades as the human population is projected to grow significantly and reach 9 billion by 2050 (Hickey et al. 2017). Krishnappa et al. (2021) give a comprehensive review of the previous work done using genomic selection for the use of improving crop outcomes.

This chapter will discuss the use of machine learning for the possible improvement of genomic prediction, using *Arabidopsis thaliana* as a model organism, and statistically benchmarking against the traditional methods that are employed for this task.

2.1.1 *Arabidopsis thaliana*

The flowering plant *Arabidopsis thaliana* is a useful model organism with which to study genomic prediction methods. It is a diploid species with a short generation time, selfing ability, and a relatively small genome. As part of the *Brassica* genus, which contains mustards, cabbages, rapeseed and broccoli, it is an excellent experimental reference for many important crops (Koornneef and Meinke 2010). A large amount of natural variation exists across a wide geographic range for many important *Arabidopsis* phenotypes. This includes flowering time, which has long been studied as a quintessential quantitative trait (Alonso-Blanco et al. 2016; Laibach 1951). Collections of naturally inbred homozygous *Arabidopsis* samples, called accessions, are widely available for genetic analysis.

2.1.2 Heritability in Plants

As explained in Section 1.1.3, the heritability of a trait is defined as the ratio of the genetic variance to the total phenotypic variation in a population. In plants, many complexities can arise in defining and estimating this parameter and there are a number of factors that may have to be considered such as reproduction method (sexual, asexual and self-fertilization), ploidy, and the existence of large clonal populations (Nyquist 1991; Schmidt et al. 2019). In general, as with human genetics, pedigrees can be used to estimate the h^2 and H^2 heritability values. However, as clonal populations can be grown and experimented upon, each with the same genotype but measured across different environments, the H^2 can be calculated using a phenotypic mean of a collective genotype, which is naturally in contrast to human and animal calculations.

The discussion on the classical partition of genetic variances in Section 1.1.4 is also relevant for genomic selection and prediction in plants. In animal breeding programmes, one is almost always interested in the breeding value of an individual, reflected in its progeny, and which is calculated using the additive component of genetic variation as this is the component that is transferable across generations. However, in crop improvement programmes, one tends to be more interested in the total genotypic value[15] (which includes non-additive effects) of the individual plants, which may then be clonally propagated as a final product (Bernardo 2020). This is not to say that one is not interested in the breeding potential of plants, especially during early selection cycles, but that the ultimate measure of interest will often be the total genotypic value of an individual, not specifically its breeding value, as this will be more reflective of the worth of the final commercial entity (Crossa et al. 2017). For this reason, the potential to exploit non-additive variance must be considered in any well-designed crop breeding programme.

2.1.3 Performance Metrics

As explained in Section 1.2.4, for quantitative traits, the coefficient of determination is a standard goodness-of-fit metric by which to evaluate model performance. However, the interpretation of the R^2 as the amount of variance explained by the model does not hold for non-linear models, and caution is warranted when employing this metric on such models (Kvålseth 1985; Sapra 2014; Spiess and Neumeyer 2010). In order to interpret R^2 as variance explained, it is essential that the SSE and ESS sum to TSS, as seen in Equation 1.5 however, this is only true in a linear regression setting.

[15] Genotypic value is defined as "The expectation of the performance of the candidate in a target population of environments" in Bernardo (2020).

This is unfortunate, as the R^2 lies on the same scale as the heritability and would be useful in benchmarking performance against the maximum prediction theoretically possible through using genomic information. Instead, an alternative metric that is often employed in plant and animal genetics is Pearson's correlation coefficient (ρ) between predictions and measurements (Tong and Nikoloski 2021). It is a good measure of the predictivity of a model and will be employed in this chapter.

2.1.4 Previous Work and Motivation

There has been some interest in the field of plant genetics on the use of machine learning to improve genomic prediction (see Liu and Wang (2017); Ma et al. (2018); Montesinos-López et al. (2018, 2019) and others). Studies on several different species have been published although there remains debate on whether or not ML consistently improves upon traditional methods, and if so, which specific family of learning algorithms works best. Tong and Nikoloski (2021) give a current review of previous work done on this topic. In general, the sample sizes of these studies number in the hundreds and improvements are modest. This chapter will focus on adding to the body of research on the question of machine learning for genomic prediction with a large global sample of *Arabidopsis thaliana* across a number of traits. Strategies for feature selection and SNP-set size will be discussed as well as the challenges in understanding the trained models, statistically comparing between methods, and the potential bias introduced by population structure.

2.2 Materials and Methods

2.2.1 Data

Open-access *Arabidopsis* SNP data was accessed from the 1001 *Arabidopsis* Project across a variety of phenotypes (Alonso-Blanco et al. 2016). This dense dataset contains 10,707,430

SNPs across all four chromosomes. Accessions are from a global sample which are highly inbred, resulting in homozygosity at every position in the genome. "Time to first flowering" was chosen as the main trait of interest as the genetics of the trait have been extensively studied and had the largest sample size of both genotype and phenotype information. This trait was studied across two laboratory temperatures: 10°C and 16°C. The other traits examined are dried seed yield in grams (g) and leaf area in centimeters (cm) squared. All accessions were grown under standard laboratory procedures and conditions, constraining the environmental component of trait heritability. More detailed information on the exact protocols for genotyping, phenotyping and the quality control (QC) of this collection is described in Alonso-Blanco et al. (2016).

2.2.2 Heritability Estimation

GREML analysis was used to estimate the additive SNP-based heritability (h^2_{SNP}) for each trait using GCTA software version 1.92.0 (Evans et al. 2018; Yang et al. 2010, 2011). The GRM used for this analysis was built using a set of 585,375 pruned SNPs. The pruning procedure was conducted in PLINK2 and used a 250kb window, 5bp step-size and a LD ρ^2 threshold of 0.05 (Purcell et al. 2007). A minimum minor allele count of 6 was chosen for SNP inclusion.

2.2.3 Experimental Approach

Using genomic data from the samples of *Arabidopsis*, the objective of this analysis was to compare the relative performance of the various machine learning methods to the baseline linear models. In addition to baseline gBLUP, the performance of LASSO regression, ridge regression, random forests, support vector machines (linear and non-linear), feed-forward neural networks and convolutional neural networks was assessed.

The influence of feature set size and effect of prior linear feature selection were also investigated in this analysis. Prediction accuracies of machine learning methods were compared against gBLUP prediction as a baseline, and for each trait the SNP-based heritability was also calculated to gauge the overall performance of the models.

2.2.4 Feature Selection

Two feature selection strategies were adopted in order to select the input SNP set from the dense genotype dataset, with feature selection performed independently on all inner and outer loops of this analysis. The first strategy was to use a GWAS to identify the SNPs most linearly associated with the trait. The results were subsequently clumped to generate the set of top SNPs free from strong linkage disequilibrium. The GWASs conducted on the various *Arabidopsis* phenotypes were performed using GCTA-MLMA (Yang et al. 2011, 2014). The GRMs were computed using all SNPs and all relevant training individuals in each analysis (Yang et al. 2014). No explicit covariates were included in the MLMA (aside from the population structure controlling effect of the GRM) as per the original Arabidopsis study protocol (Alonso-Blanco et al. 2016).

The resulting SNPs were clumped using the following parameters: significance threshold of 0.05, secondary significance threshold of 0.1, LD threshold of 0.05 and distance threshold of 250kb.

The second approach was to omit an initial GWAS and select SNPs randomly from across the genome. SNPs were first pruned down to an LD-independent set using a 250kb window, 5bp step-size and an LD ρ^2 threshold of 0.05. A minimum allele count of 6 was used.

2.2.5 Model Optimization

In order to implement nested-cross validation a modified version of the Python NestedCV package was used, allowing for logging and plotting of inner loop test results and pre-

specified random cross-validation splits. Cross-validation splits were made with pre-shuffling using Scikit-learn's KFold function (Pedregosa et al. 2011).

gBLUP prediction was performed in GCTA using the previously described GRMs and the subset of feature selected SNPs (Yang et al. 2010, 2014).

All regression, tree and kernel methods were performed using the Scikit-learn library in Python3. Neural networks were implemented using Keras on a TensorFlow backend (Abadi et al. 2015; Chollet et al. 2015). The mean absolute error (MAE) loss metric was chosen for all machine learning algorithms during fitting to prioritize robustness (Friedman et al. 2009).

Genotypes were loaded in as binary variables along with corresponding phenotypes. Phenotype values were standardized and scaled based on the training set prior to learning. Random grid searching and manual evaluation on the inner loops were implemented to optimize the hyperparameters (see Appendix A for details of the hyperparameters chosen for optimization).

2.2.6 Performance Assessment

In order to compare predictivity across all methods used, Pearson's correlation coefficient on the test set was calculated.

A multiple-comparison Dunnett test was used in order to test for significant differences in performance between the machine learning models and the baseline gBLUP model (Japkowicz and Shah 2011; Maxwell et al. 2018). This was implemented in a non-parametric and small sample size design using the nparcomp and mlt packages in R-3.6.3 (Hothorn and Kluxen 2019; Hothorn 2020; Konietschke et al. 2015).

Table 2.1 **SNP-Based Narrow-Sense Trait Heritabilities (h^2) of Four *Arabidopsis Thaliana* Traits**

Phenotype	**N**	**h^2_{SNP}**	**s.e.**
Flowering Time (10°C)	1058	0.943	0.057
Flowering Time (16°C)	1021	0.904	0.057
Seed Yield	384	0.236	0.101
Leaf Area	445	0.300	0.093

2.3 Results

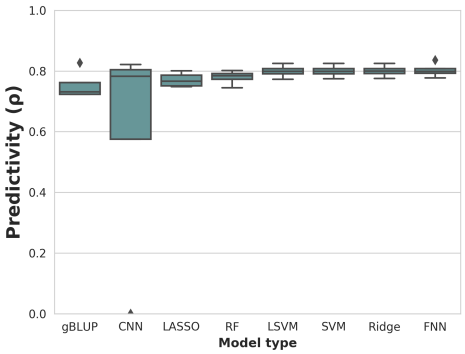
2.3.1 Heritability Estimation

The SNP-based narrow-sense heritabilities for each trait are shown in Table 2.1. For the flowering time traits, the values are relatively high, which may be explained by the fact that non-genetic variance was constrained by the laboratory conditions under which the plants grew. Given that nearly all variation in flowering time is captured by genetic effects, it is likely that genomic prediction models could be useful in accounting for trait variance in the population.

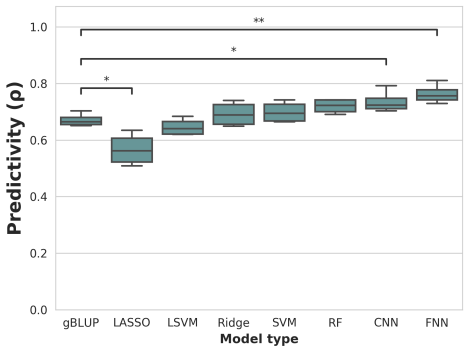
For the other two traits, seed yield and leaf area, the heritabilities were lower. It is possible that a larger element of random or developmental variation underlie the traits.

2.3.2 Genomic Prediction

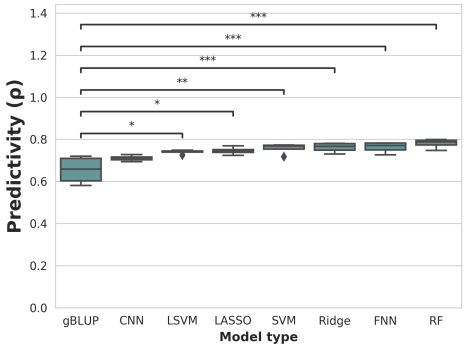
The results for the flowering traits can be seen in Figures 2.1 and 2.2. Some machine learning families consistently outperform the gBLUP method, often significantly so. Importantly, this was not true for all families and shows that the linear gBLUP serves as an important baseline upon which to improve. For context in terms of heritability and of the amount of variation being explained by the models, the R^2 of the gBLUP models are given in the figure legends.



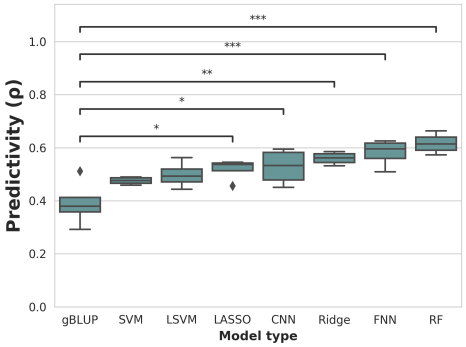
(a) Top 10,000 GWAS SNPs (clumped). Mean R^2 of gBLUP model is 0.57.



(b) Random 10,000 SNPs (pruned). Mean R^2 of gBLUP model is 0.45.



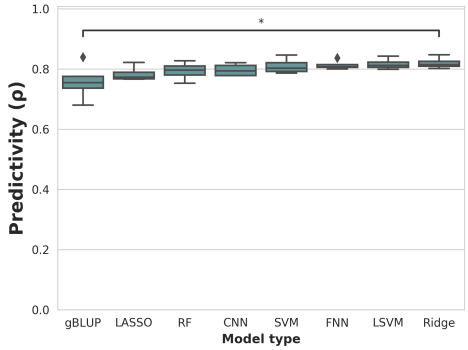
(c) Top 1,000 GWAS SNPs (clumped). Mean R^2 of gBLUP model is 0.43.



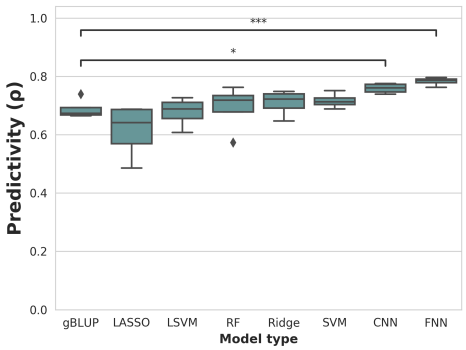
(d) Random 1,000 SNPs (pruned). Mean R^2 of gBLUP model is 0.16.

Fig. 2.1 **Nested Cross-Validation Results for Flowering Time (10°C)**[16]. Models are sorted by mean performance. Dunnett test p-value annotations: $*, p < 0.05$; $**, p < 0.01$; $***, p < 0.001$. gBLUP, genomic best linear unbiased prediction; CNN, convolutional neural network; RF, random forests; LASSO, least absolute shrinkage and selection operator regression; SVM, support vector machine; LSVM, linear SVM; Ridge, ridge regression; FNN, feed-forward neural network.

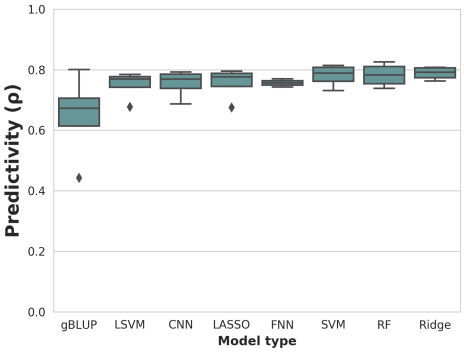
[16]Created using Matplotlib in Python3 (Hunter 2007).



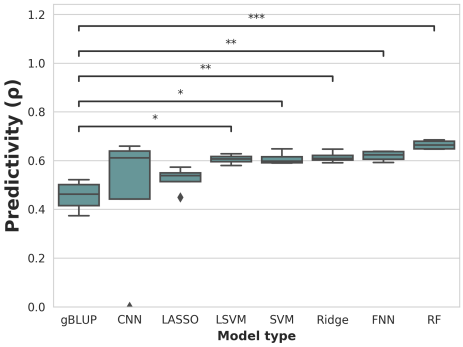
(a) Top 10,000 GWAS SNPs (clumped). Mean R^2 of gBLUP model is 0.58.



(b) Random 10,000 SNPs (pruned). Mean R^2 of gBLUP model is 0.47.



(c) Top 1,000 GWAS SNPs (clumped). Mean R^2 of gBLUP model is 0.44.



(d) Random 1,000 SNPs (pruned). Mean R^2 of gBLUP model is 0.21.

Fig. 2.2 **Nested Cross-Validation Results for Flowering Time (16°C)**[16]. Models are sorted by mean performance. Dunnett test p-value annotations: $*, p < 0.05$; $**, p < 0.01$; $***, p < 0.001$. gBLUP, genomic best linear unbiased prediction; CNN, convolutional neural network; RF, random forests; LASSO, least absolute shrinkage and selection operator regression; SVM, support vector machine; LSVM, linear SVM; Ridge, ridge regression; FNN, feed-forward neural network.

Feed-forward neural networks were the best method on average in terms of overall ranking, as well as the model that was significantly improved from the baseline the most times. For this reason, they could be seen as the machine learning family that has the most potential to improve over gBLUP. Although neural networks tend to have a lot of tuned hyperparameters, and are thus prone to overfitting, this was not seen to a detrimental extent in the outer loops of the nested cross-validation. Ridge regression, which is a linear model, also performed generally quite well.

Random forests were seen to be the most predictive model in some of the sets, but also often failed to improve over the gBLUP, making it hard to strongly recommend. Convolutional neural networks also tended to have good predictive performance except in the cases where there was clear failure on one of the cross-validation sets due to model misspecification (see Figs. 2.1a, 2.2d). This instability in prediction is an important observation and would not have been seen if not for the nested design. For this reason, in future model-building, it would be recommended to look closely for full robustness on any validation sets before moving to a final test set.

Two overall trends are also clear from the plots: that using a larger SNP-set size leads to improved performance; and that feature selection through linear association (GWAS) also aids in decreasing predictive error. Examples of prediction scatter plots for each of the traits under consideration are given in Fig. 2.3.

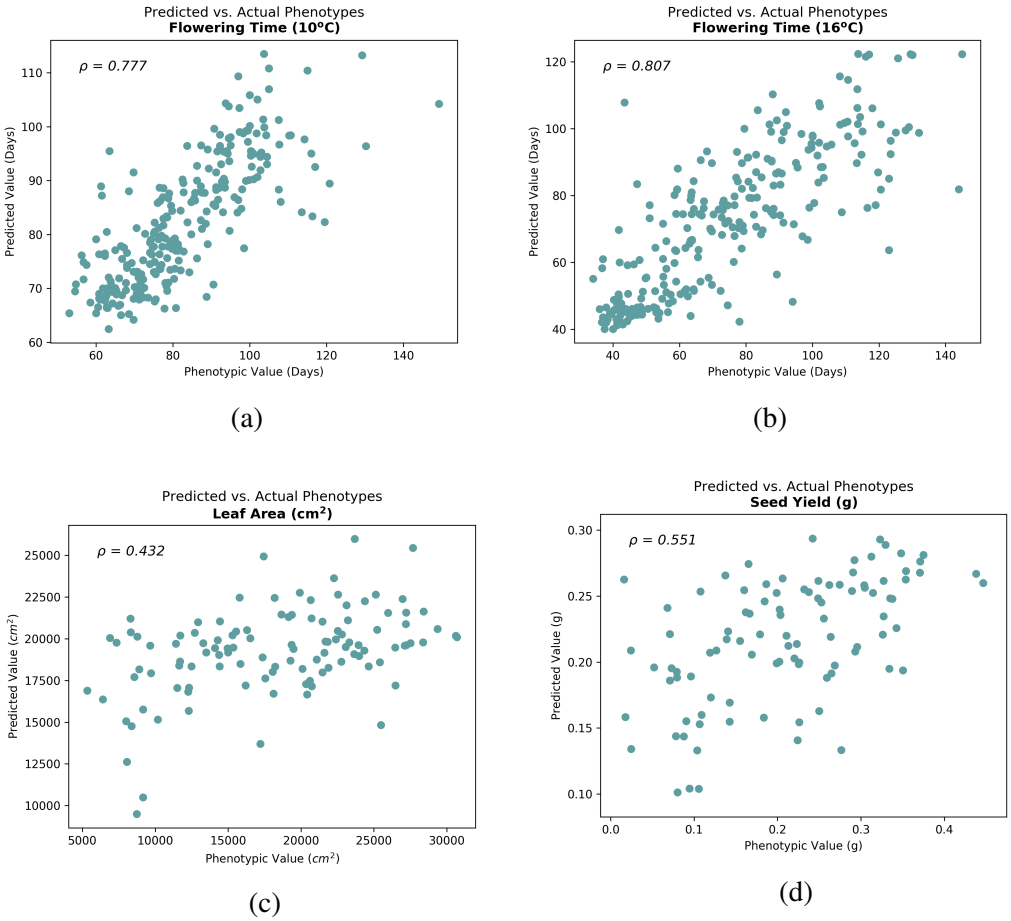
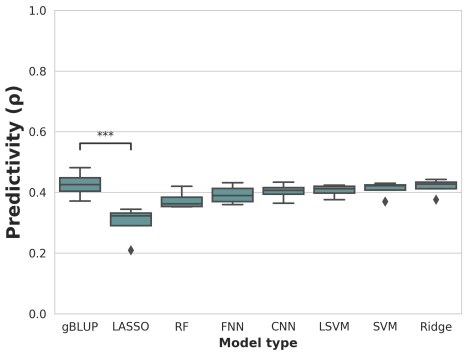
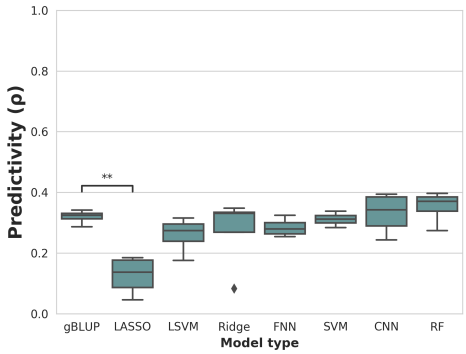


Fig. 2.3 **Scatter Plots of Predicted vs. Actual Phenotypic Values for Four *Arabidopsis Thaliana* Traits**[16]. These graphs illustrate the results of a single outer validation loop using feed-forward neural networks as predictors. Pearson's Correlation Coefficient values (ρ) of the predictions are shown.

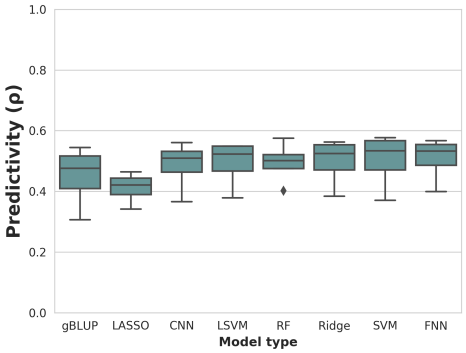
The statistical comparisons between models for the other two traits, seed yield and leaf area, can be seen in Figure 2.4. Unlike the flowering traits, none of the machine learning models was able to significantly outperform the baseline gBLUP method. In fact, some of the models significantly underperformed, particularly LASSO regression.



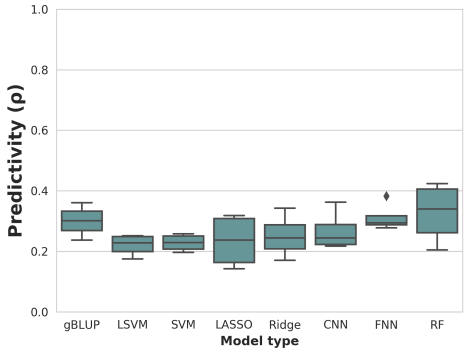
(a) Top 10,000 GWAS SNPs (clumped) of Leaf area. Mean R^2 of gBLUP model is 0.18.



(b) Random 10,000 SNPs (pruned) of Leaf area. Mean R^2 of gBLUP model is 0.10.



(c) Top 10,000 SNPs (clumped) of Seed yield. Mean R^2 of gBLUP model is 0.21.



(d) Random 10,000 SNPs (pruned) for Seed yield. Mean R^2 of gBLUP model is 0.09.

Fig. 2.4 **Nested Cross-Validation Results for Seed Yield and Leaf Area**[16]. Models are sorted by mean performance. Dunnett test p-value annotations: $*, p < 0.05$; $**, p < 0.01$; $***, p < 0.001$. gBLUP, genomic best linear unbiased prediction; CNN, convolutional neural network; RF, random forests; LASSO, least absolute shrinkage and selection operator regression; SVM, support vector machine; LSVM, linear SVM; Ridge, ridge regression; FNN, feed-forward neural network.

2.4 Discussion

The results presented here show the potential for non-linear models to improve upon linear genomic prediction in *Arabidopsis*. The possibility for substantial improvement is worth pursuing given the struggles in the field to close the missing heritability gap despite ever

denser marker sets. Improvements in prediction were dependent on model type, feature set size, feature selection method, and trait.

Overall, the baseline gBLUP model yielded reasonable predictions in comparison to the machine learning models, a not unexpected result from prior research. However, for flowering traits, some machine learning families could be seen to perform consistently better than the baseline model, such as neural networks and ridge regression. This improvement in performance was often significant, despite the lower statistical power the cross-validation design gave for performance assessment. Several previous studies on machine learning in genomic prediction have omitted statistical tests, presumably on the basis of low power, although it is an important component of assessing improvements in performance.

For the other two traits, none of the machine learning models was able to significantly outperform the baseline model. These traits had lower heritabilities and smaller sample sizes than flowering time which could be contributing factors to this result. This adds to the existing body of evidence that improvement in performance using machine learning is not guaranteed and is highly dependent on trait and dataset.

Performance in prediction generally improved when using more SNP markers. Despite a much smaller number of samples than predictors, overfitting with so many variables did not seem to be a major issue, as might happen in $P \gg N$ problems.

Additionally, prior linear association for feature selection was found to be a useful step in this study, which is not guaranteed when modelling non-additive interactions (Hua et al. 2009; Wolff et al. 2019); this performance boost was not extreme, however.

Interestingly, the performance improvement for the best non-linear models over the linear models was more pronounced when using random SNPs, which, unlike the top GWAS SNPs, were not chosen for their prior linear association. It is perhaps understandable that the most linearly associated features tend to explain the most variance when combined in a linear fashion; however, it is evident that a substantial amount of variation stands to be captured

from the rest of the genome, where non-linear effects may play a larger role (Krakovska et al. 2019; Urbanowicz et al. 2018). For this reason, it is often noted, and this study reiterates, that linear feature selection may overlook important variables that combine non-linearly. Notwithstanding this possibility, due to computational limitations, a subset of SNPs had to be chosen as input in this study and it was found from this study that the optimal input features were the most significant QTLs obtained from a GWAS. As computing power grows, however, one should remain open-minded as to the input search space when using genotypic information to predict complex traits.

For example, it is possible that widespread biological and statistical epistasis exists for these traits which is why the non-linear models were sometimes successful in explaining more trait heritability. Therefore, a better understanding of the total genetic architecture of a trait could aid in building more powerful models (Ober et al. 2015). One of the drawbacks of machine learning, however, is that the models are often "black boxes" that are not readily interpretable (Friedman et al. 2009). Although improved prediction can hint towards a greater role for epistasis in the variance of complex traits, elucidating the specific complex interactions is not a simple task. One possible extension of this study would be to find putative epistatic interactions through linear methods in the feature selection step.

Although many machine learning models are black boxes in terms of causal inference, it might still be an important goal of a prediction tool to identify relevant genes, variants and pathways that are being used by the model. By examining the best-performing SNP input sets some insight into genetic architecture may be obtained, however, machine learning approaches will most likely remain less useful as tools to understand the biological mechanisms underlying genomic prediction.

Genomic prediction is concerned not just with performance, but also with ensuring that methods and models are robust enough to be reliable, consistent and transferable across datasets. This study used a nested-cross validation approach for hyperparameter tuning in

order to assure the reliability of the results. This approach was useful in identifying a pitfall of hyperparameter tuning using *k*-fold cross-validation, namely the potential to miss serious misspecification on the final models despite success on validation sets.

Limitations of this analysis include the modest sample sizes for the various traits and the lack of a fully independent external test dataset. In the field, there would also be a larger non-genetic component of variation, as well as genotype-environment interactions, which could possibly affect the relative performances of the various machine learning models. Furthermore, the accessions in this analysis were inbred to homozygosity which simplified the input genotype matrices and removed intra-allelic effects such as dominance. Any of these considerations could cause a difference in model performance in natural populations (Charmantier et al. 2014).

As the feature selection step was linear in this analysis, the GRM employed in the MLMA should have mitigated the effect of population structure as per the original study protocol described in Alonso-Blanco et al. (2016), yet the issue of confounding cannot be fully discounted.

In the field of machine learning, the most appropriate methods for handling confounding variables which are often needed in genetic studies, such as through using PCs as linear covariates in a regression, have not been well established which is discussed further in Chapter 4 (He et al. 2019; Li et al. 2011; Whalen et al. 2022).

The accessions in this analysis were grown under standardized laboratory conditions which theoretically removes the effect of confounding by means of a correlation between genotype and environmental exposure (e.g. differences in soil type and mineral concentrations). However, to the extent that there is a correlation between ancestry informative SNPs and causal loci (and ignoring genotype-environment interactions), fully accounting for this population structure may reduce predictive performance in such a way as to be undesirable (Barton et al. 2019; Wray et al. 2013). As the final model would ultimately be predicting

phenotype from genotype, such information should not necessarily be excluded when trying to maximize the explained variance across real populations in potentially varied environments.

From this research, it would appear that machine learning should continue to be explored in the context of genomic prediction in plants. Further work is necessary, to quantify the exact magnitude of the potential benefit depending on the optimization of the techniques.

Chapter 3

Genomic Prediction in Amyotrophic Lateral Sclerosis

3.1 Introduction

3.1.1 Human Complex Trait Genetics

The results of heritability estimates from the twentieth century show that almost any human trait under consideration has at least some genetic component (Turkheimer 2000). In the twenty-first century, with the advent of large-scale genotyping tools such as SNP-chips and whole-genome sequencing, the genetics of many complex traits and disorders have been more extensively studied and previous heritability calculations generally recapitulated as accurate (Visscher et al. 2017; Wainschtein et al. 2022). One of the major findings from these last two decades has been that the majority of human traits and disorders are massively polygenic in their architecture.

These observations have led to the development of an "omnigenic model" that can be seen as an extension of the infinitesimal model introduced in Section 1.1.1 (Boyle et al. 2017). The infinitesimal model allowed for essentially all genetic variants to have a very

small effect on the variation seen in a phenotype. The omnigenic model constrains this by assuming that heritability is only conferred by those genetic variants that are within gene regulatory networks connected to core trait genes expressed in the biologically relevant cell types[17]. Although core genes do indeed contribute to phenotypic variation, the sheer scale and interconnectedness of cellular pathways result in the majority of variation stemming from "peripheral" genes. This hypothesis helps to explain both the vast polygenicity of complex traits while also accounting for the observation that heritability tends to be enriched in genomic regions relevant to known disease pathways (Kundaje et al. 2015; Maurano et al. 2012). The extent to which the omnigenic model holds varies by disease, but it is a useful conceptual framework with which to understand the general results from massive GWAS analyses of complex traits.

Despite the problems that polygenicity naturally creates for the detailed understanding of complex traits, the genetic architecture of many phenotypes has become more well-elucidated and major headway has been made in being able to account for variance in the population, with the aim of utilizing PRS in the clinic (Lewis and Green 2021). For some traits, such as standing height, thousands of independent loci have been significantly associated with the trait. These variants are enriched in trait-relevant pathways, in line with the omnigenic model, and the majority of heritability has recently been accounted for (Yengo et al. 2022). When assessing the predictive performance of a PRS built from the latest height GWAS even a within-siblings analysis explains much of the observed variation. This suggests that population structure may not be playing a major confounding role in this study (see Section 1.3.4).

[17] Boyle et al. (2017) define core genes as those "modest number of genes or gene pathways with specific roles in [trait] etiology, as well as their direct regulators".

Although height has been found to be particularly amenable to genetic quantitative analysis, the ability to explain a large part of population-level variation of human disorders offers an exciting opportunity for personalized medicine. For example, a polygenic score developed for coronary artery disease (CAD) can identify individuals in the tail end of the score distribution that are at a three-fold higher risk of CAD than the general population (Khera et al. 2018). This level of risk increase is similar to that seen in monogenic cases of CAD. As sample sizes of disease GWASs grow, it is possible that many common disorders could be anticipated in advance based on a genotyping array, the cost of which continues to decrease. It has also been shown that genetic information can be a useful tool in predicting a patient's response to treatment (Johnson et al. 2022). Even in the cases where individual-level prediction is not accurate enough to justify prophylactic treatment, PRS could be an efficient and cost-effective population-level tool for choosing earlier ages of screening for disease. This would be in addition to using variables such as age and/or family history if such effects can be shown to be independent (Lewis and Green 2021).

3.1.2 Amyotrophic Lateral Sclerosis

Amyotrophic Lateral Sclerosis (ALS) is a progressive and fatal neurodegenerative disease, the symptoms of which are ultimately caused by the degeneration and death of upper and lower motor neurons (Hardiman et al. 2017; Mejzini et al. 2019). It is classified under the larger umbrella of Motor Neurone Diseases (MNDs) and can also be referred to as Lou Gehrig's Disease.

The first symptoms of ALS occur at an average age of onset of 65 years - with that figure being slightly lower in Asia and South America than in Europe. A lifetime ALS risk of approximately 1 in 350 individuals has been found across populations (Al-Chalabi and Hardiman 2013). Disease progression can be rapid, with variable sites of first onset, and the disorder manifests with difficulty with motor function, swallowing, and speech impairment.

Death generally occurs within two to five years from first onset, usually due to failure of respiratory muscle function. Cognitive impairment can be seen in up to half of all patients. Furthermore, in around 10% of cases, the disease manifests with concomitant Frontotemporal Dementia (FTD) (Phukan et al. 2012).

ALS is generally found to be sporadic in origin, although 10% of cases show a clear familial inheritance. The genetic contribution across both forms is known to be high, with heritability estimates generally being around 50-60% (Al-Chalabi et al. 2010; Ryan et al. 2019). More success has been had in finding causative genes in familial ALS, although many of these genes have also been linked to sporadic cases.

Overall, the genetic origin of the disease is known in around 40-55% of familial cases and less than 10% of sporadic cases (Al-Chalabi et al. 2017; Zou et al. 2017). In European patients with both sporadic and familial ALS, a repeat expansion in the *C9orf72* gene is the most common cause of disease. This expansion accounts for 5% of familial cases and 30% of sporadic cases in the European population (DeJesus-Hernandez et al. 2011; Renton et al. 2011). Other major genes known to be involved in ALS include *SOD1*, *TARDDP*, *FUS*, and *ATXN2*.

Although monogenic inheritance is seen in some cases, and much progress has been made in the last twenty years in identifying causative genes, the full genetic architecture of ALS is largely unknown. This is complicated by the incomplete penetrance of some mutations, low prevalence of disease, and the large extent to which rare variation is thought to play a role. The exact contributions from oligogenic and polygenic variation, either as drivers or modifiers of disease have not been fully elucidated (McCann et al. 2020; Renton et al. 2014). The largest GWAS to date was conducted with 29,612 cases and 122,656 controls, identifying 15 significant risk loci (Van Rheenen et al. 2021). The narrow-sense SNP-based heritability estimate was 3%, significantly lower than the previous estimate of 9% (Van Rheenen et al. 2016). The most significant locus associated with ALS is the region

containing the aforementioned pathogenic repeat-expansion in *C9orf72*. Other highlighted loci include previously described ALS genes including *SOD1* while also implicating some previously unlinked genes.

Examinations into the function of these genes using cellular, tissue-based, and whole-organism models point towards a complex, heterogeneous pathology underlined by aberrations in a variety of processes including RNA processing, immune regulation, oxidative stress response, nucleocytoplasmic transport, DNA repair and mitochondrial function (Hardiman et al. 2017; Mejzini et al. 2019). Interestingly, the TDP-43 protein from *TARDDP* has been seen to be a part of cytoplasmic inclusions that are found in 97% of postmortem inspections of patient motor neurons. These inclusions are similar to the aggregations seen in other neurodegenerative conditions such as Alzheimer's and Parkinson's disease (Mackenzie et al. 2007; Maekawa et al. 2009). It is not yet known whether these insoluble inclusions are a byproduct of disease, or a cause of it. Furthermore, it remains a possibility that prion-like spread of proteinaceous inclusions could be an underappreciated disease mechanism.

There is no known cure for ALS and currently available drug treatments have minimal impact on overall life expectancy (Hardiman et al. 2017). For this reason, it is imperative that more is known about the etiology of the disease so as to improve the chance of finding effective or novel drug targets and treatments.

3.1.3 Imbalanced Data

Problems can arise when utilizing ML techniques on imbalanced data where the frequency of one class is much greater than the other (Krawczyk 2016). This class imbalance is a common occurrence in population-level biobanks of human genetic data for rare diseases. To illustrate the issue, if the majority class comprises 99% of the examples in a dataset, a classifier can easily achieve 99% accuracy by ignoring all input features and by simply outputting the majority class status to all samples.

In a disease setting, if there is an over-representation of controls relative to cases, these problems inherent to imbalanced data may arise. Generally, during data collection, an effort is made to oversample the minority class but, depending on the metric in use, a roughly 50:50 ratio may be most appropriate. For example, the AUC (see Section 1.2.4), the most commonly used metric in complex disease genetics, can be quite sensitive to class imbalance (Cook 2007).

A simple way to overcome this problem would be to discard excess majority class examples, although this is clearly a waste of data. Similarly, under-represented classes may be oversampled until a balanced ratio is achieved, which is generally preferable. Newer techniques that can generate novel examples of the under-represented class using feature combination methods are also possible, although their appropriateness to complex data involving LD and unknown genetic architecture is not clear (Krawczyk 2016).

3.1.4 Bayesian Optimization

Exploring which hyperparameter combinations are best suited to fit a model through grid-searching can be a time-consuming process, especially in large datasets. To overcome some of the challenges brought about by combinatorial testing, a Bayesian approach can be taken when tuning in order to systematically prioritize certain hyperparameter combinations when training and evaluating models (Géron 2019). Bayesian hyperparameter optimization keeps a record of previous combination results; and can be more efficient in exploring the overall search space than naively grid-searching through the full suite of possible combinations[18].

[18]Generally, if time is a limiting factor, random grid-searching is used, whereby random combinations of hyperparameters are chosen and evaluated to get a broad picture of the optimal values, rather than full grid-searching.

3.1.5 Previous Work

Several previous studies have been published on the use of machine learning using human genetic data to predict complex diseases (see reviews by Bracher-Smith et al. (2021); Katsaouni et al. (2021)). While it is clear that ML techniques do indeed have the potential to out-compete the standard linear PRS approaches; the exact magnitude of the benefit remains unclear. Many questions also remain unanswered as to best practices when approaching this task such as algorithm choice, model comparison, feature selection methods, interpretability, and bias reduction. Specifically with regards to ALS, important previous work using ML techniques has pointed towards a role for non-additive interactions in the genetic architecture of the disease. The authors recommend using genome-scale data to further interrogate the genotype-phenotype map (Yin et al. 2019).

3.2 Materials and Methods

3.2.1 Data

The ALS genetic data come from a cohort of 12,577 case and 23,475 control individuals of European ancestries that were used in a previously published GWAS (Van Rheenen et al. 2016). QC was done on this data to recapitulate the prior GWAS as faithfully as possible. As there is an excess of controls in this dataset oversampling was used to balance the sets in an equal ratio of cases to controls. Oversampling was performed through random duplication of individuals. A nested cross-validation approach was also employed with the outer split being 10-fold and the inner split being 4-fold. In order to first extract the most linearly associated

SNPs during feature selection, a GWAS was performed using PLINK2, with and without correcting for the top three principal components[19].

Only autosomal chromosomes were considered in these analyses. Clumping was performed on SNPs using the summary statistics files from the GWAS. The clumping parameters were as follows: significance threshold of 0.005, secondary significance threshold of 0.05, LD ρ^2 threshold of 0.1 and distance threshold of 100kb. These SNPs were then fed into the machine learning pipeline described below.

3.2.2 Experimental Approach

Using genomic data from the samples of ALS, the objective of this analysis was to compare the relative performance of the various machine learning methods to the baseline linear models. In addition to baseline PRS, the performance of logistic LASSO, Ridge classification, random forests, linear support vector machines, feed-forward neural networks and convolutional neural networks was assessed[20]. The influence of feature set size and effect of prior linear feature selection were also investigated in this analysis. Prediction accuracies of machine learning methods were compared against PRS prediction as a baseline.

3.2.3 Feature Selection

The genotype dataset consisted of 1,250,041 HapMap SNPs. Two feature selection strategies were adopted in order to select the input SNP set from the genotype dataset, with feature

[19]This number of components was chosen as per Van Rheenen et al. as well as the inflection point from a scree plot (see Fig. 4.4a in Chapter 4). The first two principal components of variation are plotted against one another in Fig. A.1a

[20]Scikit-learn's non-linear SVM's were impractical for the ALS dataset as fit-time scales at least quadratically with the number of samples. See https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html for details.

selection performed independently on all inner and outer loops of this analysis. The first strategy was to use a GWAS to identify the SNPs most linearly associated with the trait. The results were subsequently clumped to generate the set of top SNPs free from strong linkage disequilibrium. A GWAS was performed on the ALS phenotype using PLINK2 association (Chang et al. 2015).

The second approach was to create a set of SNPs without prior association by random selection.

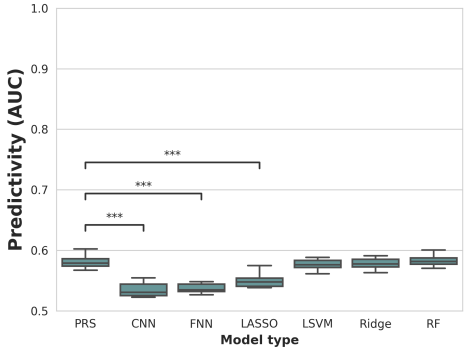
3.2.4 Model Optimization

In order to implement nested-cross validation a modified version of the Python NestedCV[21] package was used, allowing for logging and plotting of inner loop test results and pre-specified random cross-validation splits. Cross-validation splits were made with pre-shuffling using Scikit-learn's KFold function (Pedregosa et al. 2011). All regression, tree, and kernel methods were performed using the Scikit-learn library in Python3. The liblinear solver was chosen for LASSO construction. Neural networks were implemented using Keras on a TensorFlow backend using binary cross-entropy (logistic) loss (Abadi et al. 2015; Chollet et al. 2015). Genotypes dosages (coded as 0, 0.5 or 1) were loaded in along with corresponding phenotypes. Random grid searching and manual evaluation on the inner loops were implemented to optimize the hyperparameters (see Appendix A for details of the hyperparameters chosen for optimization). Polygenic risk scoring was conducted using PRSice software version 2.1.11 (Euesden et al. 2015). SNP effect-size estimates from the prior GWAS were used as the base statistics for the PRS analysis (including for randomly selected SNPs). P-value thresholding was not implemented for the PRS analysis to ensure

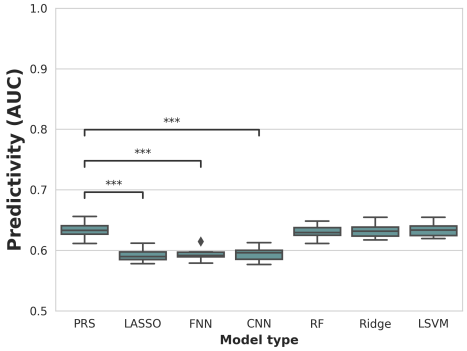
[21] Available from: https://github.com/casperbh96/Nested-Cross-Validation

that the SNP-set sizes were equal between the baseline approach and the machine learning models. Calibration was assessed by visual inspection of reliability diagrams (see Section 1.2.4) (Van Calster et al. 2019, 2016). Calibration enhancement was conducted using Scikit-learn's CalibratedClassifierCV method with sigmoid Platt scaling fitted on the training data using ten bins.

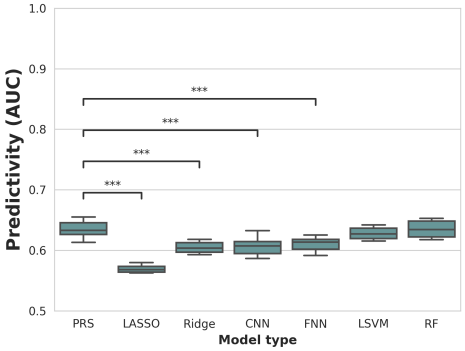
3.3 Results



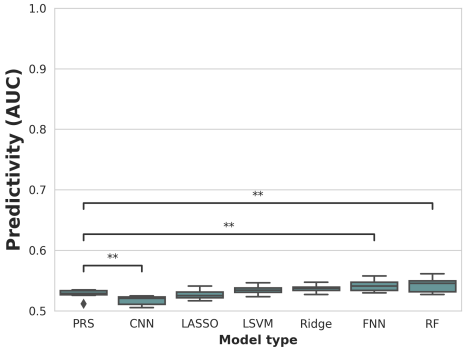
(a) Random 1,000 SNPs



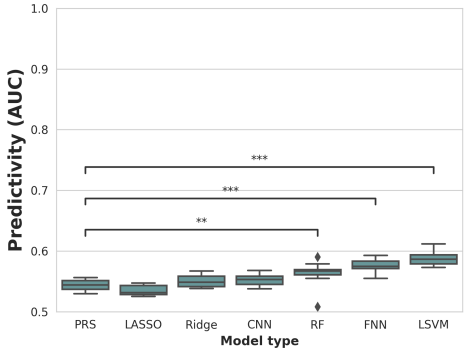
(b) Top 1,000 GWAS SNPs



(c) Top 10,000 GWAS SNPs

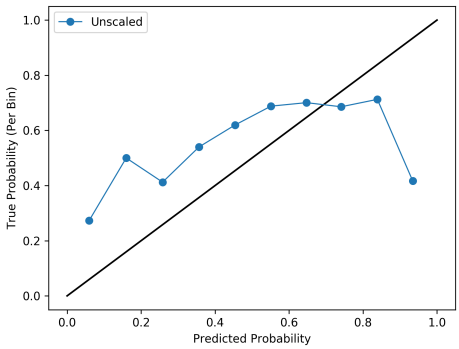


(d) Top 1,000 GWAS SNPs (PC-adjusted)

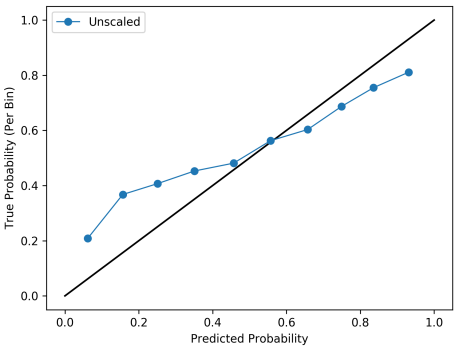


(e) Top 10,000 GWAS SNPs (PC-adjusted)

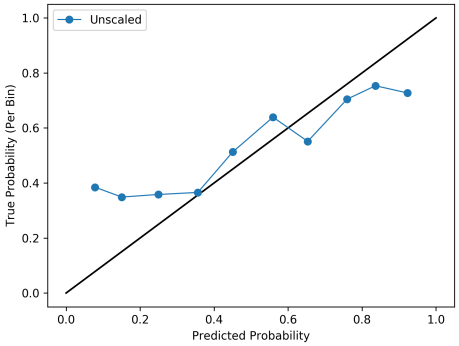
Fig. 3.1 **Nested Cross-Validation Results for ALS:**[16] Dunnett test results are shown between models for various SNP-selection strategies. Models are sorted by mean performance. Dunnett test p-value annotations: $*, p < 0.05$; $**, p < 0.01$; $***, p < 0.001$. PRS, polygenic risk score; CNN, convolutional neural network; RF, random forests; LASSO, least absolute shrinkage and selection operator regression; LSVM, linear support vector machine; Ridge, ridge regression; FNN, feed-forward neural network.



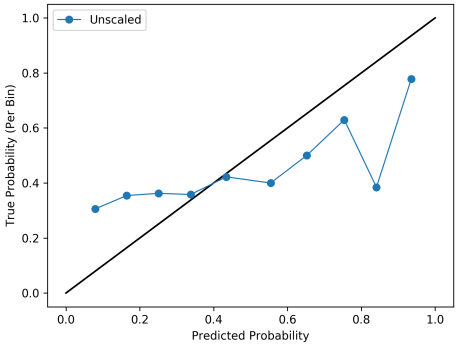
(a) Random 1,000 SNPs



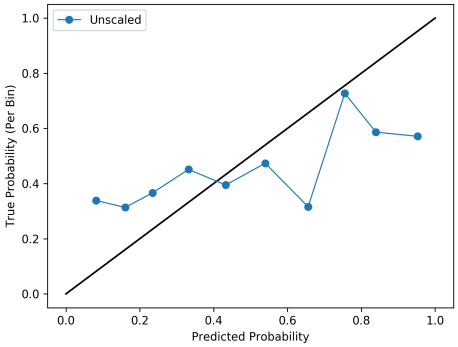
(b) Top 1,000 GWAS SNPs



(c) Top 10,000 GWAS SNPs

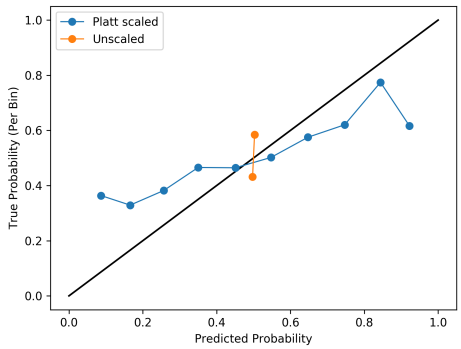


(d) Top 1,000 GWAS SNPs (PC-adjusted)

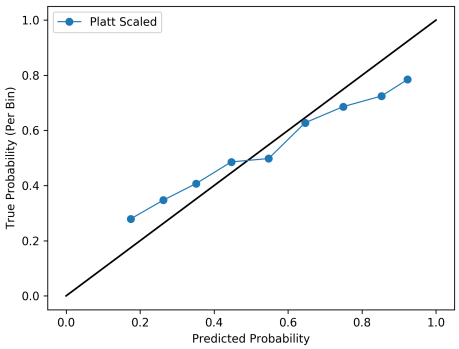


(e) Top 10,000 GWAS SNPs (PC-adjusted)

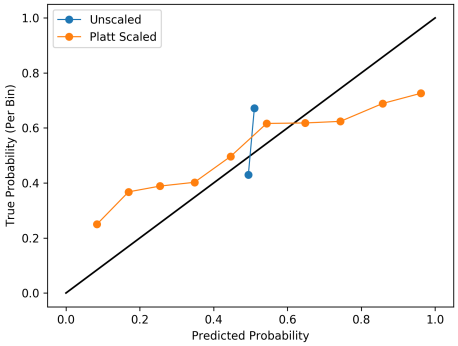
Fig. 3.2 **Calibration Results for PRS Models:**[16] Representative calibration results are shown for PRS models for various SNP-selection strategies. A well-calibrated model lies on the diagonal. Prediction outputs are grouped into separate probability bins and the actual case rate in each bin is assessed.



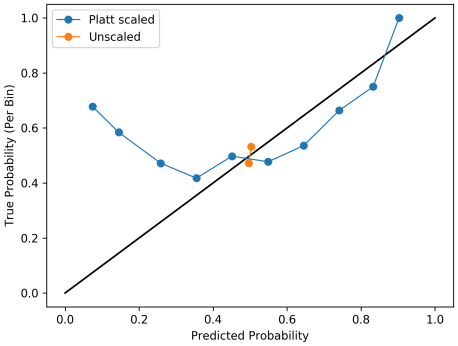
(a) **RF:** Random 1,000 SNPs



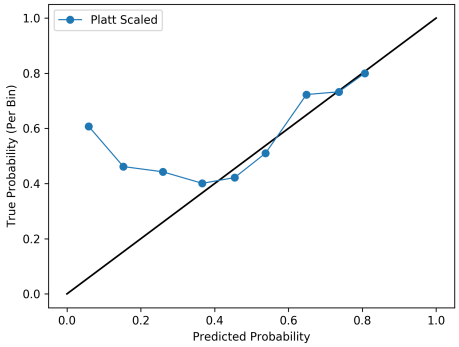
(b) **LSVM:** Top 1,000 GWAS SNPs



(c) **RF:** Top 10,000 GWAS SNPs

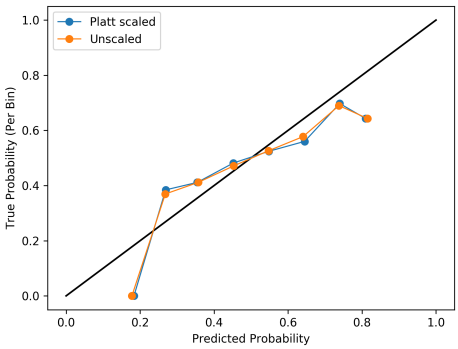


(d) **RF:** Top 1,000 GWAS SNPs (PC-adjusted)

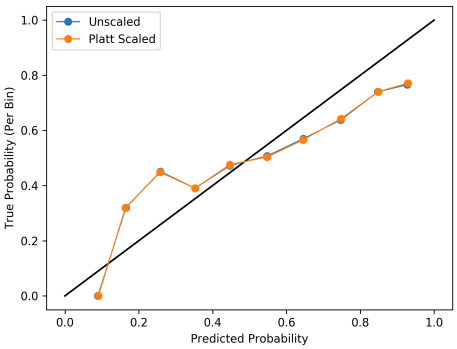


(e) **LSVM:** Top 10,000 GWAS SNPs (PC-adjusted)

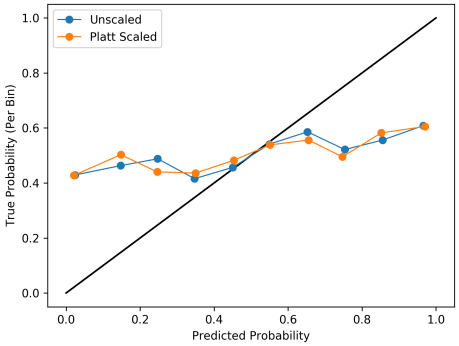
Fig. 3.3 **Calibration Results for Best-Performing ALS Models:**[16] Representative calibration results are shown for the best performing models for various SNP-selection strategies. A well-calibrated model lies on the diagonal. Prediction outputs are grouped into separate probability bins and the actual case rate in each bin is assessed. RF, random forests; LSVM, linear support vector machine.



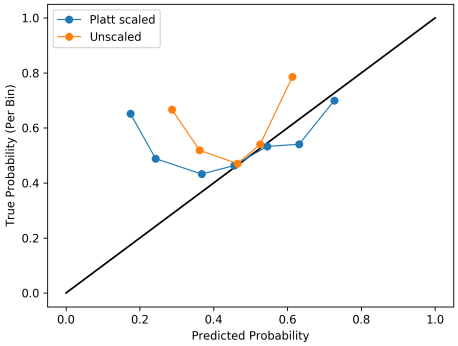
(a) **CNN:** Random 1,000 SNPs



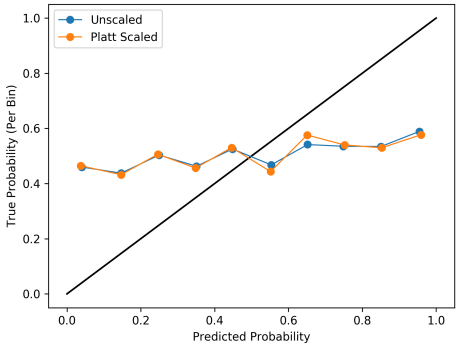
(b) **LASSO:** Top 1,000 GWAS SNPs



(c) **LASSO:** Top 10,000 GWAS SNPs



(d) **CNN:** Top 1,000 GWAS SNPs (PC-adjusted)



(e) **LASSO:** Top 10,000 GWAS SNPs (PC-adjusted)

Fig. 3.4 **Calibration Results for Worst-Performing ALS Models:**[16] Representative calibration results are shown for the worst performing models for various SNP-selection strategies. A well-calibrated model lies on the diagonal. Prediction outputs are grouped into separate probability bins and the actual case rate in each bin is assessed. CNN, convolutional neural network; LASSO, least absolute shrinkage and selection operator regression.

3.3.1 Predictive Performance

The Dunnett test results comparing model types for each SNP-selection strategy can be seen in Figure 3.1. Random forests were both the most commonly best-performing model as well as the model type that significantly improved upon the baseline PRS the most times (tied with feed-forward neural networks). However, although random forests never performed substantially worse than the baseline model, the improvements were generally modest when they did occur. As in previous work, PRS is an important baseline against which to gauge performance. Although only significantly outperforming the baseline model only once, linear SVMs performed generally quite well.

As stated above, feed-forward neural networks were tied with random forests as the model type that significantly improved upon the baseline PRS the most times, however, they also significantly underperformed relative to PRS multiple times making it hard to strongly recommend over baseline. The CNN models generally underperformed relative to the baseline PRS while the median LASSO AUC was lower than that of the PRS model in every experiment.

As also observed in Section 2.3.2, using prior linear association was an effective method to improve prediction performance relative to a randomly selected subset of independent SNPs. However, the inclusion of principal components as covariates during the initial GWAS step could be seen to substantially decrease the general performance of the models (compare Figs 3.1b and 3.1d with one another, as well as 3.1c and 3.1e). Interestingly, increasing the SNP-set size only significantly improved performance in the PC-adjusted experiments (Figs. 3.1d and 3.1e).

Examples of AUC plots for each of the best-performing models are given in Fig. 3.5. Overall, there is no large discrimination between cases and controls achieved, even from the best-performing models.

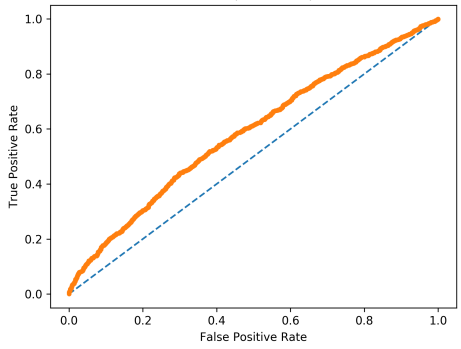
3.3.2 Calibration

As explained in Section 1.2.4, predictive performance using the AUC discrimination metric is only one aspect of a model's performance that one can analyze. Model calibration may be seen as equally important, depending on the exact task at hand (calibration is especially relevant if risk estimates are given to individual patients).

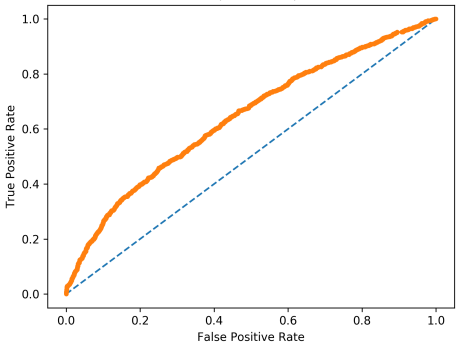
For context in terms of baseline performance, representative reliability diagrams are shown in Fig. 3.2. Calibration is generally good, in Fig. 3.2b the risk probabilities of the output generally correspond to the observed case events in their respective bins. However, in Fig. 3.2e the observed case numbers remain constant across predicted probabilities of up to ~ 0.7.

Representative calibration plots from the best-performing machine learning models can be seen in Figure 3.3. After Platt scaling to output probabilities, random forests achieved generally good calibration (see Figs. 3.3a and 3.3c). However, random forests were not always optimally calibrated as can be seen in Fig. 3.3d, where there was under-calibration except for those predicted to be at the highest level of risk. Likewise, a similar calibration problem for the LSVM model can be seen in Fig. 3.3e. Nevertheless, the calibration of the top-performing models was generally respectable.

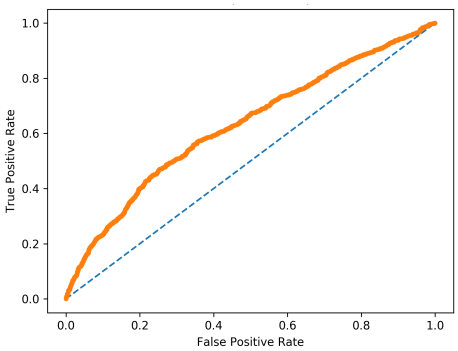
For the worst-performing models, this poor performance in discrimination did not necessarily translate to inferior calibration as can be seen from Figs. 3.4a and 3.4b where the CNN and LASSO models have quite reasonable calibrations. However, very poor calibration can be observed in Figs. 3.4c and 3.4e where the observed case rate remained almost constant over the entire predicted risk probability levels.



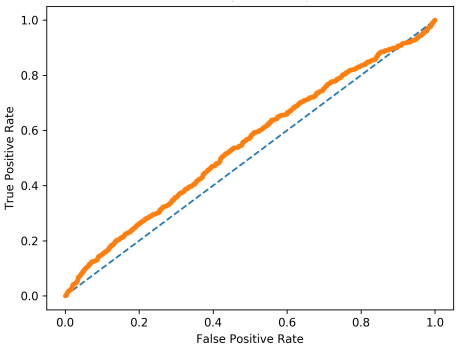
(a) **RF:** Random 1,000 SNPs



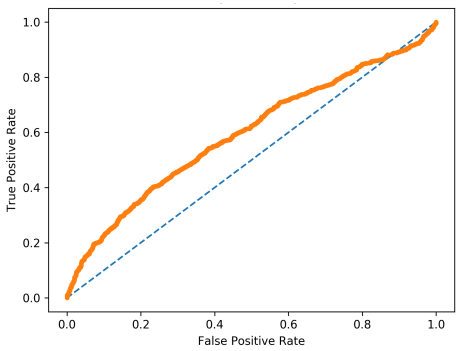
(b) **LSVM:** Top 1,000 GWAS SNPs



(c) **RF:** Top 10,000 GWAS SNPs



(d) **RF:** Top 1,000 GWAS SNPs (PC-adjusted)



(e) **LSVM:** Top 10,000 GWAS SNPs (PC-adjusted)

Fig. 3.5 **AUC Samples of ALS Models:**[16] AUC results are shown for the best performing model across various SNP-selection strategies. RF, random forests; LSVM, linear support vector machine.

3.4 Discussion

The results presented here show that, regardless of model type, the creation of genomic prediction models using SNP data for this ALS sample does not result in very high discrimination between cases and controls. However, it is possible that space remains for incorporating non-additive interactions when calculating risk for ALS as random forest models were often able to significantly improve upon the baseline PRS models. It is difficult to ascertain the exact source of the improvement in discrimination, whether or not non-additive interactions between or within genes are being exploited. A similar result has been described in some previous research, improvements upon linear models are generally modest when they are achieved. Nevertheless, given the high heritability of ALS and the difficulties thus far of accounting for much of the estimated variance, even small improvements upon previous efforts should be welcomed.

As this work compared several different machine learning algorithms, it highlights the difficulty in predicting ahead of time which model type may be the most suitable for the task at hand. Algorithm performance may depend on trait architecture and specific sample details such as linkage or SNP-set size.

The calibration seen from the top-performing models was generally acceptable, and Platt scaling was seen here to be effective in improving probability estimates. Evaluation of calibration may sometimes be neglected but it is especially important when the relative performance of machine learning families, of which there are many, are gauged. Especially in the case of severe disease, proper calibration is clinically essential as under-forecasting of risk may lead to missed opportunities for screening and treatment, and over-forecasting may cause emotional distress to patients and result in wasted healthcare expenditure.

Although the sample size here was relatively large (12,577 cases and 23,475 controls), larger databases of ALS genomic samples exist and it is possible that predictive performance may improve as these databases are used for similar investigations. However, the observed

AUCs of approximately 0.6 fall quite short of the minimum value of 0.75 which is said to be clinically desirable. Although ALS is generally not predictable outside of familial settings, it might one day be useful for the polygenic component to be assessed, even if only as a modifier of disease onset or severity.

ALS may not be the ideal disorder within which to study omnigenic effects using SNP-data as rare variation is known to play a large role in disease etiology. However, estimates of the polygenicity of the trait have differed across studies and the role of genetic interactions had remained under-explored. Nevertheless, it would appear that there is indeed a polygenic component being captured by these models as the 10,000 SNP models substantially improved upon the 1,000 SNP models, even after correction for population structure.

Although it would be interesting to move beyond SNP data to investigate the role of epistasis in ALS using machine learning, computational power will remain an obstacle to incorporating more and more variation into model-building pipelines. However, focusing on the variation within previously described ALS genes to build genetic risk scores could be a worthwhile future approach. Furthermore, it may be beneficial for more difficult traits such as ALS to see more concrete agreement reached in the literature, using more extensively evaluated traits and disorders, on best machine learning procedures and optimal parameter search spaces. There has been recent growth in large-scale genomic databases that capture many rare and common diseases in the human population such as the *UK BioBank* and the *All of Us* research programme (Sankar and Parker 2017; Sudlow et al. 2015). These biobanks may facilitate further insight into the best approaches to modeling non-additive interactions in genetic data.

For simplicity, and as is common in large-scale genomic association and prediction, sex chromosomes were excluded from this analysis. However, future work could build on more recent frameworks that have been developed to deal with sex contributions to genetic architecture and genomic risk profiling (Bernabeu et al. 2021).

The issue of addressing population structure and other confounding effects in genomic prediction using machine learning has remained unresolved. Initially, during feature selection, to choose an appropriate SNP-set as input to the model, one can make use of linear covariates to try and limit the bias introduced by nuisance variables. This was seen to generally decrease the performance of the resultant models, suggesting that some confounding is being accounted for by this process. However, it is entirely possible that the linear procedure to remove confounding effects during feature selection is not enough to prevent machine learning algorithms (especially non-linear models) from exploiting unwanted information to aid in improving prediction. This is further discussed in Chapter 4 and some methods to detect and address this problem are explored.

A limitation of many machine learning methods is that even with valuable predictive performance, this will not necessarily translate to insights into disease etiology, as the models can be "black boxes" in this regard. For ALS, whose pathogenesis is not yet fully understood, this is unfortunate, although the best SNP-set size and amenability to non-linear modeling may give some insight into the genetic architecture of a trait. If further work can be done to improve the interpretability of machine learning models this may one day lead to a greater understanding of disease pathogenesis.

Although the predictive performance of these models did not approach utility, there remains some debate around whether or not black box models such as those developed from machine learning frameworks can be implemented clinically if the details of what the risk prediction is detecting are not understood. Patients may be uncomfortable with the lack of transparency and clinicians may not be certain that the model is free from various sources of bias.

There also remains debate in the field of complex trait genetics as to the actual usefulness of generating and reporting polygenic scores developed for diseases and disorders for which

there are largely no preventative measures one can undertake or effective prophylactic treatments, as is the case with ALS.

Overall, with this dataset, machine learning methods were not able to facilitate high discrimination between ALS cases and controls using polymorphic markers, however, there was some improvement seen over the baseline PRS using random forests. Discrimination may improve with larger and denser datasets and machine learning algorithms may in future be able to exploit non-additive interactions and the complexity that lies behind the genetic architecture of this devastating disease.

Chapter 4

Bias Mitigation in Deep Learning Approaches to Genomic Prediction

4.1 Introduction

The results from the previous two chapters demonstrate that machine learning techniques are competitive with, if not potentially superior to, traditional genomic prediction techniques. It is likely that interest in their use will remain high over the next decade as computational resources increase and the user-friendliness of their implementation improves. The focus of this chapter is on how best to account for the confounding effects encountered in large-scale genetic studies during the development of these machine learning models. Specifically, this chapter focuses on the evaluation of several bias-mitigating approaches to neural network training that can result in prediction tools that have minimal confounding. Below, some of the traditional methods of accounting for sources of confounding, such as population structure, are revisited and other available techniques are introduced.

4.1.1 PCA

As described in Section 1.3.1, one of the most common methods for accounting for the confounding effects of population structure is through the use of a dimensionality-reduction technique called principal component analysis. The method of including the top PCs in a GWAS or PRS is currently best practice, although there remains some debate with regard to how best to choose the exact number of PCs included, and how to evaluate the overall performance of the method in reducing confounding from population structure (Abegaz et al. 2019; Peloso and Lunetta 2011). One can make use of a scree plot, which plots the eigenvalues (which represent the relative variance explained by each component) against the relevant component numbers, to visually identify an inflection point of variance explained by the PCs (see Figure 4.1). Another approach is to include only those PCs that collectively explain a pre-determined amount of variance e.g. 90% of variation. Additionally, statistical tests can be utilized, such as the Tracy-Widom test, to assess the significance of the eigenvalues. Some of these tests are implemented in common software packages.

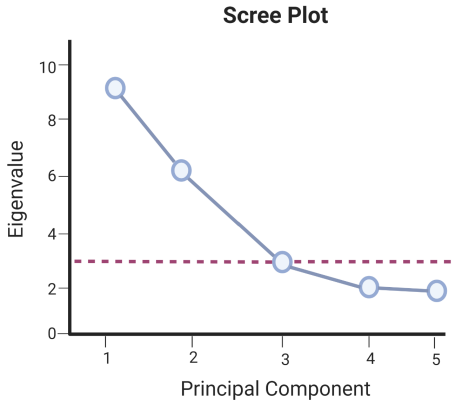


Fig. 4.1 **Scree Plot Example**[2]**:** In order to choose the most informative PCs one can plot the eigenvalues against the ordered principal components. The dashed line represents the elbow inflection point which would be a reasonable stopping point for PC inclusion. In this example, with three principal components one can account for a large portion of the overall variation.

As covariates in linear regression, PCs are a useful tool to control confounding. However, when moving out of a linear setting such as in machine learning, consensus on best practice with regard to how to use such covariates, and how to interpret their performance remains lacking. For many ML methods, the incorporation of covariates is not simple, although some novel techniques have been developed to address this known issue, especially as more ML models move into real-world medical application (Parikh et al. 2019; Vokinger et al. 2021).

4.1.2 Data Collection

A critical stage in ensuring a model will not be biased is to ensure that data collection is conducted appropriately and is representative of the intended target sample. Heavily labeled data can be difficult to obtain and all the relative confounding factors challenging to identify in advance. However, it is important to be aware of the possible introduction of bias at this stage even if it cannot be immediately addressed by changes in the collection protocol. In human disease data, common characteristics that must be balanced between cases and controls could include sex, age, ethnicity, environmental exposures (e.g. smoking status, occupation), and socioeconomic status. For genetic data, genotyping platform is also a known confounder in many analyses.

4.1.3 Two-Stage Regression

One method that has been employed to correct for confounding in human genomic prediction studies is the two-stage regression approach (also known as "regression on the residuals") (Bellot et al. 2018). This involves first regressing the phenotype against known covariates; then taking the residuals from this regression as an "adjusted" phenotype with which to use as input to the genetic association. The assumption is that the adjusted phenotype is now free from confounding and so the prediction model built from it will also be bias-free. The appeal of this method in the context of machine learning models is that the phenotypes can

be adjusted before learning commences, which would circumvent the problem of including covariates during model training.

Even in a standard linear regression (a one-stage approach), however, the two-stage approach is known to be inferior to including covariates simultaneously during β estimation and has received specific criticism for its use in genetic association (Che et al. 2012; Demissie and Cupples 2011; Freckleton 2002). Furthermore, to the extent that it is appropriate to treat the adjusted phenotype as confounder-free; it is not clear that such appropriateness would hold in reducing bias from a *non-linear* model making use of such input (Dinga et al. 2020). The risk that confounding effects may affect a non-linear model's prediction is not eliminated. Therefore, in this thesis, the two-stage regression approach was not used as a bias-mitigating strategy for machine learning approaches to genomic prediction and alternatives were explored.

4.1.4 Bias Mitigation Techniques for Deep Learning

Neural networks can be particularly complex in their construction and operation, with each successive hidden layer potentially extracting higher-order features, and so care must be taken to ensure that the final prediction is free from bias. As the interpretability of their input-output transformations is typically poor, it can be challenging to identify the form such complex biases may take in the model. Therefore, it is important to be proactive in reducing the capacity for known confounding factors to play a role in prediction generation. Unfortunately, there is no gold-standard approach to achieve this for deep learning tasks, although a number of different methods have been proposed.

Adversarial Networks

The dual-goal task of bias minimization and performance maximization for prediction is not a simple one. However, the use of domain-adversarial neural networks (DANNs) may

be a suitable approach. These networks were first introduced by Ganin et al. (2016) and their general architecture is depicted in Figure 4.2. Briefly, a DANN is given two separate learning tasks, with two losses being simultaneously calculated and pitted against one another [22]. While the goal of the classifier[23] is to minimize the loss in order to improve predictive performance, the loss from the adversary is multiplied by a negative lambda factor to create a competitive learning environment within the network (see Equation 4.1). Therefore, the ancillary goal of the network, in addition to classification accuracy, is to maximize the prediction error between the extracted features and the secondary output labels. A tuning parameter λ can be used to set the relative importance of the tasks. These networks are useful when there are multiple competing tasks to be accomplished, however, they are notoriously difficult to train (Kasieczka and Shih 2020; Maekawa et al. 2021). The general instability often seen during training is due to the adversarial nature of the objectives with possibly conflicting goals competing against one another.

$$\mathscr{L}_{Total} = \mathscr{L}_{Classifier} - \lambda \mathscr{L}_{Adversary} \tag{4.1}$$

[22] DANNs were developed by building on the adversarial network work of Goodfellow et al. (2014), and as a method of *domain adaptation* where a model trained on labelled data generalizes well despite shifts in data probability distributions in the test data. Shifts in distributions are not exclusively due to confounding from nuisance variables such as age and sex but could also be due to data collection differences, background time conditions of data etc. (Farahani et al. 2021). The Adeli et al. (2021) approach tackles domain adaptation specifically in the context of confounding and introduces relevant changes and additions to the original DANN framework.

[23] For the sake of simplicity, the primary component of a DANN (i.e. that which outputs the phenotype prediction) is referred to as the "classifier" in this introduction. This is the term generally seen in the literature, although neural networks can of course output continuous values and act as a "regressor" instead of a classifier. Both use cases are explored in this chapter.

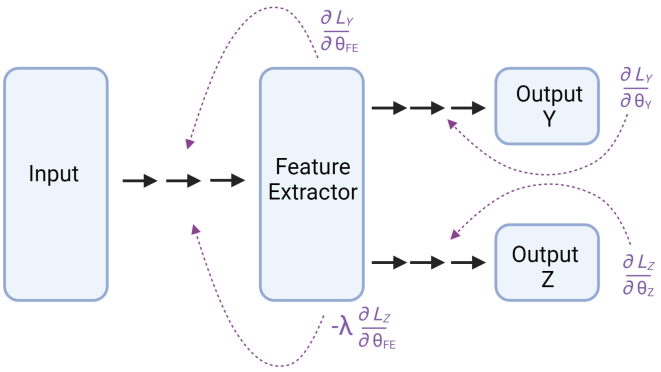


Fig. 4.2 **Domain-Adversarial Neural Network[2]:** A feature extractor (FE) first learns representation features from the input. This feature extractor is used for two separate tasks; prediction of Y and prediction of Z. In the context of this thesis, Y may represent the phenotype and Z may represent a confounding variable(s). Unbroken arrows represent the network's hidden layers. The dashed lines represent back-propagation of the change (δ) in loss ($\mathscr{L}$) with respect to change in the network weights (θ) . The λ value sets the relative importance of the Z prediction task relative to Y. The negative sign of λ creates the adversarial component of the network.

BR-Net

Adeli et al. (2021) adapt this DANN framework to the task of mitigating bias by known continuous confounders, such as genomic PCs, with their bias-resilient neural network scheme (BR-Net). They also introduce a novel adversarial loss function whose minimization encourages statistical mean independence with regard to the predictions and the confounding variables. This property entails the predictions being both linearly and non-linearly independent from the dependent variable. Statistical independence can be measured by calculating the distance correlation (DC) between the target variable(s) and the confounding variable(s). Distance correlation is akin to Pearson's correlation ρ but in multi-dimensional and non-linear space (Edelmann et al. 2019; Székely et al. 2007). A DC of zero between the predictions of a model and the confounding variables would imply statistical mean independence between

them. Naturally, this property is attractive for a machine learning framework that may introduce complex non-linearities in high-dimensional feature space (Hou et al. 2022)[24].

DisCo-Regularized Network

Kasieczka and Shih (2020) also propose the use of distance correlation during model training except as a regularization term rather than as part of a dedicated adversarial network. Here, the classifier performance is regularized by the addition of the DC value[25] between the predictions and the confounding factors to the overall loss (see Equation 4.2). Again, a λ factor controls the strength of penalization.

$$\mathscr{L}_{Total} = \mathscr{L}_{Classifier} - \lambda DC_{(\hat{Y},Z)} \tag{4.2}$$

Pivotal Adversarial Neural Network

The pivotal adversarial neural network (PANN) used by Shimmin et al. (2017), based on the concept of Louppe et al. (2017), modifies the traditional structure of a DANN. Instead of bifurcating the dual-task network, the output of the classifier is used as the starting input of the secondary network (see Fig. 4.3). Learned features from this input are used to predict the confounding factors and the loss fed back to both feature extractors, adversarially in the case of the classifier. The name of the network refers to the ideal property of a classifier, whereby it is pivotal to (i.e. independent of) nuisance parameters.

[24]Precedent for the use of distance correlation in biological applications also includes Chiu et al. (2018), Norman et al. (2019), Zhu et al. (2021) and Omberg et al. (2022).

[25]The DC loss in a TensorFlow compatible form is available from: https://github.com/gkasieczka/DisCo

Pre-training

Aside from the magnitude of the λ parameter, another method that can be used to encourage bias removal is to pre-train using only the adversarial component of the network i.e. giving a head-start by refraining from weight updating based on the classifier loss information for a number of epochs. This might allow the network to be primed with features already independent from confounding factors. This is no guarantee that the network will not go on to learn biased features, nevertheless, it has been shown to be a viable strategy (Shimmin et al. 2017).

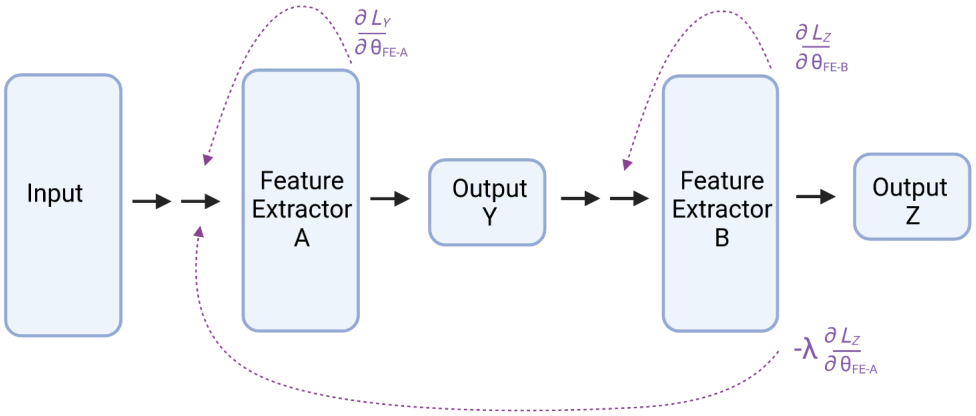


Fig. 4.3 **Pivotal Adversarial Neural Network[2]:** The pivotal network follows a similar procedure to the general DANN seen in Fig. 4.2 except that the output of the classifier is used as the input to the secondary network. The loss from the adversary is back-propagated in an adversarial fashion to the primary classification network's layers. Unbroken arrows represent the network's hidden layers. The dashed lines represent back-propagation of the change (δ) in loss ($\mathscr{L}$) with respect to change in the network weights (θ).

4.2 Materials and Methods

4.2.1 Approach

The goal of this chapter is to assess the effectiveness of various novel methods for confounder handling in neural network approaches to genomic prediction. To this end, the approach

taken is to optimize the proposed methods with respect to prediction performance as well as minimizing bias, while comparing against a baseline feed-forward neural network. Bias is measured as the distance correlation between the final target predictions of each network and the corresponding confounding variables.

4.2.2 Data

Arabidopsis thaliana

The *Arabidopsis thaliana* data used in this chapter and previously described in Section 2.2.1 has no confounding environmental variables to account for as individuals were grown under standardized techniques.

Genomic principal components were calculated on this data using GCTA software version 1.93.2 (Yang et al. 2011). The GRMs for this analysis were made using a set of 1,049,629 independent SNPs created from PLINK's LD-based pruner. The pruning used a 50kb window, 5bp step-size and a ρ^2 threshold of 0.2 (Purcell et al. 2007). The first two principal components of variation are plotted against one another in Fig. A.1b. The flowering time (16°C) was chosen as the phenotype of interest in this work. The top 10,000 SNPs from the GWAS described in Section 2.2.4 were used as the input variables to all networks with the homozygous genotypes coded as binary variables.

ALS

The ALS data used in this chapter and previously described in Section 3.2.1 has limited demographic information on cases and controls although individuals had their sex recorded and were divided into 27 separate strata on the basis of nationality and genotyping platform (Van Rheenen et al. 2016). Nationality here is likely a good proxy for ethnicity/ancestry as individuals not of European ancestries were excluded from the analysis (see Section 1.3). Furthermore, PCs from the genetic data of individuals can be calculated for more fine-grained

ancestry approximation. These were calculated using a set of 131,883 independent SNPs created from PLINK's LD-based pruner. The pruning used a 50kb window, 5bp step-size and a ρ^2 threshold of 0.2. A scree plot was then created from the resultant PCs calculated by PLINK. The top 1,000 SNPs from the GWAS described in Section 3.2.3 were used as the input variables to all networks, with genotypes coded as 0, 0.5, or 1.

4.2.3 Network Descriptions

Neural networks were implemented using TensorFlow's Python API (Abadi et al. 2015). Binary cross-entropy (logistic loss) or MAE was used to calculate the phenotypic output loss for the ALS and *Arabidopsis* data respectively. Pearson's ρ or AUC was used as the phenotypic performance measure. Unless otherwise stated, binary cross-entropy, categorical cross-entropy, or MAE was used for binary, categorical, or continuous confounder losses respectively.

All networks made use of the Adam optimizer (Kingma and Ba 2014). Early stopping of the networks was implemented after four 10-epoch intervals of less than 1% gain in performance. Genotypes were loaded in as dosages (0, 0.5, or 1) along with corresponding phenotypes. All continuous phenotype and confounder values were standardized and scaled based on the training set prior to learning. The BR-Net, PANN, and DisCo-regularized networks were permitted to pre-train using only the adversarial component of the loss during hyperparameter tuning. Learning rate, network depth, λ strength, pre-training epochs, and training epochs were optimized by inspection.

BR-Net

The correlation coefficient loss described in Adeli et al. (2021) was implemented for the adversary. In the case of multiple independent confounders, the DC-based loss was implemented instead. Refer to Figure 4.2 for a representation of the training procedure. For

each epoch, the phenotypic (Y) loss was first back-propagated to update the Y output layer's weights as well as the feature extractor. The adversary's loss was then back-propagated to update the bias (Z) predictor weights. Finally, the adversary's loss was maximized and back-propagated to update the feature extractor's weights.

DisCo-Regularized Networks

This approach follows a similar learning procedure as a standard feed-forward network except the total loss back-propagated to the weights was calculated as shown in Equation 4.2.

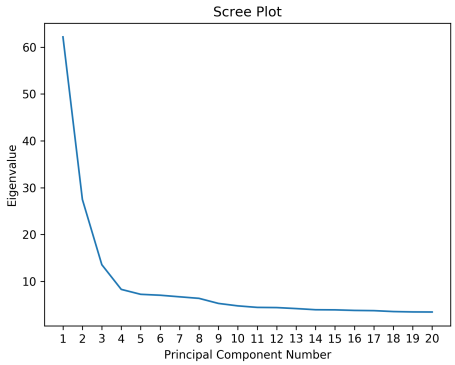
PANNs

The correlation coefficient loss described in Adeli et al. (2021) was investigated for the adversary in addition to the standard binary cross-entropy, categorical cross-entropy, or MAE for binary, categorical and continuous confounder losses respectively. In the case of multiple independent confounders, the DC-based loss was implemented instead. Refer to Figure 4.3 for a representation of the training procedure. For each epoch, the Y and Z loss are used to update the first feature extractor's weights. The loss was then used to update the second feature extractor's weights.

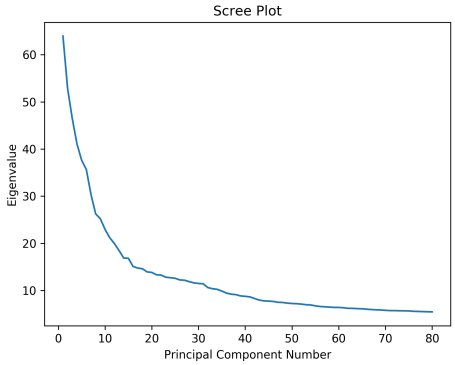
4.2.4 Assessment

Training, test, and validation splits were conducted using a nested cross-validation design (see Section 1.4.6). Distance correlation between predictions and the nuisance variables was measured using the dcor Python package (Ramos-Carreño 2022). In order to select the final models of interest after hyperparameter tuning, models were initialized repeatedly and the top-performing models were chosen based on their performance in terms of prediction and DC minimization.

4.3 Results



(a) **ALS Case/Control Data:** An inflection point in explained variance can be observed at $\sim$ 3 principal components.



(b) ***Arabidopsis thaliana*:** An inflection point in explained variance can be observed at $\sim$ 15 principal components.

Fig. 4.4 **Scree Plots for ALS and *Arabidopsis Thaliana* Data**[16].

4.3.1 *Arabidopsis thaliana*

The experiments were run using the first principal component of variation, as well as the full 15 PCs deemed relevant for inclusion as estimated by a scree plot (see Fig. 4.4b). As a baseline with which to compare, the results from a non-adversarial feed-forward neural network can be seen in Fig. 4.5. As learning progresses, the distance correlation between the predictions and the nuisance variable increases. This occurs both in the training and in the test sets. In general, to achieve a test-set phenotypic prediction ρ of $\sim$ 0.6, the distance correlation grows to around 0.12. Although there can be differences in magnitude between the train DC and the test DC, there is a tight correlation during training. A Pearson correlation of $\sim$ 0.87 was measured between the final train DC values and the hold-out validation DC (calculated using $\sim$ 500 different runs, across all four model architectures). Without relying on incorporating new methods, one can choose between many iterations of an FNN and choose the model that has a relatively low DC, an example of such is given in Fig. 4.7a ($\rho \approx 0.6$, DC ≈ 0.04).

To demonstrate the feasibility of the dual-task adversarial approach one can observe that, when using an overly large λ strength value, learning does not commence at all (see Fig. 4.6a). In this example, a DisCo network simply prioritizes the minimization of the distance correlation between the output predictions and the first principal component of variation. Likewise, when using a large negative λ value (changing the task to maximization of the distance correlation), the distance correlation becomes extremely high without any concomitant growth in predictive performance on the phenotype (see Fig. 4.6b). However, the ultimate goal of the experiment is to demonstrate that there exists a λ value that optimally maximizes predictive performance while also forcing the DC to be relatively low. Ideally, the DC between the predictions and the confounding variables(s) would be zero as this would imply statistical mean independence between them.

As was expected, a trade-off often exists between predictive performance and bias reduction using the adversarial networks. Setting the λ too high does not allow for high predictive performance yet after hyperparameter tuning, there were λ values that appeared to be effective in keeping prediction high while keeping DC low. This was particularly seen in the case of DisCo networks (see Fig. 4.7c). Depending on how strictly one might set that maximum acceptable DC threshold, the top-performing DisCo network was able to achieve a ρ of ~ 0.4 with a distance correlation of close to zero.

To a lesser extent, PANNs were also able to achieve a noticeable decrease in the overall DC needed to achieve high prediction accuracy (see Fig. 4.7b). However, even after λ tuning and pre-training, BR-Nets were not able to substantially achieve lower DC than the best baseline FNN models (see Figs. 4.7a and 4.7d).

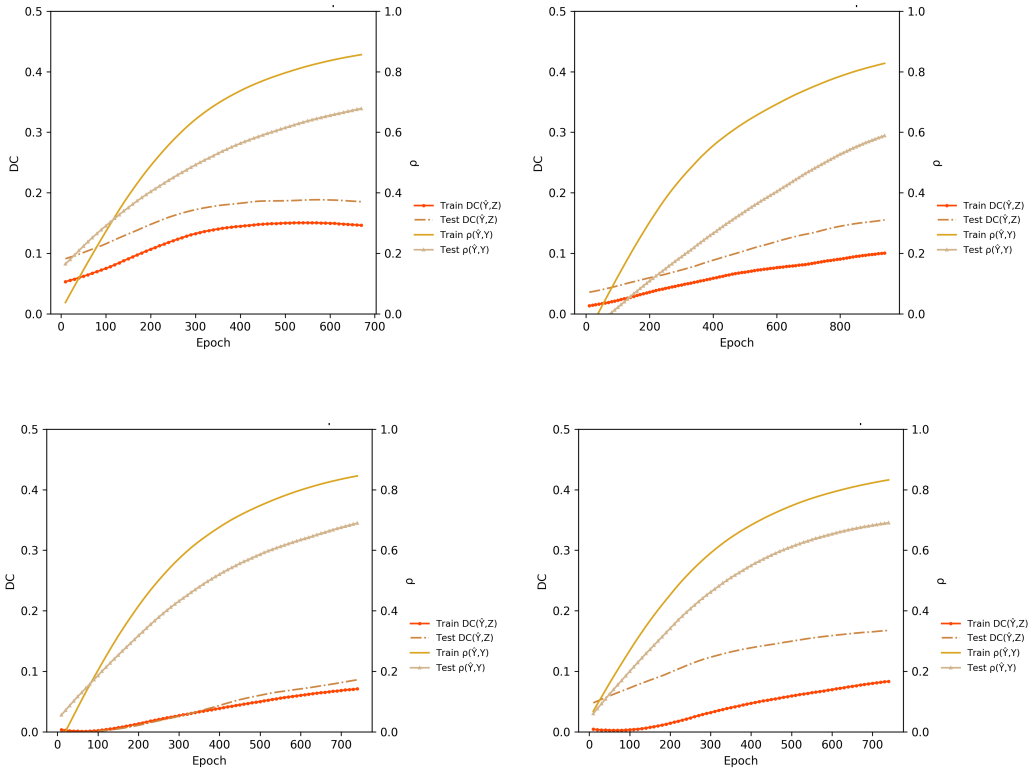
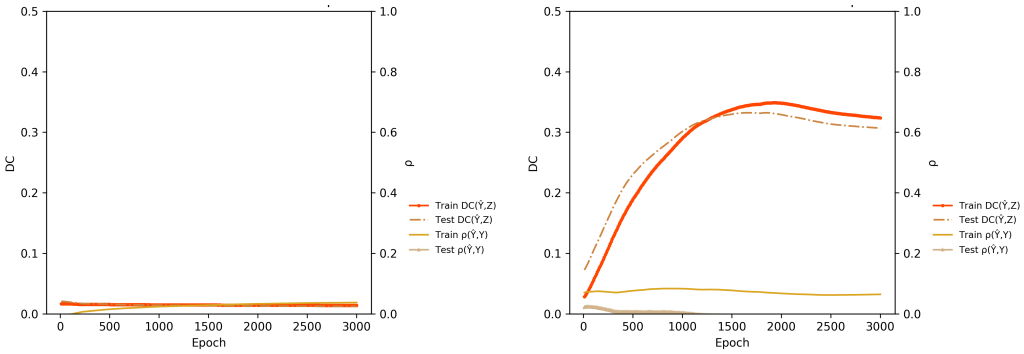
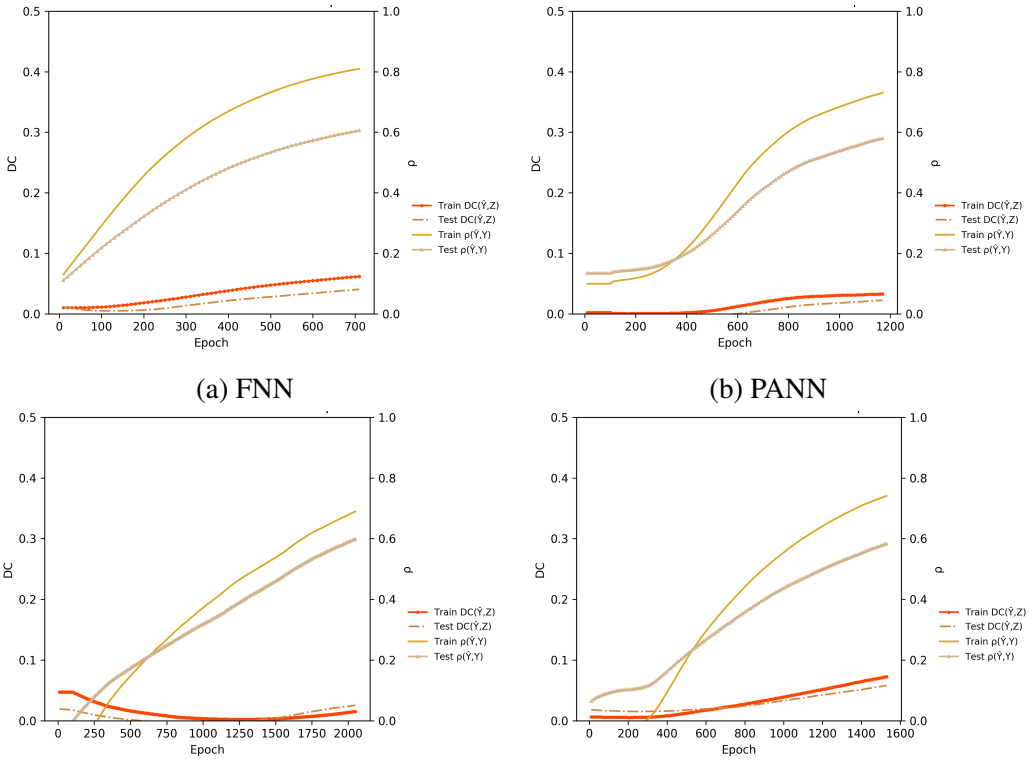


Fig. 4.5 **Representative Results Using 1 PC (*Arabidopsis* Flowering Time Trait)**[16]. The predictive performance (ρ) of the training and test sets is displayed on the right-hand y-axis. The left-hand y-axis measures the distance correlation (DC) between outputted predictions and the confounder variable(s). Ideally, a genomic prediction model should have high ρ and low DC.



(a) λ=5 (b) λ=-5

Fig. 4.6 **Representative Results Using Sub-Optimal λ Values for 1 PC Using a DisCo Network (*Arabidopsis* Flowering Time).**[16]. The predictive performance (ρ) of the training and test sets is displayed on the right-hand y-axis. The left-hand y-axis measures the distance correlation (DC) between outputted predictions and the confounder variable(s). Ideally, a genomic prediction model should have high ρ and low DC.



(a) FNN (b) PANN

(c) DisCo (d) BR-Net

Fig. 4.7 **Best Results Using 1 PC (*Arabidopsis* Flowering Time Trait)**[16]. The predictive performance (ρ) of the training and test sets is displayed on the right-hand y-axis. The left-hand y-axis measures the distance correlation (DC) between outputted predictions and the confounder variable(s). Ideally, a genomic prediction model should have high ρ and low DC.

In general, pre-training of the adversarial networks on only the confounding variables was found to be an effective method with which to encourage low bias during subsequent training including the phenotypic loss.

When using the full 15 PCs as confounder variables, the average DC of the most predictive models increased, which was to be expected. Representative examples of the best-performing baseline FNN models (selecting those with low DC and high ρ) are shown in Fig 4.8. As before, there was a close correlation between the train DC and the test DC. Across all model architectures, using data from 120 iterations, the correlation between the training DC and the hold-out validation DC was 0.93.

Unlike as when using a single nuisance variable, the PANN networks struggled to keep the DC any lower than the baseline model when achieving high prediction accuracy (see Fig. 4.9). As can be seen in Fig. 4.9c, the DC could be kept low for a certain amount of learning, but in order to achieve high prediction accuracies the DC values needed to increase in tandem with ρ.

As before, the BR-Net class of models failed to distinguish themselves from the general performance of the best baseline models (see Fig. 4.11). Using a strong λ parameter, DC can be kept low but this puts a ceiling of ~ 0.2 on the predictive performance (see Fig. 4.11d). However, in Fig. 4.8c, it can be observed that the baseline model can also achieve this result at an early training stage.

Finally, the results from the DisCo networks (see Fig. 4.10) were slightly more promising, although they failed to recapitulate the impressive results from Fig. 4.7c. As can be seen in Fig. 4.10d, a predictive ρ of ~ 0.5 can be achieved while keeping the DC below a threshold of 0.05 (compared to a ρ of around 0.4 for Fig. 4.8c). An inflection point often existed using the DisCo method, beyond which higher prediction accuracy could not be achieved without dramatically increasing the distance correlation between predictions and the confounding variables.

4.3.2 ALS

The baseline feed-forward network results using the top three principal components (based on the scree plot in Fig. 4.4a) as confounding variables are shown in Fig. 4.12. As with the *Arabidopsis* data, the distance correlation between the predictions and the bias variables grows steadily as training progresses. The correlation between the final training DC value and the validation DC was found to be extremely high at 0.99.

When using some of the bias-mitigating techniques, it can be seen that there is a clear reduction in overall DC when using both PANNs (see Fig. 4.13) and DisCo networks (see

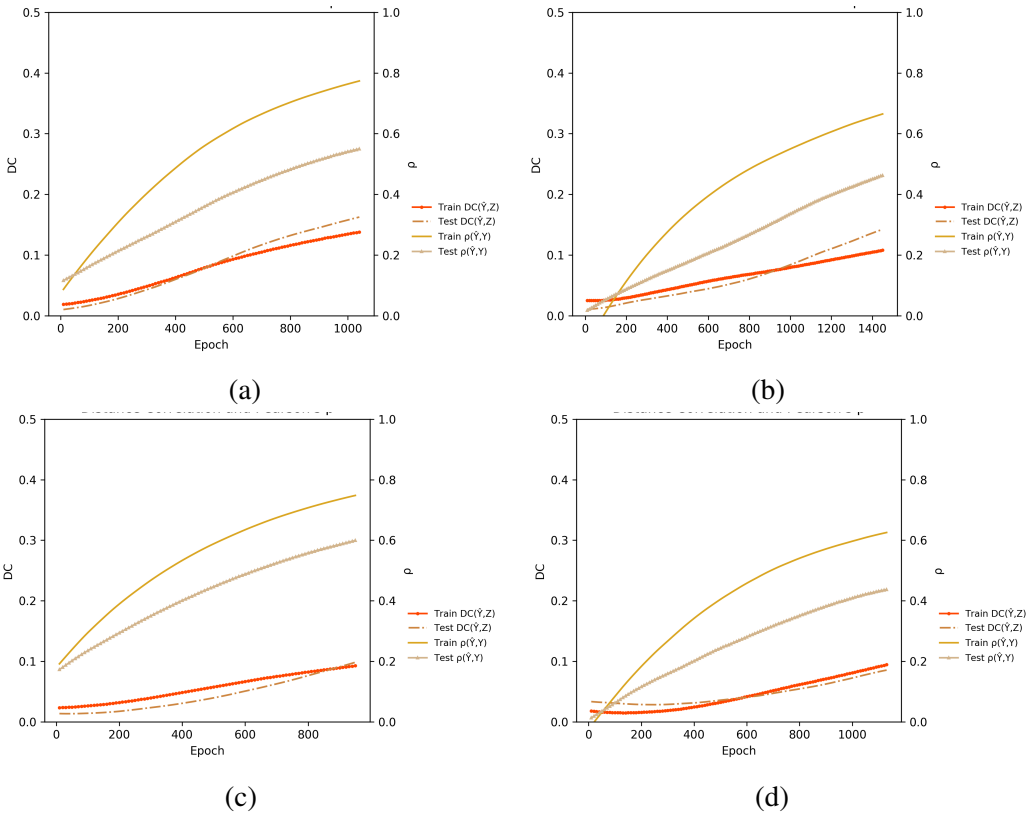


Fig. 4.8 **Best FNN Results Using 15 PCs (*Arabidopsis* Flowering Time Trait)**[16]. The predictive performance (ρ) of the training and test sets is displayed on the right-hand y-axis. The left-hand y-axis measures the distance correlation (DC) between outputted predictions and the confounder variable(s). Ideally, a genomic prediction model should have high ρ and low DC.

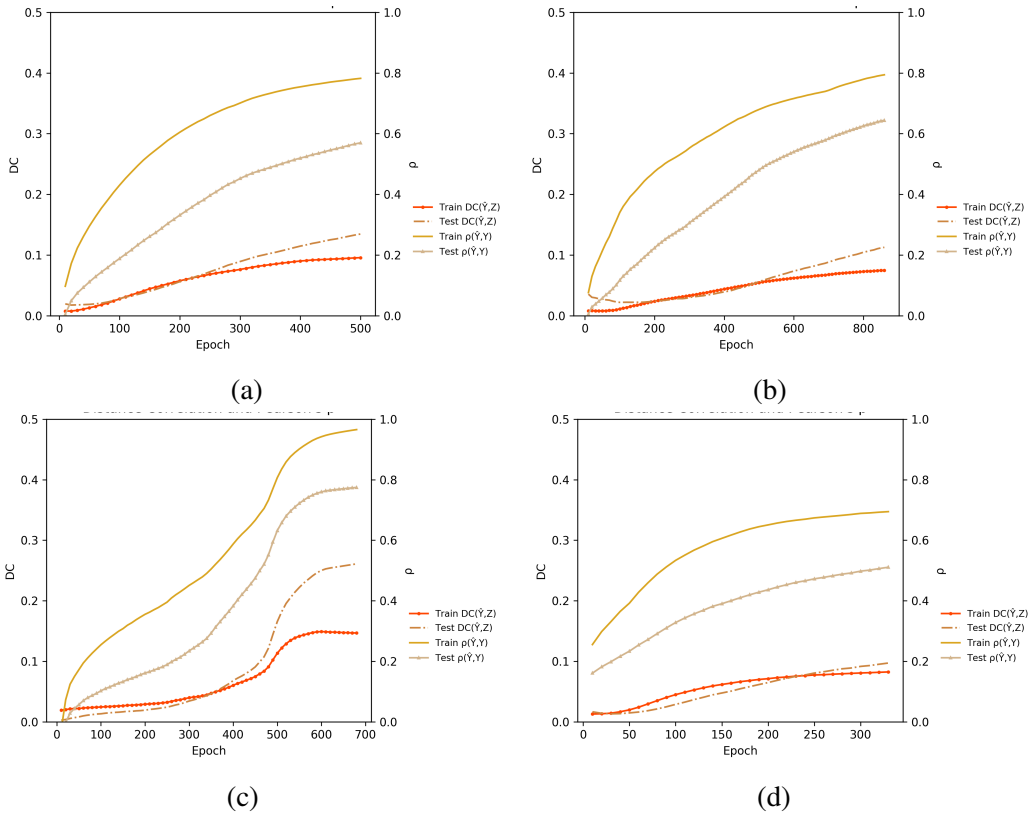


Fig. 4.9 **Best PANN Results Using 15 PCs (*Arabidopsis* Flowering Time Trait)**[16]. The predictive performance (ρ) of the training and test sets is displayed on the right-hand y-axis. The left-hand y-axis measures the distance correlation (DC) between outputted predictions and the confounder variable(s). Ideally, a genomic prediction model should have high ρ and low DC.

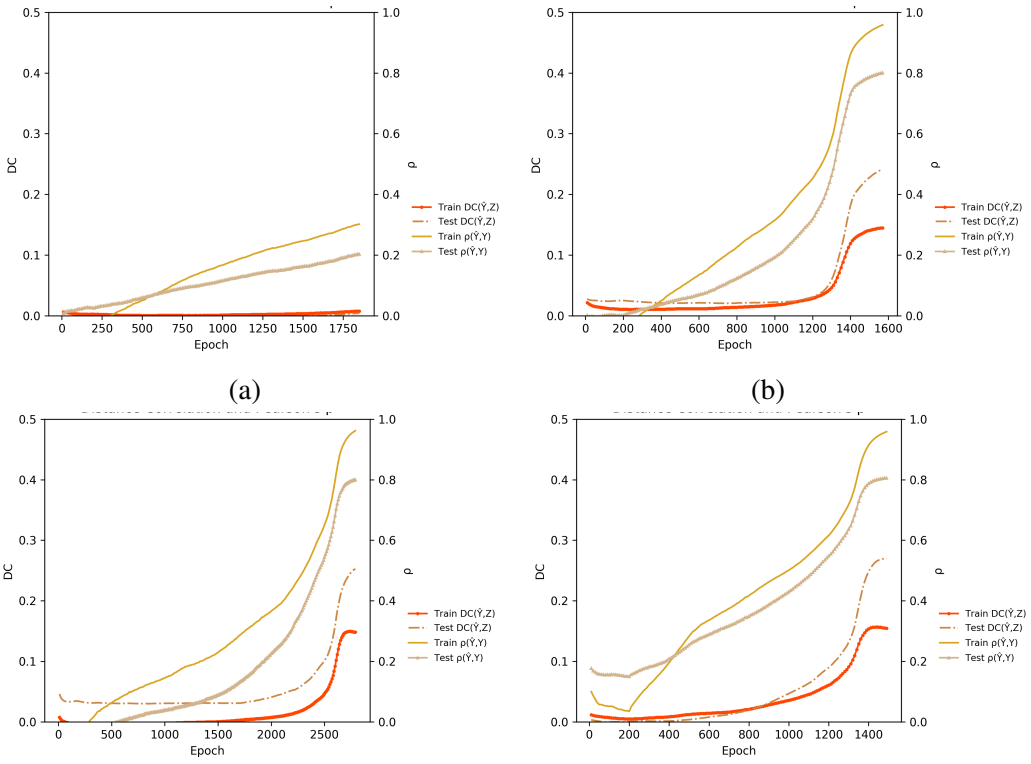


Fig. 4.10 **Best DisCo Results Using 15 PCs (*Arabidopsis* Flowering Time Trait)**[16]. The predictive performance (ρ) of the training and test sets is displayed on the right-hand y-axis. The left-hand y-axis measures the distance correlation (DC) between outputted predictions and the confounder variable(s). Ideally, a genomic prediction model should have high ρ and low DC.

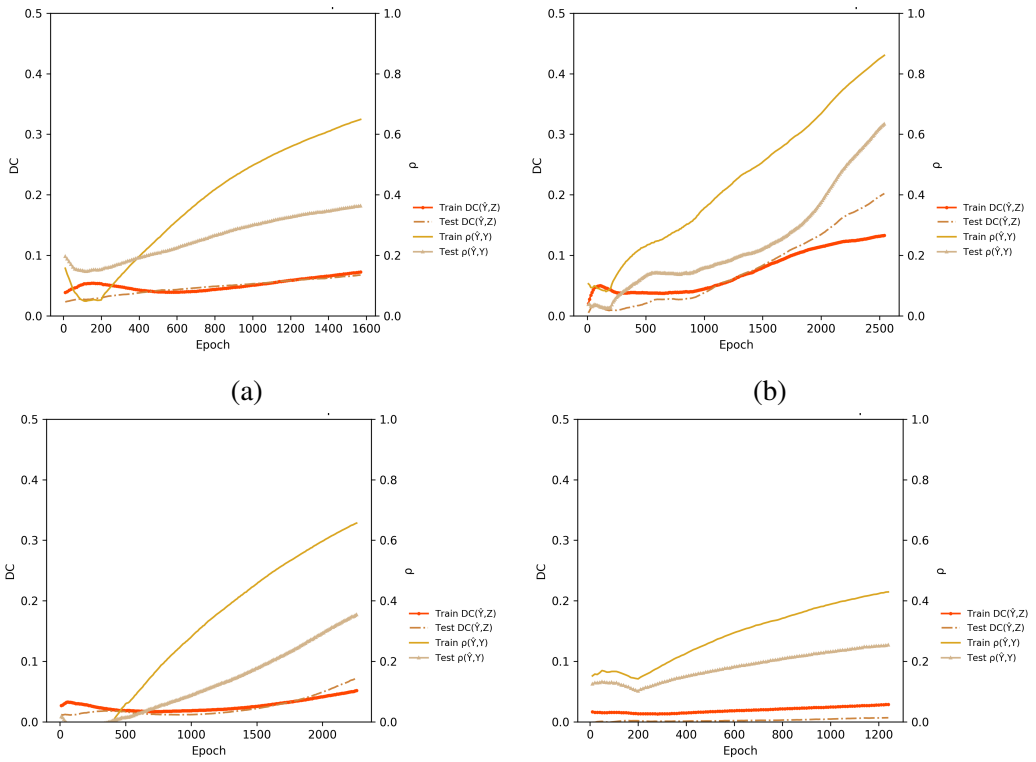


Fig. 4.11 **Best BR-Net Results Using 15 PCs (*Arabidopsis* Flowering Time Trait)**[16]. The predictive performance (ρ) of the training and test sets is displayed on the right-hand y-axis. The left-hand y-axis measures the distance correlation (DC) between outputted predictions and the confounder variable(s). Ideally, a genomic prediction model should have high ρ and low DC.

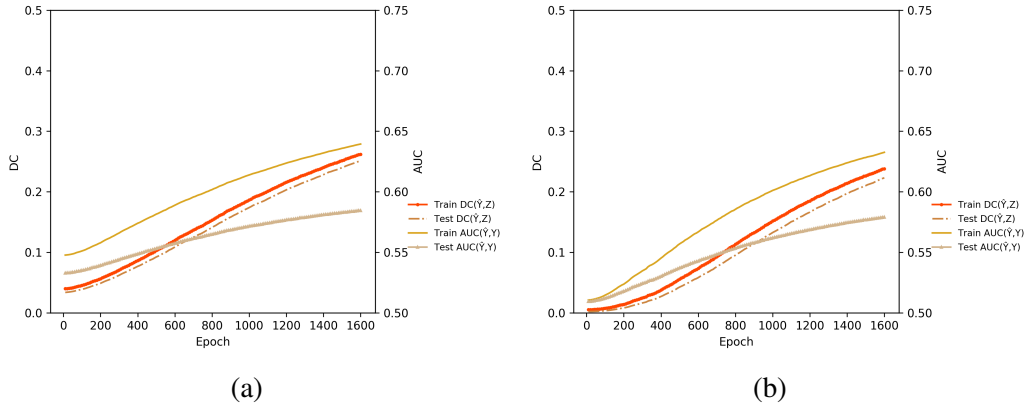
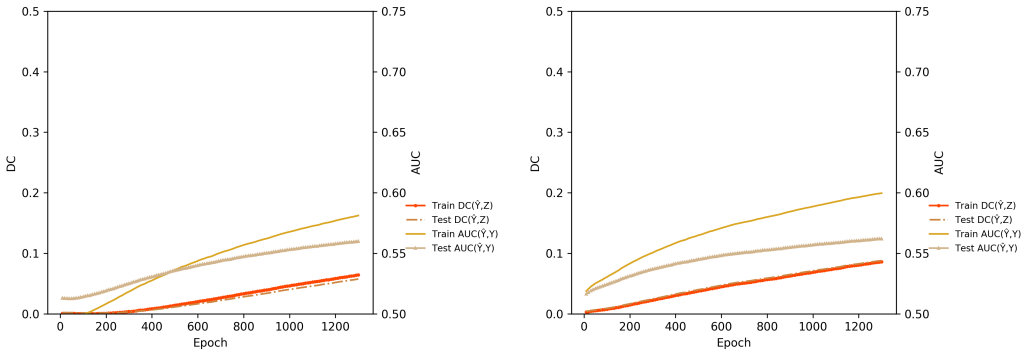
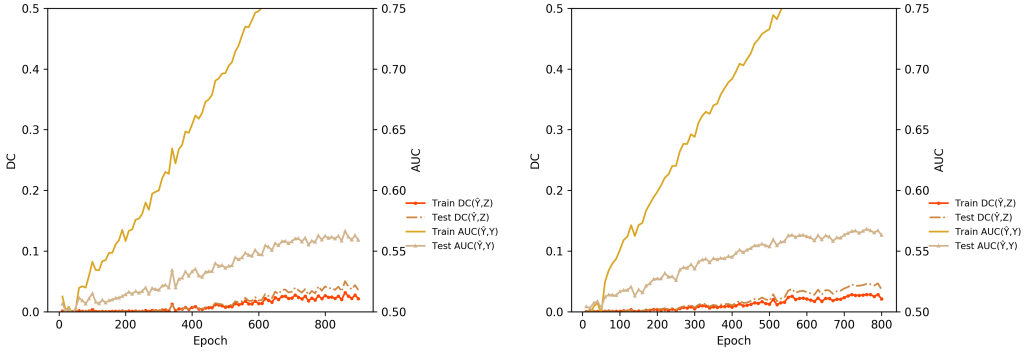


Fig. 4.12 **Best FNN Results Using 3 PCs** (ALS)[16]. The predictive performance (ρ) of the training and test sets is displayed on the right-hand y-axis. The left-hand y-axis measures the distance correlation (DC) between outputted predictions and the confounder variable(s). Ideally, a genomic prediction model should have high ρ and low DC.



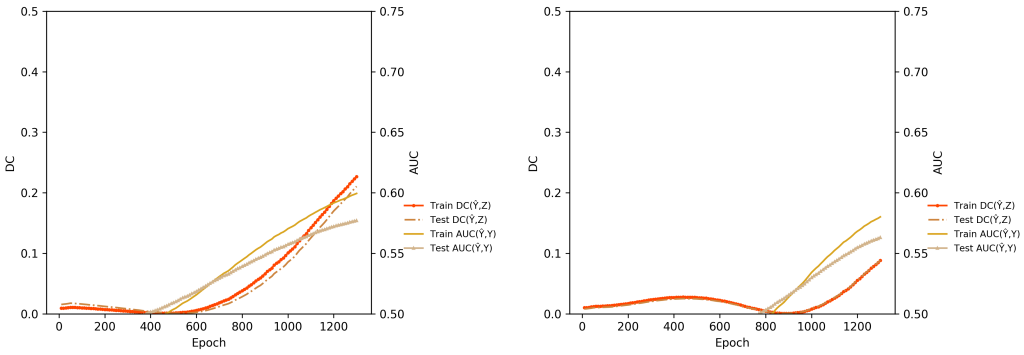
(a) (b)

Fig. 4.13 **Best PANN Results Using 3 PCs** (ALS)[16]. The predictive performance (ρ) of the training and test sets is displayed on the right-hand y-axis. The left-hand y-axis measures the distance correlation (DC) between outputted predictions and the confounder variable(s). Ideally, a genomic prediction model should have high ρ and low DC.



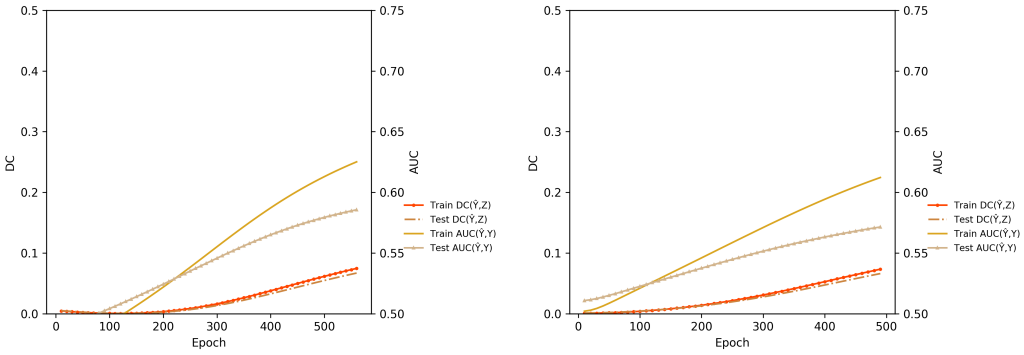
(a) (b)

Fig. 4.14 **Best DisCo Results using 3 PCs** (ALS)[16]. The predictive performance (ρ) of the training and test sets is displayed on the right-hand y-axis. The left-hand y-axis measures the distance correlation (DC) between outputted predictions and the confounder variable(s). Ideally, a genomic prediction model should have high ρ and low DC.



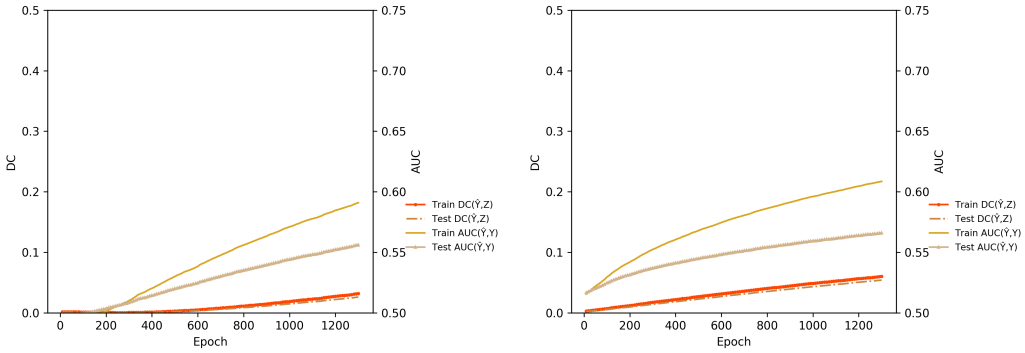
(a) (b)

Fig. 4.15 **Best BR-Net Results using 3 PCs** (ALS)[16]. The predictive performance (ρ) of the training and test sets is displayed on the right-hand y-axis. The left-hand y-axis measures the distance correlation (DC) between outputted predictions and the confounder variable(s). Ideally, a genomic prediction model should have high ρ and low DC.



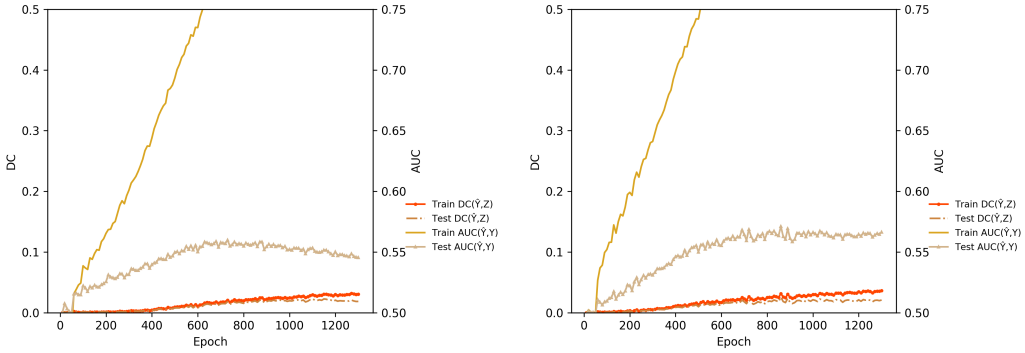
(a) (b)

Fig. 4.16 **Best FNN Results Using Strata Information** (ALS)[16]. The predictive performance (ρ) of the training and test sets is displayed on the right-hand y-axis. The left-hand y-axis measures the distance correlation (DC) between outputted predictions and the confounder variable(s). Ideally, a genomic prediction model should have high ρ and low DC.



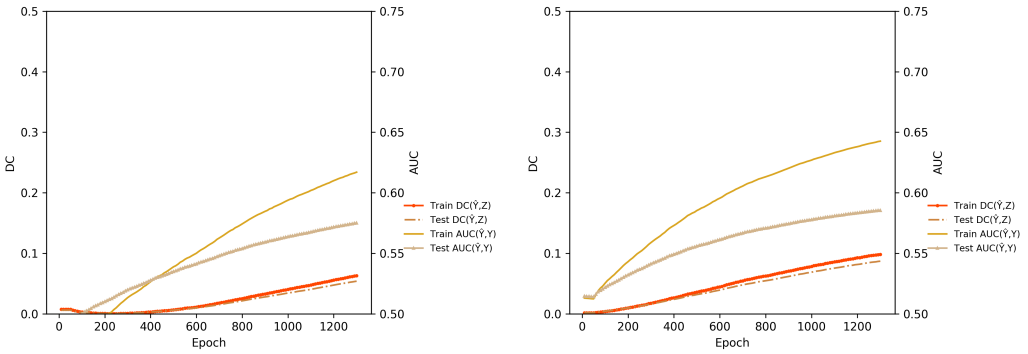
(a) (b)

Fig. 4.17 **Best PANN Results Using Strata Information** (ALS)[16]. The predictive performance (ρ) of the training and test sets is displayed on the right-hand y-axis. The left-hand y-axis measures the distance correlation (DC) between outputted predictions and the confounder variable(s). Ideally, a genomic prediction model should have high ρ and low DC.



(a) (b)

Fig. 4.18 **Best DisCo Results Using Strata Information** (ALS)[16]. The predictive performance (ρ) of the training and test sets is displayed on the right-hand y-axis. The left-hand y-axis measures the distance correlation (DC) between outputted predictions and the confounder variable(s). Ideally, a genomic prediction model should have high ρ and low DC.



(a) (b)

Fig. 4.19 **Best BR-Net Results Using Strata Information** (ALS)[16]. The predictive performance (ρ) of the training and test sets is displayed on the right-hand y-axis. The left-hand y-axis measures the distance correlation (DC) between outputted predictions and the confounder variable(s). Ideally, a genomic prediction model should have high ρ and low DC.

Fig. 4.14). However, for both these techniques there was a slightly lower ceiling for the predictive performance of networks.

The BR-Net networks did not greatly improve upon the baseline performance set by the FNN models (see Fig. 4.15), except at very early training stages when the AUC metric is quite low.

When using the strata covariate (which combines information on geographic origin and genotyping platform) the results were less promising. It was not clear that any of the bias-mitigating models (see Fig 4.17, 4.19 and 4.18) outperform the baseline feed-forward model (see Fig. 4.16). As before, the correlation between training DC and validation DC was also very high at 0.98.

4.4 Discussion

The results in this chapter demonstrate the utility of tracking known and measurable biases during deep learning and offer some insight into the feasibility of adjusting network architectures to encourage the minimization of confounding while keeping performance

high. As shown in Fig. 4.7a, in comparison with Fig. 4.5, one does not necessarily need to make use of adversarial approaches or loss adjustment in order to make a measurable reduction in the model's final DC between predictions and confounder variables. The inherent stochasticity of network initialization and learning allows one to choose between various training iterations, some of which may output predictions that have a final DC value that falls above the acceptable threshold. Observing the DC during training also allows for the identification of certain performance inflection points, beyond which performance does not improve without clearly increasing bias as measured by distance correlation.

However, the use of adversarial or loss-adjusted networks may allow a researcher to actively encourage less statistical dependence between confounders and predictions during training. As with most adversarial approaches, this is not an easy task and training may prove to be difficult, although a reduced level of distance correlation was sometimes observed in the series of experiments detailed here; with DisCo networks proving to be the most promising.

The DisCo network is incredibly simple to implement in a deep learning setting as it only involves adding a penalty term to the normal loss metric. The tuning of the λ strength parameter is more difficult, however, it is probably worth investigating if measured confounders are at hand during training. In both the *Arabidopsis* and ALS experiments, BR-Nets seemed to be more limited in their utility and ease of use while PANNs were only sometimes more successful than the baseline approach. However, all three methods did indeed show an ability to prioritize DC minimization, though a difference was observed in how effectively the three architectures could minimize DC while also maximizing predictive performance.

Future work on this project could include the creation of an adaptable λ framework. The inflection points that were observed during training show that the penalty factor loses its ability to keep the DC value low beyond a certain predictive point, however, if the strength

of λ could vary depending on contextual factors during training, this could result in a more successful model.

The DisCo networks using ALS data, although successful in reducing bias relative to the feed-forward networks, had a clear problem with over-fitting (see Figs. 4.14 and 4.18). It is not clear why this should be the case although future, more comprehensive, hyperparameter tuning involving regularization may aid in reducing this effect (and possibly improve predictive accuracy on the test set). This may aid performance across all model types although the problem was particularly pronounced in the DisCo networks.

The ALS analysis was limited by the polygenic predictivity of the dataset. Future work will need to be done on other phenotypes to see if the patterns described here hold more generally.

In both datasets, the measured bias in the training data closely correlated with the bias observed in both the test and validation set, but this is of course not guaranteed in other datasets with different nuisance variables, so it is advisable to ensure a high correlation exists, at least between training and testing sets, before interpreting results from these methods.

Unfortunately, there is no gold-standard approach by which to gauge the performance of these results beyond the calculation of distance correlation. This benchmarking problem is also true of PC-adjustment during standard genomic prediction as debate remains on how best to show that no further bias adjustment is needed. Although distance correlation is a useful tool to measure statistical independence in biological applications, it is beyond the scope of this thesis to prove that the predictions outputted by low DC models are truly unbiased with respect to the confounders. Distance correlation is only one measure of bias and it is unclear how one could irrefutably demonstrate a true lack of bias in a model. Ultimately, as with PC-adjustment and the addition of further principal components, there may be diminishing returns with continual reduction of the maximum DC threshold. This may prove to be task-specific and somewhat heuristic, however, the creation of clear confounding guiding

principles could be an important goal for the field of genomic prediction with machine learning. So far, this issue has remained somewhat under-explored relative to the amount of information on covariate adjustment in linear genomic prediction tasks.

With human data, when one is specifically interested in avoiding predictive power stemming from certain specific confounders (e.g. sex or ethnicity), it is usually imperative that the predictions themselves are known to be independent of those factors. Therefore, metrics that can detect non-independence and encourage mitigation of it are an important contextual addition during the evaluation of the model. A predictor that improves *upon* confounding may be of less interest than a predictor that is known to be truly free *from* confounding, even if the former has more predictive power overall. When using the ALS data one can see a clear general decrease in bias from the baseline (Fig. 4.12) with the DisCo (Fig. 4.14) network, although the overall predictivity remains lower with the bias-mitigating model. It is not fully clear how one should balance prioritizing high predictivity with low observed bias. When dealing with human data there may be a higher emphasis placed on low-bias models for example.

A limitation of this analysis is that as neural networks tend to defy interpretability, it is difficult to recognize when learned features have truly transformed from a biased representation to an unbiased one. As mentioned previously, any DC threshold will largely be arbitrary, but it would be useful to be able to recognize state changes between model representations. For example, in Fig. 4.10a training does not proceed over a predictive ρ of 0.2 due to a strong λ parameter, but it is unclear if this prediction stems from representations that are different from the representations that constitute the ρ of 0.2 achieved by the FNN at an early epoch (in Fig 4.8b for example.).

This chapter dealt solely with the use of confounder adjustment in the case of neural networks but the general problem of biased predictions exists across many non-linear machine learning models. Various other techniques have been proposed, some of which are general

and some of which are model specific, and it is likely that major headway will be made in this area over the next decade as the importance of reducing bias in ML-driven tasks in medicine and genomics is put to the fore.

Chapter 5

General Discussion and Conclusion

This chapter will attempt to synthesize the new results presented in this thesis and place them within the broader context of the field, as well as introduce future methods by which this work can be expanded upon and the directions machine learning approaches to genomic prediction might take. It will end with a brief conclusion as to the results and value of this thesis.

5.1 General Discussion

Chapter 1 introduced the two fields of genomic prediction and quantitative genetics. It highlighted the fact that much of the genetic variation in complex traits estimated to exist across populations has yet to be accounted for, and that the genetic architectures of these traits are not yet fully understood. A potential reason for this knowledge gap is that complex non-linear interactions are not well modeled by current techniques in genomic prediction, which are largely linear in their construction. However, it is known that non-additive interactions can indeed be captured by the additive component of variance usually targeted by researchers, and so the extent to which non-linear modeling holds the key to closing the missing heritability gap remains debated.

Machine learning is an attractive option if one wishes to model non-linear interactions and several ML families are also introduced in Chapter 1. However, the complexity of machine learning has a drawback in that methods that account for confounding, an eternal problem in population-level genomic data, are under-developed and gold-standard approaches are lacking.

Chapter 2 makes use of a global *Arabidopsis thaliana* sample to explore the benefits of applying machine learning techniques to the task of genomic prediction, benchmarking against the standard gBLUP method. Using a nested cross-validation approach, a Dunnett test was conducted on the predictive performances as a statistical comparison between the models and the baseline. Formal statistical tests are often left out of comparisons between methods in genomic prediction and this was considered a strength of this analysis. The experiment found that even after extensive hyperparameter tuning, the standard linear method performed relatively well against the machine learning methods although some families, namely feed-forward neural networks and ridge regression, often outperformed it. Although feed-forward neural networks sometimes improved upon the gBLUP method, it is unclear to what extent this was due to the incorporation of non-linear effects, as Ridge regression is a linear model and also performed well. One of the drawbacks of using machine learning methods is that they are often “black boxes" whose input-output transformations prove difficult to interpret. Much work is being done on improving the interpretability of machine learning models, but it is unclear to what extent new techniques may aid in interpreting genomic prediction outputs. This is one area of future research in the field that could provide useful context for the results presented in this thesis.

Chapter 3 takes a similar methodological approach as Chapter 2, except that it uses a large European sample of ALS cases and controls. Although ALS is known to have a large contribution from rare variation, previous work had estimated that almost 10% of the heritability was due to the effect of common polymorphisms in the population (overall $h^2 \approx$

0.5) (Van Rheenen et al. 2016). Furthermore, a role for epistasis in the genetic architecture of this disorder had also been proposed. Similar to the results in Chapter 2, statistically significant improvement over a standard PRS was not regularly obtained, and the magnitudes of any improvements were generally small. Overall, the random forest class of predictors performed the strongest in this setting. A calibration analysis, which is often omitted in similar studies, found that poor calibration was not an issue for the baseline PRS, or for the top-performing models across each feature set.

The discrimination between cases and controls achieved by all of the ALS models was found to be quite low, even when using thousands of the most strongly associated SNPs, suggesting that in this sample there is not a large polygenic component to trait architecture. Although bigger and more diverse sample sizes are always useful in genomics, a larger GWAS conducted during the course of this research revised downward the estimate of the SNP-based heritability in ALS to only 3% (Van Rheenen et al. 2021). This may limit the current practicality of further insight into ALS genetic architecture using only SNP-data, although advances in ML techniques may allow for further interesting work, even if only on rarer forms of variation.

The rise in the accuracy of polygenic scoring on human traits and disorders offers up as many questions as it does exciting prospects. On the one hand, precision medicine will be greatly facilitated by the growth of cheaply available polygenic risk scores. Persons identified to be at high risk of certain diseases can be selected for early-screening programs, especially important in cancers, or given appropriate access to treatments, such as preventative statins for CAD. Furthermore, treatment-responsive sub-groups may be more easily identified during clinical trials, and the role of pharmacogenomics in medicinal dosing regimens is becoming ever larger (Johnson et al. 2022). However, information on genetic disease risk can often be emotionally burdensome to patients, as it is largely immutable and transferable across generations. There is much debate around the cost-benefit of the routine supply of genomic

risk information of serious disorders, without effective preventative measures or treatments, such as is the case in Alzheimer's disease (Baker and Escott-Price 2020). In the context of ALS, in the absence of involvement in a research program dedicated to early treatment investigation, it is doubtful that a publicly available ALS polygenic test would be of net benefit to the population. However, as mentioned previously, genetic profiling to identify treatment-responsive sub-groups in a clinical trial setting could be one potential use case.

The development of human PRS has also led to an increased interest in the use of polygenic embryo selection. However, there are a number of ethical questions related to this practice, notwithstanding debate regarding the actual utility of such an approach to prevent disease (Karavani et al. 2019; Lencz et al. 2022; Turley et al. 2021).

The first two experimental chapters took an approach that has been common in previous research, which was not to fully address the problem of confounding and population structure when developing machine learning models for genomic prediction. Other than adjusting for population structure at the feature selection stage (through the use of a GRM-based MLMA or using PCs as covariates in GWAS), no formal evaluation of bias was made on the outputted predictions. However, in Chapter 4, an attempt was made to develop an approach future work in this area could take in identifying the extent of confounding, and in the use of tools to mitigate bias during neural network training. A metric deemed especially useful in this task is the distance correlation which allows for the extent of statistical independence to be measured in multi-dimensional space. A DC value of zero establishes the desirable property of statistical mean independence between two vectors; in this case as measured between genomic predictions and any number of confounding variables.

Without any modification to network architecture, this metric can be useful as a tool to measure bias during training and the establishment of an appropriate threshold can inform a training stopping point. Furthermore, if one is using a test set during model training and

hyperparameter optimization, as is almost always the case, it is further contextual information one can use in selecting the final model(s) sent for validation.

The distance correlation was also explored in the context of examining several new methods that have been proposed to mitigate bias during network training. These dual-task approaches, which include adversarial designs, are notoriously difficult to train but have shown promise in other contexts in reducing confounding in predictions. The DisCo networks proposed by Kasieczka and Shih were found to be the simplest and most effective method to implement and often successfully reduced the prediction DC in comparison to standard feed-forward neural networks. However, statistical mean independence proved elusive when including many confounding factors and a trade-off existed between predictive performance and bias reduction. This reduction in bias using DisCo networks was not observed when using patient strata information in the ALS dataset however, which may point towards a fundamental limitation of the method in some use cases, although further optimization may improve the performance.

As noted in Section 1.3.3, the exact point at which environmental effects become unwanted in risk profiling is contentious and will probably be case-specific. In fact, the issue of removing environmental effects strikes right at the heart of the nature versus nurture debate as exemplified by Equation 1.1. In reality, nature and nurture are interwoven, perhaps implacably so, and it is evident that such a complex web will not easily be disentangled.

Even the gold-standard approach of a family-based design (see Section 1.3.4) cannot account for every form of unwanted bias. These problems, including dynastic effects and geographic self-clustering (and which are beyond the scope of this thesis), are most acute for human behavioral traits, however, they may one day prove relevant to more proximally "biological" traits as certain behaviors (e.g. smoking and diet choices) are known risk factors for a myriad of disorders (Abdellaoui et al. 2022; Abdellaoui and Verweij 2021).

Chapter 4 is concerned only with implementing novel approaches to confounder handling in the case of deep learning, although the problem exists across many machine learning frameworks. Future research in this vein would naturally extend to benchmarking similar DC-based mitigation methods in other settings. Furthermore, within each machine learning family, there have been myriad methods proposed to deal with confounding, each of which could be investigated should they look feasible and promising. With no gold-standard method with which to truly gauge the success of each method, the task is a daunting one, however, if machine learning based genomic predictions are to be implemented clinically it is essential that they are trustworthy and free from bias. For this reason, it should be a large priority for quantitative genetics to grapple with confounder handling as the interest in and capabilities of machine learning is only growing.

Although this work attempted to gauge the potential improvements to be made using a broad variety of machine learning families, there are many other algorithms that could be included in a future analysis. Limitations of this thesis include not having access to external validation datasets for either the *Arabidopsis thaliana* data or the ALS patient cohort. The use of a nested cross-validation design protected against the unwanted effects of overfitting and should give a reasonable estimation of actual relative performance on other datasets, however, access to external data is an invaluable resource to measure performance transferability, but this was unfortunately not available here.

Additionally, computational power limited experimental investigation beyond the use of 10,000 SNPs, though there is no real limit to the number of unlinked SNPs one would want to investigate with the availability of unlimited resources. Likewise, the performance of random grid-searching and Bayesian searching is to some extent a function of time, and so future analysis with larger datasets and more resources could yield additional insights.

Comparisons between multiple factors can be limited by statistical power, and this was certainly the case here as there was a trade-off between outer cross-validation folds being

numerous enough to run a powerful Dunnett test, and each fold containing enough samples to have generalizable results.

5.2 Conclusion

In conclusion, this thesis has added to the body of knowledge regarding the potential performance of machine learning approaches in the task of genomic prediction. It lays out a framework of how to approach and implement this task, the potential benefits and limitations of machine learning, as well as statistical methods by which to analyze the results. It also highlights the importance of detecting and accounting for bias in non-linear frameworks and offers some insight into the benefits and practicality of recently proposed techniques to handle confounding during model training, a serious problem in genomic prediction.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Abdellaoui, A., Dolan, C. V., Verweij, K. J. H., and Nivard, M. G. (2022). Gene–environment correlations across geographic regions affect genome-wide association studies. *Nature Genetics*, 54(9):1345–1354.

Abdellaoui, A. and Verweij, K. J. H. (2021). Dissecting polygenic signals from genome-wide association studies on human behaviour. *Nature Human Behaviour*.

Abegaz, F., Chaichoompu, K., Génin, E., Fardo, D. W., König, I. R., Mahachie John, J. M., and Van Steen, K. (2019). Principals about principal components in statistical genetics. *Briefings in Bioinformatics*, 20(6):2200–2216.

Adeli, E., Zhao, Q., Pfefferbaum, A., Sullivan, E. V., Fei-Fei, L., Niebles, J. C., and Pohl, K. M. (2021). Representation learning with statistical independence to mitigate bias. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2512–2522.

Al-Chalabi, A., Fang, F., Hanby, M. F., Leigh, P. N., Shaw, C. E., Ye, W., and Rijsdijk, F. (2010). An estimate of amyotrophic lateral sclerosis heritability using twin data. *Journal of Neurology, Neurosurgery & Psychiatry*, 81(12):1324–1326.

Al-Chalabi, A. and Hardiman, O. (2013). The epidemiology of ALS: a conspiracy of genes, environment and time. *Nature Reviews Neurology*, 9(11):617–628.

Al-Chalabi, A., van den Berg, L. H., and Veldink, J. (2017). Gene discovery in amyotrophic lateral sclerosis: implications for clinical management. *Nature Reviews Neurology*, 13(2):96–104.

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., Cao, J., Chae, E., Dezwaan, T. M., Ding, W., Ecker, J. R., Exposito-Alonso, M., Farlow, A., Fitz, J., Gan, X., Grimm, D. G., Hancock, A. M., Henz, S. R., Holm, S., Horton, M., Jarsulic, M., Kerstetter, R. A., Korte, A., Korte, P., Lanz, C., Lee, C.-R., Meng, D., Michael, T. P., Mott, R., Muliyati, N. W., Nägele, T., Nagler, M., Nizhynska, V., Nordborg, M., Novikova, P. Y., Picó, F. X., Platzer, A., Rabanal, F. A., Rodriguez, A., Rowan, B. A.,

Salomé, P. A., Schmid, K. J., Schmitz, R. J., Seren, Ü., Sperone, F. G., Sudkamp, M., Svardal, H., Tanzer, M. M., Todd, D., Volchenboum, S. L., Wang, C., Wang, G., Wang, X., Weckwerth, W., Weigel, D., and Zhou, X. (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell*, 166(2):481–491.

Altshuler, D., Daly, M. J., and Lander, E. S. (2008). Genetic Mapping in Human Disease. *Science*, 322(5903):881–888.

Baker, E. and Escott-Price, V. (2020). Polygenic risk scores in alzheimer's disease: Current applications and future directions. *Frontiers in Digital Health*, 2.

Barton, N., Hermisson, J., and Nordborg, M. (2019). Population Genetics: Why Structure Matters. *eLife*, 8.

Bateson, W. (1909). *Mendel's Principles of Heredity*. Cambridge University Press.

Bellot, P., de Los Campos, G., and Pérez-Enciso, M. (2018). Can Deep Learning Improve Genomic Prediction of Complex Human Traits? *Genetics*, 210(3):809–819.

Bernabeu, E., Canela-Xandri, O., Rawlik, K., Talenti, A., Prendergast, J., and Tenesa, A. (2021). Sex differences in genetic architecture in the UK biobank. *Nature Genetics*, 53(9):1283–1289.

Bernardo, R. (2020). Reinventing quantitative genetics for plant breeding: something old, something new, something borrowed, something BLUE. *Heredity*, 125(6):375–385.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer Science + Business Media, 1st edition.

Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*, 169(7):1177–1186.

Bracher-Smith, M., Crawford, K., and Escott-Price, V. (2021). Machine learning for genetic prediction of psychiatric disorders: a systematic review. *Molecular Psychiatry*, 26(1):70–79.

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., and Neale, B. M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3):291–295.

Carey, G. (2003). Quantitative Genetics: II - Advanced Topics. In *Human Genetics for the Social Sciences*, chapter 19. Sage Publications, London.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1):7.

Charmantier, A., Garant, D., and Kruuk, L. E. B. (2014). *Quantitative Genetics in the Wild*. Oxford University Press, Oxford.

Che, R., Motsinger-Reif, A. A., and Brown, C. C. (2012). Loss of Power in Two-Stage Residual-Outcome Regression Analysis in Genetic Association Studies. *Genetic Epidemiology*, pages 890–894.

Chen, Z., Boehnke, M., Wen, X., and Mukherjee, B. (2021). Revisiting the genome-wide significance threshold for common variant GWAS. *G3 Genes|Genomes|Genetics*, 11(2).

Chiu, H.-S., Somvanshi, S., Patel, E., Chen, T.-W., Singh, V. P., Zorman, B., Patil, S. L., Pan, Y., Chatterjee, S. S., Sood, A. K., Gunaratne, P. H., Sumazin, P., Caesar-Johnson, S. J., Demchok, J. A., Felau, I., Kasapi, M., Ferguson, M. L., Hutter, C. M., Sofia, H. J., Tarnuzzer, R., Wang, Z., Yang, L., Zenklusen, J. C., Zhang, J. J., Chudamani, S., Liu, J., Lolla, L., Naresh, R., Pihl, T., Sun, Q., Wan, Y., Wu, Y., Cho, J., DeFreitas, T., Frazer, S., Gehlenborg, N., Getz, G., Heiman, D. I., Kim, J., Lawrence, M. S., Lin, P., Meier, S., Noble, M. S., Saksena, G., Voet, D., Zhang, H., Bernard, B., Chambwe, N., Dhankani, V., Knijnenburg, T., Kramer, R., Leinonen, K., Liu, Y., Miller, M., Reynolds, S., Shmulevich, I., Thorsson, V., Zhang, W., Akbani, R., Broom, B. M., Hegde, A. M., Ju, Z., Kanchi, R. S., Korkut, A., Li, J., Liang, H., Ling, S., Liu, W., Lu, Y., Mills, G. B., Ng, K.-S., Rao, A., Ryan, M., Wang, J., Weinstein, J. N., Zhang, J., Abeshouse, A., Armenia, J., Chakravarty, D., Chatila, W. K., de Bruijn, I., Gao, J., Gross, B. E., Heins, Z. J., Kundra, R., La, K., Ladanyi, M., Luna, A., Nissan, M. G., Ochoa, A., Phillips, S. M., Reznik, E., Sanchez-Vega, F., Sander, C., Schultz, N., Sheridan, R., Sumer, S. O., Sun, Y., Taylor, B. S., Wang, J., Zhang, H., Anur, P., Peto, M., Spellman, P., Benz, C., Stuart, J. M., Wong, C. K., Yau, C., Hayes, D. N., Parker, J. S., Wilkerson, M. D., Ally, A., Balasundaram, M., Bowlby, R., Brooks, D., Carlsen, R., Chuah, E., Dhalla, N., Holt, R., Jones, S. J., Kasaian, K., Lee, D., Ma, Y., Marra, M. A., Mayo, M., Moore, R. A., Mungall, A. J., Mungall, K., Robertson, A. G., Sadeghi, S., Schein, J. E., Sipahimalani, P., Tam, A., Thiessen, N., Tse, K., Wong, T., Berger, A. C., Beroukhim, R., Cherniack, A. D., Cibulskis, C., Gabriel, S. B., Gao, G. F., Ha, G., Meyerson, M., Schumacher, S. E., Shih, J., Kucherlapati, M. H., Kucherlapati, R. S., Baylin, S., Cope, L., Danilova, L., Bootwalla, M. S., Lai, P. H., Maglinte, D. T., Van Den Berg, D. J., Weisenberger, D. J., Auman, J. T., Balu, S., Bodenheimer, T., Fan, C., Hoadley, K. A., Hoyle, A. P., Jefferys, S. R., Jones, C. D., Meng, S., Mieczkowski, P. A., Mose, L. E., Perou, A. H., Perou, C. M., Roach, J., Shi, Y., Simons, J. V., Skelly, T., Soloway, M. G., Tan, D., Veluvolu, U., Fan, H., Hinoue, T., Laird, P. W., Shen, H., Zhou, W., Bellair, M., Chang, K., Covington, K., Creighton, C. J., Dinh, H., Doddapaneni, H., Donehower, L. A., Drummond, J., Gibbs, R. A., Glenn, R., Hale, W., Han, Y., Hu, J., Korchina, V., Lee, S., Lewis, L., Li, W., Liu, X., Morgan, M., Morton, D., Muzny, D., Santibanez, J., Sheth, M., Shinbrot, E., Wang, L., Wang, M., Wheeler, D. A., Xi, L., Zhao, F., Hess, J., Appelbaum, E. L., Bailey, M., Cordes, M. G., Ding, L., Fronick, C. C., Fulton, L. A., Fulton, R. S., Kandoth, C., Mardis, E. R., McLellan, M. D., Miller, C. A., Schmidt, H. K., Wilson, R. K., Crain, D., Curley, E., Gardner, J., Lau, K., Mallery, D., Morris, S., Paulauskis, J., Penny, R., Shelton, C., Shelton, T., Sherman, M., Thompson, E., Yena, P., Bowen, J., Gastier-Foster, J. M., Gerken, M., Leraas, K. M., Lichtenberg, T. M., Ramirez, N. C., Wise, L., Zmuda, E., Corcoran, N., Costello, T., Hovens, C., Carvalho, A. L., de Carvalho, A. C., Fregnani, J. H., Longatto-Filho, A., Reis, R. M., Scapulatempo-Neto, C., Silveira, H. C., Vidal, D. O., Burnette, A., Eschbacher, J., Hermes, B., Noss, A., Singh, R., Anderson, M. L., Castro, P. D., Ittmann, M., Huntsman, D., Kohl, B., Le, X., Thorp, R., Andry, C., Duffy, E. R., Lyadov, V., Paklina, O., Setdikova, G., Shabunin, A., Tavobilov, M., McPherson, C., Warnick, R., Berkowitz, R., Cramer, D., Feltmate, C., Horowitz, N., Kibel, A., Muto, M., Raut, C. P., Malykh, A., Barnholtz-Sloan,

J. S., Barrett, W., Devine, K., Fulop, J., Ostrom, Q. T., Shimmel, K., Wolinsky, Y., Sloan, A. E., De Rose, A., Giuliante, F., Goodman, M., Karlan, B. Y., Hagedorn, C. H., Eckman, J., Harr, J., Myers, J., Tucker, K., Zach, L. A., Deyarmin, B., Hu, H., Kvecher, L., Larson, C., Mural, R. J., Somiari, S., Vicha, A., Zelinka, T., Bennett, J., Iacocca, M., Rabeno, B., Swanson, P., Latour, M., Lacombe, L., Têtu, B., Bergeron, A., McGraw, M., Staugaitis, S. M., Chabot, J., Hibshoosh, H., Sepulveda, A., Su, T., Wang, T., Potapova, O., Voronina, O., Desjardins, L., Mariani, O., Roman-Roman, S., Sastre, X., Stern, M.-H., Cheng, F., Signoretti, S., Berchuck, A., Bigner, D., Lipp, E., Marks, J., McCall, S., McLendon, R., Secord, A., Sharp, A., Behera, M., Brat, D. J., Chen, A., Delman, K., Force, S., Khuri, F., Magliocca, K., Maithel, S., Olson, J. J., Owonikoko, T., Pickens, A., Ramalingam, S., Shin, D. M., Sica, G., Van Meir, E. G., Zhang, H., Eijckenboom, W., Gillis, A., Korpershoek, E., Looijenga, L., Oosterhuis, W., Stoop, H., van Kessel, K. E., Zwarthoff, E. C., Calatozzolo, C., Cuppini, L., Cuzzubbo, S., DiMeco, F., Finocchiaro, G., Mattei, L., Perin, A., Pollo, B., Chen, C., Houck, J., Lohavanichbutr, P., Hartmann, A., Stoehr, C., Stoehr, R., Taubert, H., Wach, S., Wullich, B., Kycler, W., Murawa, D., Wiznerowicz, M., Chung, K., Edenfield, W. J., Martin, J., Baudin, E., Bubley, G., Bueno, R., De Rienzo, A., Richards, W. G., Kalkanis, S., Mikkelsen, T., Noushmehr, H., Scarpace, L., Girard, N., Aymerich, M., Campo, E., Giné, E., Guillermo, A. L., Van Bang, N., Hanh, P. T., Phu, B. D., Tang, Y., Colman, H., Evason, K., Dottino, P. R., Martignetti, J. A., Gabra, H., Juhl, H., Akeredolu, T., Stepa, S., Hoon, D., Ahn, K., Kang, K. J., Beuschlein, F., Breggia, A., Birrer, M., Bell, D., Borad, M., Bryce, A. H., Castle, E., Chandan, V., Cheville, J., Copland, J. A., Farnell, M., Flotte, T., Giama, N., Ho, T., Kendrick, M., Kocher, J.-P., Kopp, K., Moser, C., Nagorney, D., O'Brien, D., O'Neill, B. P., Patel, T., Petersen, G., Que, F., Rivera, M., Roberts, L., Smallridge, R., Smyrk, T., Stanton, M., Thompson, R. H., Torbenson, M., Yang, J. D., Zhang, L., Brimo, F., Ajani, J. A., Gonzalez, A. M. A., Behrens, C., Bondaruk, J., Broaddus, R., Czerniak, B., Esmaeli, B., Fujimoto, J., Gershenwald, J., Guo, C., Lazar, A. J., Logothetis, C., Meric-Bernstam, F., Moran, C., Ramondetta, L., Rice, D., Sood, A., Tamboli, P., Thompson, T., Troncoso, P., Tsao, A., Wistuba, I., Carter, C., Haydu, L., Hersey, P., Jakrot, V., Kakavand, H., Kefford, R., Lee, K., Long, G., Mann, G., Quinn, M., Saw, R., Scolyer, R., Shannon, K., Spillane, A., Stretch, J., Synott, M., Thompson, J., Wilmott, J., Al-Ahmadie, H., Chan, T. A., Ghossein, R., Gopalan, A., Levine, D. A., Reuter, V., Singer, S., Singh, B., Tien, N. V., Broudy, T., Mirsaidi, C., Nair, P., Drwiega, P., Miller, J., Smith, J., Zaren, H., Park, J.-W., Hung, N. P., Kebebew, E., Linehan, W. M., Metwalli, A. R., Pacak, K., Pinto, P. A., Schiffman, M., Schmidt, L. S., Vocke, C. D., Wentzensen, N., Worrell, R., Yang, H., Moncrieff, M., Goparaju, C., Melamed, J., Pass, H., Botnariuc, N., Caraman, I., Cernat, M., Chemencedji, I., Clipca, A., Doruc, S., Gorincioi, G., Mura, S., Pirtac, M., Stancul, I., Tcaciuc, D., Albert, M., Alexopoulou, I., Arnaout, A., Bartlett, J., Engel, J., Gilbert, S., Parfitt, J., Sekhon, H., Thomas, G., Rassl, D. M., Rintoul, R. C., Bifulco, C., Tamakawa, R., Urba, W., Hayward, N., Timmers, H., Antenucci, A., Facciolo, F., Grazi, G., Marino, M., Merola, R., de Krijger, R., Gimenez-Roqueplo, A.-P., Piché, A., Chevalier, S., McKercher, G., Birsoy, K., Barnett, G., Brewer, C., Farver, C., Naska, T., Pennell, N. A., Raymond, D., Schilero, C., Smolenski, K., Williams, F., Morrison, C., Borgia, J. A., Liptay, M. J., Pool, M., Seder, C. W., Junker, K., Omberg, L., Dinkin, M., Manikhas, G., Alvaro, D., Bragazzi, M. C., Cardinale, V., Carpino, G., Gaudio, E., Chesla, D., Cottingham, S., Dubina, M., Moiseenko, F., Dhanasekaran, R., Becker, K.-F., Janssen, K.-P., Slotta-Huspenina, J., Abdel-Rahman, M. H., Aziz, D., Bell, S., Cebulla, C. M., Davis, A., Duell, R., Elder, J. B., Hilty, J., Kumar, B., Lang, J., Lehman, N. L., Mandt, R., Nguyen, P., Pilarski, R., Rai, K., Schoenfield, L., Senecal, K., Wakely,

P., Hansen, P., Lechan, R., Powers, J., Tischler, A., Grizzle, W. E., Sexton, K. C., Kastl, A., Henderson, J., Porten, S., Waldmann, J., Fassnacht, M., Asa, S. L., Schadendorf, D., Couce, M., Graefen, M., Huland, H., Sauter, G., Schlomm, T., Simon, R., Tennstedt, P., Olabode, O., Nelson, M., Bathe, O., Carroll, P. R., Chan, J. M., Disaia, P., Glenn, P., Kelley, R. K., Landen, C. N., Phillips, J., Prados, M., Simko, J., Smith-McCune, K., VandenBerg, S., Roggin, K., Fehrenbach, A., Kendler, A., Sifri, S., Steele, R., Jimeno, A., Carey, F., Forgie, I., Mannelli, M., Carney, M., Hernandez, B., Campos, B., Herold-Mende, C., Jungk, C., Unterberg, A., von Deimling, A., Bossler, A., Galbraith, J., Jacobus, L., Knudson, M., Knutson, T., Ma, D., Milhem, M., Sigmund, R., Godwin, A. K., Madan, R., Rosenthal, H. G., Adebamowo, C., Adebamowo, S. N., Boussioutas, A., Beer, D., Giordano, T., Mes-Masson, A.-M., Saad, F., Bocklage, T., Landrum, L., Mannel, R., Moore, K., Moxley, K., Postier, R., Walker, J., Zuna, R., Feldman, M., Valdivieso, F., Dhir, R., Luketich, J., Pinero, E. M. M., Quintero-Aguilo, M., Carlotti, C. G., Dos Santos, J. S., Kemp, R., Sankarankuty, A., Tirapelli, D., Catto, J., Agnew, K., Swisher, E., Creaney, J., Robinson, B., Shelley, C. S., Godwin, E. M., Kendall, S., Shipman, C., Bradford, C., Carey, T., Haddad, A., Moyer, J., Peterson, L., Prince, M., Rozek, L., Wolf, G., Bowman, R., Fong, K. M., Yang, I., Korst, R., Rathmell, W. K., Fantacone-Campbell, J. L., Hooke, J. A., Kovatich, A. J., Shriver, C. D., DiPersio, J., Drake, B., Govindan, R., Heath, S., Ley, T., Van Tine, B., Westervelt, P., Rubin, M. A., Lee, J. I., Aredes, N. D., and Mariamidze, A. (2018). Pan-Cancer Analysis of lncRNA Regulation Supports Their Targeting of Cancer Genes in Each Tumor Context. *Cell Reports*, 23(1):297–312.e12.

Choi, S. W., Mak, T. S.-H., and O'Reilly, P. F. (2020). Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols*, 15:2759–2772.

Chollet, F. et al. (2015). Keras: The python deep learning library. *Available from: https://github.com/fchollet/keras*.

Cook, N. R. (2007). Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction. *Circulation*, 115(7):928–935.

Cope, J. L., Baukmann, H. A., Klinger, J. E., Ravarani, C. N. J., Böttinger, E. P., Konigorski, S., and Schmidt, M. F. (2021). Interaction-Based Feature Selection Algorithm Outperforms Polygenic Risk Score in Predicting Parkinson's Disease Status. *Frontiers in Genetics*, 12.

Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human molecular genetics*, 11(20):2463–8.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker, S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, M., Rutkoski, J., and Varshney, R. K. (2017). Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends in Plant Science*, 22(11):961–975.

Dai, Z., Long, N., and Huang, W. (2020). Influence of Genetic Interactions on Polygenic Prediction. *G3 Genes|Genomes|Genetics*, 10(1):109–115.

DeJesus-Hernandez, M., Mackenzie, I. R., Boeve, B. F., Boxer, A. L., Baker, M., Rutherford, N. J., Nicholson, A. M., Finch, N. A., Flynn, H., Adamson, J., Kouri, N., Wojtas, A., Sengdy, P., Hsiung, G.-Y. R., Karydas, A., Seeley, W. W., Josephs, K. A., Coppola, G.,

Geschwind, D. H., Wszolek, Z. K., Feldman, H., Knopman, D. S., Petersen, R. C., Miller, B. L., Dickson, D. W., Boylan, K. B., Graff-Radford, N. R., and Rademakers, R. (2011). Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron*, 72(2):245–56.

Demissie, S. and Cupples, L. A. (2011). Bias due to two-stage residual-outcome regression analysis in genetic association studies. *Genetic epidemiology*, 35(7):592–6.

Dinga, R., Schmaal, L., Penninx, B. W. J. H., Veltman, D. J., and Marquand, A. F. (2020). Controlling for effects of confounding variables on machine learning predictions. *biorXiv:2020.08.17.255034*.

Dunnett, C. W. (1955). A Multiple Comparison Procedure for Comparing Several Treatments with a Control. *Journal of the American Statistical Association*, 50(272):1096–1121.

Edelmann, D., Fokianos, K., and Pitsillou, M. (2019). An Updated Literature Review of Distance Correlation and Its Applications to Time Series. *International Statistical Review*, 87(2):237–262.

Euesden, J., Lewis, C. M., and O'Reilly, P. F. (2015). PRSice: Polygenic Risk Score software. *Bioinformatics*, 31(9):1466–1468.

Evans, L. M., Tahmasbi, R., Vrieze, S. I., Abecasis, G. R., Das, S., Gazal, S., Bjelland, D. W., de Candia, T. R., Goddard, M. E., Neale, B. M., Yang, J., Visscher, P. M., and Keller, M. C. (2018). Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nature Genetics*, 50(5):737–745.

Falconer, D. and Mackay, T. (1995). *Introduction to Quantitative Genetics*. Pearson Education, 4th edition.

Falconer, D. S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics*, 29(1):51–76.

Farahani, A., Voghoei, S., Rasheed, K., and Arabnia, H. R. (2021). A Brief Review of Domain Adaptation. In Stahlbock, R., Weiss, G., Abou-Nasr, M., Yang, C., H.R., A., and Deligiannidis, L., editors, *Advances in Data Science and Information Engineering*, pages 877–894. Springer.

Fisher, R. A. (1919). The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*.

Freckleton, R. P. (2002). On the misuse of residuals in ecology: regression of residuals vs. multiple regression. *Journal of Animal Ecology*, 71(3):542–545.

Friedman, J., Hastie, T., and Tibshirani, R. (2009). *The Elements of Statistical Learning*. Springer Science & Business Media, New York, 2nd edition.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.

Gola, D., Erdmann, J., Müller-Myhsok, B., Schunkert, H., and König, I. R. (2020). Polygenic risk scores outperform machine learning methods in predicting coronary artery disease status. *Genetic Epidemiology*, 44(2):125–138.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. In *Advances in Neural Information Processing Systems, 27*, pages 2672–2680.

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc., 2nd edition.

Habier, D., Fernando, R. L., and Dekkers, J. C. M. (2007). The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. *Genetics*, 177(4):2389–2397.

Hardiman, O., Al-Chalabi, A., Chio, A., Corr, E. M., Logroscino, G., Robberecht, W., Shaw, P. J., Simmons, Z., and van den Berg, L. H. (2017). Amyotrophic lateral sclerosis. *Nature Reviews Disease Primers*, 3(1):17071.

Hayes, B. (2013). *Overview of Statistical Methods for Genome-Wide Association Studies (GWAS)*, pages 149–169. Humana Press, Totowa, NJ.

He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., and Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine*, 25(1):30–36.

Hemani, G., Knott, S., and Haley, C. (2013). An evolutionary perspective on epistasis and the missing heritability. *PLoS Genetics*, 9:e1003295.

Hickey, J. M., Chiurugwi, T., Mackay, I., and Powell, W. (2017). Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nature Genetics*, 49(9):1297–1303.

Hill, W. G., Goddard, M. E., and Visscher, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics*, 4:e1000008.

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, pages 278–282.

Hothorn, L. A. and Kluxen, F. M. (2019). Robust multiple comparisons against a control group with application in toxicology. *arXiv:1905.01838v1 [stat.AP]*.

Hothorn, T. (2020). Most likely transformations: The mlt package. *Journal of Statistical Software*, 92(1):1–68.

Hou, J., Ye, X., Feng, W., Zhang, Q., Han, Y., Liu, Y., Li, Y., and Wei, Y. (2022). Distance correlation application to gene co-expression network analysis. *BMC Bioinformatics*, 23(1):81.

Hua, J., Tembe, W. D., and Dougherty, E. R. (2009). Performance of feature-selection methods in the classification of high-dimension data. *Pattern Recognition*, 42(3):409–424.

Huang, W. and Mackay, T. F. C. (2016). The Genetic Architecture of Quantitative Traits Cannot Be Inferred from Variance Component Analysis. *PLOS Genetics*, 12(11):e1006421.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R*, volume 2. Springer Science.

Janssens, A. C. J. W., Moonesinghe, R., Yang, Q., Steyerberg, E. W., van Duijn, C. M., and Khoury, M. J. (2007). The impact of genotype frequencies on the clinical validity of genomic profiling for predicting common chronic diseases. *Genetics in Medicine*, 9(8):528–535.

Japkowicz, N. and Shah, M. (2011). *Evaluating Learning Algorithims: A Classification Perspective*. Cambridge University Press, New York.

Johnson, D., Wilke, M. A., Lyle, S. M., Kowalec, K., Jorgensen, A., Wright, G. E., and Drögemöller, B. I. (2022). A Systematic Review and Analysis of the Use of Polygenic Scores in Pharmacogenomics. *Clinical Pharmacology & Therapeutics*, 111(4):919–930.

Johnson, R. C., Nelson, G. W., Troyer, J. L., Lautenberger, J. A., Kessing, B. D., Winkler, C. A., and O'Brien, S. J. (2010). Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics*, 11(1):724.

Kaplan, J. M. and Fullerton, S. M. (2022). Polygenic risk, population structure and ongoing difficulties with race in human genetics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1852).

Karavani, E., Zuk, O., Zeevi, D., Barzilai, N., Stefanis, N. C., Hatzimanolis, A., Smyrnis, N., Avramopoulos, D., Kruglyak, L., Atzmon, G., Lam, M., Lencz, T., and Carmi, S. (2019). Screening human embryos for polygenic traits has limited utility. *Cell*, 179(6):1424–1435.e8.

Kasieczka, G. and Shih, D. (2020). Robust jet classifiers through distance correlation. *Physical Review Letters*, 125:122001.1–122001.7.

Katsaouni, N., Tashkandi, A., Wiese, L., and Schulz, M. H. (2021). Machine learning based disease prediction from genotype data. *Biological Chemistry*, 402(8):871–885.

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., Natarajan, P., Lander, E. S., Lubitz, S. A., Ellinor, P. T., and Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50(9):1219–1224.

Kingma, D. P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980v9 [cs.LG]*.

Konietschke, F., Placzek, M., Schaarschmidt, F., and Hothorn, L. A. (2015). nparcomp: An R software package for nonparametric multiple comparisons and simultaneous confidence intervals. *Journal of Statistical Software*, 64(9):1–17.

Koornneef, M. and Meinke, D. (2010). The development of Arabidopsis as a model plant. *The Plant Journal*, 61(6):909–921.

Krakovska, O., Christie, G., Sixsmith, A., Ester, M., and Moreno, S. (2019). Performance comparison of linear and non-linear feature selection methods for the analysis of large survey datasets. *PloS one*, 14(3):e0213584.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232.

Krishnappa, G., Savadi, S., Tyagi, B. S., Singh, S. K., Mamrutha, H. M., Kumar, S., Mishra, C. N., Khan, H., Gangadhara, K., Uday, G., Singh, G., and Singh, G. P. (2021). Integrated genomic selection for rapid improvement of crops. *Genomics*, 113(3):1070–1086.

Krstajic, D., Buturovic, L. J., Leahy, D. E., and Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics*, 6(1):10.

Kundaje, A., , Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., Jager, P. L. D., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T., and Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330.

Kvålseth, T. O. (1985). Cautionary note about R^2. *The American Statistician*, 39(4):279–285.

LaFramboise, T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Research*, 37(13):4181–4193.

Laibach, F. (1951). Summer- and winter-annual races of A. thaliana. A contribution to the etiology of flower development. *Beiträge zur Biologie der Pflanzen*, 28:173–210.

Lander, E. and Schork, N. (1994). Genetic dissection of complex traits. *Science*, 265(5181):2037–2048.

Lee, S. H., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2012). A Better Coefficient of Determination for Genetic Profile Analysis. *Genetic Epidemiology*, 36(3):214–224.

Lello, L., Avery, S. G., Tellier, L., Vazquez, A. I., de los Campos, G., and Hsu, S. D. H. (2018). Accurate Genomic Prediction of Human Height. *Genetics*, 210(2):477–497.

Lello, L., Raben, T. G., and Hsu, S. D. H. (2020). Sibling validation of polygenic risk scores and complex trait prediction. *Scientific Reports*, 10(1):13190.

Lencz, T., Sabatello, M., Docherty, A., Peterson, R. E., Soda, T., Austin, J., Bierut, L., Crepaz-Keay, D., Curtis, D., Degenhardt, F., Huckins, L., Lazaro-Munoz, G., Mattheisen, M., Meiser, B., Peay, H., Rietschel, M., Walss-Bass, C., and Davis, L. K. (2022). Concerns about the use of polygenic embryo screening for psychiatric and cognitive traits. *The Lancet Psychiatry*, 9(10):838–844.

Lewis, A. C. F. and Green, R. C. (2021). Polygenic risk scores in the clinic: new perspectives needed on familiar ethical issues. *Genome Medicine*, 13(1):14.

Li, L., Rakitsch, B., and Borgwardt, K. (2011). ccSVM: correcting Support Vector Machines for confounding factors in biological data classification. *Bioinformatics*, 27(13):i342–i348.

Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype Imputation. *Annual Review of Genomics and Human Genetics*, 10(1):387–406.

Liu, Y. and Wang, D. (2017). Application of deep learning in genomic selection. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2280–2280. IEEE.

Liu, Y., Wang, D., He, F., Wang, J., Joshi, T., and Xu, D. (2019). Phenotype Prediction and Genome-Wide Association Study Using Deep Convolutional Neural Network of Soybean. *Frontiers in Genetics*, 10.

Louppe, G., Kagan, M., and Cranmer, K. (2017). Learning to Pivot with Adversarial Networks. *arXiv:1611.01046v3 [stat.ML]*.

Ma, W., Qiu, Z., Song, J., Li, J., Cheng, Q., Zhai, J., and Ma, C. (2018). A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta*, 248(5):1307–1318.

Mackay, T. F. C. (2001). The Genetic Architecture of Quantitative Traits. *Annual Review of Genetics*, 35(1):303–339.

Mackenzie, I. R. A., Bigio, E. H., Ince, P. G., Geser, F., Neumann, M., Cairns, N. J., Kwong, L. K., Forman, M. S., Ravits, J., Stewart, H., Eisen, A., McClusky, L., Kretzschmar, H. A., Monoranu, C. M., Highley, J. R., Kirby, J., Siddique, T., Shaw, P. J., Lee, V. M.-Y., and Trojanowski, J. Q. (2007). Pathological TDP-43 distinguishes sporadic amyotrophic lateral sclerosis from amyotrophic lateral sclerosis with SOD1 mutations. *Annals of neurology*, 61(5):427–34.

Maekawa, S., Leigh, P. N., King, A., Jones, E., Steele, J. C., Bodi, I., Shaw, C. E., Hortobagyi, T., and Al-Sarraj, S. (2009). TDP-43 is consistently co-localized with ubiquitinated inclusions in sporadic and Guam amyotrophic lateral sclerosis but not in familial amyotrophic lateral sclerosis with and without SOD1 mutations. *Neuropathology : official journal of the Japanese Society of Neuropathology*, 29(6):672–83.

Maekawa, T., Higashide, D., Hara, T., Matsumura, K., Ide, K., Miyatake, T., Kimura, K. D., and Takahashi, S. (2021). Cross-species behavior analysis with attention-based domain-adversarial deep neural networks. *Nat. Commun.*, 12(1):5519.

Mäki-Tanila, A. and Hill, W. G. (2014). Influence of Gene Interaction on Complex Trait Variation with Multilocus Models. *Genetics*, 198(1):355–367.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F. C., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753.

Márquez-Luna, C., Loh, P.-R., and Price, A. L. (2017). Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genetic Epidemiology*, 41(8):811–823.

Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., Daly, M. J., Bustamante, C. D., and Kenny, E. E. (2017). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *The American Journal of Human Genetics*, 100(4):635–649.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., Bates, D., Hansen, R. S., Neph, S., Sabo, P. J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S. R., Kaul, R., and Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099):1190–1195.

Maxwell, S. E., Delaney, H. D., and Kelley, K. (2018). *Designing Experiments and Analyzing Data: A Model Comparison Perspective*. Routledge, New York, 3rd edition.

McCann, E. P., Henden, L., Fifita, J. A., Zhang, K. Y., Grima, N., Bauer, D. C., Chan Moi Fat, S., Twine, N. A., Pamphlett, R., Kiernan, M. C., Rowe, D. B., Williams, K. L., and Blair, I. P. (2020). Evidence for polygenic and oligogenic basis of Australian sporadic amyotrophic lateral sclerosis. *Journal of Medical Genetics*.

Mejzini, R., Flynn, L. L., Pitout, I. L., Fletcher, S., Wilton, S. D., and Akkari, P. A. (2019). ALS Genetics, Mechanisms, and Therapeutics: Where Are We Now? *Frontiers in Neuroscience*, 13.

Mills, M. C. and Rahal, C. (2020). The GWAS Diversity Monitor tracks diversity by disease in real time. *Nature Genetics*, 52(3):242–243.

Montesinos-López, A., Montesinos-López, O. A., Gianola, D., Crossa, J., and Hernández-Suárez, C. M. (2018). Multi-environment Genomic Prediction of Plant Traits Using Deep Learners With Dense Architecture. *G3 Genes|Genomes|Genetics*, 8(12):3813–3828.

Montesinos-López, O. A., Martín-Vallejo, J., Crossa, J., Gianola, D., Hernández-Suárez, C. M., Montesinos-López, A., Juliana, P., and Singh, R. (2019). A Benchmarking Between Deep Learning, Support Vector Machine and Bayesian Threshold Best Linear Unbiased Prediction for Predicting Ordinal Traits in Plant Breeding. *G3 Genes|Genomes|Genetics*, 9(2):601–618.

Moore, J. H. (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human Heredity*, 56:73–82.

Morgante, F., Huang, W., Maltecca, C., and Mackay, T. F. C. (2018). Effect of genetic architecture on the prediction accuracy of quantitative traits in samples of unrelated individuals. *Heredity*, 120(6):500–514.

Norman, T. M., Horlbeck, M. A., Replogle, J. M., Ge, A. Y., Xu, A., Jost, M., Gilbert, L. A., and Weissman, J. S. (2019). Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M., and Bustamante, C. D. (2008). Genes mirror geography within Europe. *Nature*, 456(7218):98–101.

Nyquist, W. E. (1991). Estimation of Heritability and Prediction of Selection Response in Plant Populations. *Critical Reviews in Plant Sciences*, 10(3):235–322.

Ober, U., Huang, W., Magwire, M., Schlather, M., Simianer, H., and Mackay, T. F. C. (2015). Accounting for genetic architecture improves sequence based genomic prediction for a Drosophila fitness trait. *PloS one*, 10(5):e0126880.

Ogutu, J. O., Schulz-Streeck, T., and Piepho, H.-P. (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *BMC Proceedings*, 6:S10.

Omberg, L., Chaibub Neto, E., Perumal, T. M., Pratap, A., Tediarjo, A., Adams, J., Bloem, B. R., Bot, B. M., Elson, M., Goldman, S. M., Kellen, M. R., Kieburtz, K., Klein, A., Little, M. A., Schneider, R., Suver, C., Tarolli, C., Tanner, C. M., Trister, A. D., Wilbanks, J., Dorsey, E. R., and Mangravite, L. M. (2022). Remote smartphone monitoring of Parkinson's disease and individual response to therapy. *Nature Biotechnology*, 40(4):480–487.

Parikh, R. B., Teeple, S., and Navathe, A. S. (2019). Addressing Bias in Artificial Intelligence in Health Care. *Journal of the American Medical Association*, 322(24):2377–2378.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peloso, G. M. and Lunetta, K. L. (2011). Choice of population structure informative principal components for adjustment in a case-control study. *BMC Genetics*, 12(1):64.

Phukan, J., Elamin, M., Bede, P., Jordan, N., Gallagher, L., Byrne, S., Lynch, C., Pender, N., and Hardiman, O. (2012). The syndrome of cognitive impairment in amyotrophic lateral sclerosis: a population-based study. *Journal of neurology, neurosurgery, and psychiatry*, 83(1):102–8.

Platt, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*.

Provine, W. B. (1971). *The Origins of Theoretical Population Genetics*. University of Chicago Press, Chicago.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., and Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, 81(3):559–575.

Ramos-Carreño, C. (2022). dcor: distance correlation and related E-statistics in Python (0.5.7). Zenodo. https://doi.org/10.5281/zenodo.7043097.

Renton, A. E., Chiò, A., and Traynor, B. J. (2014). State of play in amyotrophic lateral sclerosis genetics. *Nature Neuroscience*, 17(1):17–23.

Renton, A. E., Majounie, E., Waite, A., Simón-Sánchez, J., Rollinson, S., Gibbs, J. R., Schymick, J. C., Laaksovirta, H., van Swieten, J. C., Myllykangas, L., Kalimo, H., Paetau, A., Abramzon, Y., Remes, A. M., Kaganovich, A., Scholz, S. W., Duckworth, J., Ding, J., Harmer, D. W., Hernandez, D. G., Johnson, J. O., Mok, K., Ryten, M., Trabzuni, D., Guerreiro, R. J., Orrell, R. W., Neal, J., Murray, A., Pearson, J., Jansen, I. E., Sondervan, D., Seelaar, H., Blake, D., Young, K., Halliwell, N., Callister, J. B., Toulson, G., Richardson, A., Gerhard, A., Snowden, J., Mann, D., Neary, D., Nalls, M. A., Peuralinna, T., Jansson, L., Isoviita, V.-M., Kaivorinne, A.-L., Hölttä-Vuori, M., Ikonen, E., Sulkava, R., Benatar, M., Wuu, J., Chiò, A., Restagno, G., Borghero, G., Sabatelli, M., ITALSGEN Consortium, Heckerman, D., Rogaeva, E., Zinman, L., Rothstein, J. D., Sendtner, M., Drepper, C., Eichler, E. E., Alkan, C., Abdullaev, Z., Pack, S. D., Dutra, A., Pak, E., Hardy, J., Singleton, A., Williams, N. M., Heutink, P., Pickering-Brown, S., Morris, H. R., Tienari, P. J., and Traynor, B. J. (2011). A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron*, 72(2):257–68.

Ridley, M. (2004). *Evolution*. Wiley-Blackwell, Oxford, 3rd edition.

Ryan, M., Heverin, M., McLaughlin, R. L., and Hardiman, O. (2019). Lifetime Risk and Heritability of Amyotrophic Lateral Sclerosis. *JAMA neurology*, 76(11):1367–1374.

Sankar, P. L. and Parker, L. S. (2017). The precision medicine initiative's all of us research program: an agenda for research on its ethical, legal, and social issues. *Genetics in Medicine*, 19(7):743–750.

Sapra, R. (2014). Using R^2 with caution. *Current Medicine Research and Practice*, 4(3):130–134.

Schmidt, P., Hartung, J., Bennewitz, J., and Hans-Peter, P. (2019). Heritability in plant breeding on a genotype-difference basis. *Genetics*, 212(4):991–1008.

Sheskin, D. J. (2000). *Handbook of parametric and nonparametric statistical procedures, 2nd ed.* Chapman & Hall/CRC, Boca Raton, FL.

Shimmin, C., Sadowski, P., Baldi, P., Weik, E., Whiteson, D., Goul, E., and Søgaard, A. (2017). Decorrelated jet substructure tagging using adversarial neural networks. *Physical Review D*, 96:074034.

Siemiatycki, J. and Thomas, D. C. (1981). Biological models and statistical interactions: an example from multistage carcinogenesis. *International Journal of Epidemiology*, 10:383–387.

Spiess, A.-N. and Neumeyer, N. (2010). An evaluation of R^2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach. *BMC Pharmacology*, 10(1):6.

Steyerberg, E. W. (2019). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer, Switzerland, 2nd edition.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., and Collins, R. (2015). UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):e1001779.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6).

Thomas, A. (2014). *Introducing Genetics: From Mendel to Molecules*. Garland Science, New York.

Tong, H. and Nikoloski, Z. (2021). Machine learning approaches for crop improvement: Leveraging phenotypic and genotypic big data. *Journal of Plant Physiology*, 257:153354.

Torkamani, A., Wineinger, N. E., and Topol, E. J. (2018). The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, 19(9):581–590.

Turkheimer, E. (2000). Three Laws of Behavior Genetics and What They Mean. *Current Directions in Psychological Science*, 9(5):160–164.

Turley, P., Meyer, M. N., Wang, N., Cesarini, D., Hammonds, E., Martin, A. R., Neale, B. M., Rehm, H. L., Wilkins-Haug, L., Benjamin, D. J., Hyman, S., Laibson, D., and Visscher, P. M. (2021). Problems with using polygenic scores to select embryos. *New England Journal of Medicine*, 385(1):78–86.

Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., and Moore, J. H. (2018). Relief-based feature selection: Introduction and review. *Journal of Biomedical Informatics*, 85:189–203.

Vabalas, A., Gowen, E., Poliakoff, E., and Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLOS ONE*, 14(11):e0224365.

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., and Steyerberg, E. W. (2019). Calibration: the Achilles heel of predictive analytics. *BMC Medicine*, 17(1):230.

Van Calster, B., Nieboer, D., Vergouwe, Y., Cock, B. D., Pencina, M. J., and Steyerberg, E. W. (2016). A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of Clinical Epidemiology*, 74:167–176.

Van Rheenen, W., Shatunov, A., Dekker, A. M., McLaughlin, R. L., Diekstra, F. P., Pulit, S. L., van der Spek, R. A. A., Võsa, U., de Jong, S., Robinson, M. R., Yang, J., Fogh, I., van Doormaal, P. T., Tazelaar, G. H. P., Koppers, M., Blokhuis, A. M., Sproviero, W., Jones, A. R., Kenna, K. P., van Eijk, K. R., Harschnitz, O., Schellevis, R. D., Brands, W. J., Medic, J., Menelaou, A., Vajda, A., Ticozzi, N., Lin, K., Rogelj, B., Vrabec, K., Ravnik-Glavač, M., Koritnik, B., Zidar, J., Leonardis, L., Grošelj, L. D., Millecamps, S., Salachas, F., Meininger, V., de Carvalho, M., Pinto, S., Mora, J. S., Rojas-García, R., Polak, M., Chandran, S., Colville, S., Swingler, R., Morrison, K. E., Shaw, P. J., Hardy, J., Orrell, R. W., Pittman, A., Sidle, K., Fratta, P., Malaspina, A., Topp, S., Petri, S., Abdulla, S., Drepper, C., Sendtner, M., Meyer, T., Ophoff, R. A., Staats, K. A., Wiedau-Pazos, M., Lomen-Hoerth, C., Van Deerlin, V. M., Trojanowski, J. Q., Elman, L., McCluskey, L., Basak, A. N., Tunca, C., Hamzeiy, H., Parman, Y., Meitinger, T., Lichtner, P., Radivojkov-Blagojevic, M., Andres, C. R., Maurel, C., Bensimon, G., Landwehrmeyer, B., Brice, A., Payan, C. A. M., Saker-Delye, S., Dürr, A., Wood, N. W., Tittmann, L., Lieb, W., Franke, A., Rietschel, M., Cichon, S., Nöthen, M. M., Amouyel, P., Tzourio, C., Dartigues, J.-F., Uitterlinden, A. G., Rivadeneira, F., Estrada, K., Hofman, A., Curtis, C., Blauw, H. M., van der Kooi, A. J., de Visser, M., Goris, A., Weber, M., Shaw, C. E., Smith, B. N., Pansarasa, O., Cereda, C., Del Bo, R., Comi, G. P., D'Alfonso, S., Bertolin, C., Sorarù, G., Mazzini, L., Pensato, V., Gellera, C., Tiloca, C., Ratti, A., Calvo, A., Moglia, C., Brunetti, M., Arcuti, S., Capozzo, R., Zecca, C., Lunetta, C., Penco, S., Riva, N., Padovani, A., Filosto, M., Muller, B., Stuit, R. J., PARALS Registry, SLALOM Group, SLAP Registry, FALS Sequencing Consortium, SLAGEN Consortium, NNIPPS Study Group, Blair, I., Zhang, K., McCann, E. P., Fifita, J. A., Nicholson, G. A., Rowe, D. B., Pamphlett, R., Kiernan, M. C., Grosskreutz, J., Witte, O. W., Ringer, T., Prell, T., Stubendorff, B., Kurth, I., Hübner, C. A., Leigh, P. N., Casale, F., Chio, A., Beghi, E., Pupillo, E., Tortelli, R., Logroscino, G., Powell, J., Ludolph, A. C., Weishaupt, J. H., Robberecht, W., Van Damme, P., Franke, L., Pers, T. H., Brown, R. H., Glass, J. D., Landers, J. E., Hardiman, O., Andersen, P. M., Corcia, P., Vourc'h, P., Silani, V., Wray, N. R., Visscher, P. M., de Bakker, P. I. W., van Es, M. A., Pasterkamp, R. J., Lewis, C. M., Breen, G., Al-Chalabi, A., van den Berg, L. H., and Veldink, J. H. (2016). Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nature Genetics*, 48(9):1043–8.

Van Rheenen, W., van der Spek, R. A. A., Bakker, M. K., van Vugt, J. J. F. A., Hop, P. J., Zwamborn, R. A. J., de Klein, N., Westra, H.-J., Bakker, O. B., Deelen, P., Shireby, G., Hannon, E., Moisse, M., Baird, D., Restuadi, R., Dolzhenko, E., Dekker, A. M., Gawor, K., Westeneng, H.-J., Tazelaar, G. H. P., van Eijk, K. R., Kooyman, M., Byrne, R. P., Doherty, M., Heverin, M., Al Khleifat, A., Iacoangeli, A., Shatunov, A., Ticozzi, N., Cooper-Knock, J., Smith, B. N., Gromicho, M., Chandran, S., Pal, S., Morrison, K. E., Shaw, P. J., Hardy, J., Orrell, R. W., Sendtner, M., Meyer, T., Başak, N., van der Kooi, A. J., Ratti, A., Fogh, I., Gellera, C., Lauria, G., Corti, S., Cereda, C., Sproviero, D., D'Alfonso, S., Sorarù, G., Siciliano, G., Filosto, M., Padovani, A., Chiò, A., Calvo, A., Moglia, C., Brunetti, M., Canosa, A., Grassano, M., Beghi, E., Pupillo, E., Logroscino, G., Nefussy, B., Osmanovic, A., Nordin, A., Lerner, Y., Zabari, M., Gotkine, M., Baloh, R. H., Bell, S., Vourc'h, P., Corcia, P., Couratier, P., Millecamps, S., Meininger, V., Salachas, F., Mora Pardina, J. S., Assialioui, A., Rojas-García, R., Dion, P. A., Ross, J. P., Ludolph, A. C., Weishaupt, J. H., Brenner, D., Freischmidt, A., Bensimon, G., Brice, A., Durr, A., Payan, C. A. M., Saker-Delye, S., Wood, N. W., Topp, S., Rademakers, R., Tittmann, L., Lieb, W., Franke, A., Ripke, S., Braun, A., Kraft, J., Whiteman, D. C., Olsen, C. M., Uitterlinden, A. G.,

Hofman, A., Rietschel, M., Cichon, S., Nöthen, M. M., Amouyel, P., Comi, G., Riva, N., Lunetta, C., Gerardi, F., Cotelli, M. S., Rinaldi, F., Chiveri, L., Guaita, M. C., Perrone, P., Ceroni, M., Diamanti, L., Ferrarese, C., Tremolizzo, L., Delodovici, M. L., Bono, G., Canosa, A., Manera, U., Vasta, R., Bombaci, A., Casale, F., Fuda, G., Salamone, P., Iazzolino, B., Peotta, L., Cugnasco, P., De Marco, G., Torrieri, M. C., Palumbo, F., Gallone, S., Barberis, M., Sbaiz, L., Gentile, S., Mauro, A., Mazzini, L., De Marchi, F., Corrado, L., D'Alfonso, S., Bertolotto, A., Gionco, M., Leotta, D., Odddenino, E., Imperiale, D., Cavallo, R., Pignatta, P., De Mattei, M., Geda, C., Papurello, D. M., Gusmaroli, G., Comi, C., Labate, C., Ruiz, L., Ferrandi, D., Rota, E., Aguggia, M., Di Vito, N., Meineri, P., Ghiglione, P., Launaro, N., Dotta, M., Di Sapio, A., Giardini, G., Tiloca, C., Peverelli, S., Taroni, F., Pensato, V., Castellotti, B., Comi, G. P., Del Bo, R., Ceroni, M., Gagliardi, S., Corrado, L., Mazzini, L., Raggi, F., Simoncini, C., Lo Gerfo, A., Inghilleri, M., Ferlini, A., Simone, I. L., Passarella, B., Guerra, V., Zoccolella, S., Nozzoli, C., Mundi, C., Leone, M., Zarrelli, M., Tamma, F., Valluzzi, F., Calabrese, G., Boero, G., Rini, A., Traynor, B. J., Singleton, A. B., Mitne Neto, M., Cauchi, R. J., Ophoff, R. A., Wiedau-Pazos, M., Lomen-Hoerth, C., van Deerlin, V. M., Grosskreutz, J., Roediger, A., Gaur, N., Jörk, A., Barthel, T., Theele, E., Ilse, B., Stubendorff, B., Witte, O. W., Steinbach, R., Hübner, C. A., Graff, C., Brylev, L., Fominykh, V., Demeshonok, V., Ataulina, A., Rogelj, B., Koritnik, B., Zidar, J., Ravnik-Glavač, M., Glavač, D., Stević, Z., Drory, V., Povedano, M., Blair, I. P., Kiernan, M. C., Benyamin, B., Henderson, R. D., Furlong, S., Mathers, S., McCombe, P. A., Needham, M., Ngo, S. T., Nicholson, G. A., Pamphlett, R., Rowe, D. B., Steyn, F. J., Williams, K. L., Mather, K. A., Sachdev, P. S., Henders, A. K., Wallace, L., de Carvalho, M., Pinto, S., Petri, S., Weber, M., Rouleau, G. A., Silani, V., Curtis, C. J., Breen, G., Glass, J. D., Brown, R. H., Landers, J. E., Shaw, C. E., Andersen, P. M., Groen, E. J. N., van Es, M. A., Pasterkamp, R. J., Fan, D., Garton, F. C., McRae, A. F., Davey Smith, G., Gaunt, T. R., Eberle, M. A., Mill, J., McLaughlin, R. L., Hardiman, O., Kenna, K. P., Wray, N. R., Tsai, E., Runz, H., Franke, L., Al-Chalabi, A., Van Damme, P., van den Berg, L. H., and Veldink, J. H. (2021). Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. *Nature Genetics*, 53(12):1636–1648.

VanRaden, P. (2008). Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science*, 91(11):4414–4423.

Varma, S. and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(1):91.

Verweij, K. J. H., Mosing, M. A., Zietsch, B. P., and Medland, S. E. (2012). Estimating heritability from twin studies. In *Methods in Molecular Biology*, pages 151–170. Humana Press.

Vilhjálmsson, B. J. and Nordborg, M. (2013). The nature of confounding in genome-wide association studies. *Nature Reviews Genetics*, 14(1):1–2.

Visscher, P. M., Hill, W. G., and Wray, N. R. (2008). Heritability in the genomics era — concepts and misconceptions. *Nature Reviews Genetics*, 9(4):255–266.

Visscher, P. M., Medland, S. E., Ferreira, M. A. R., Morley, K. I., Zhu, G., Cornes, B. K., Montgomery, G. W., and Martin, N. G. (2006). Assumption-Free Estimation of Heritability

from Genome-Wide Identity-by-Descent Sharing between Full Siblings. *PLoS Genetics*, 2(3):e41.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics*, 101(1):5–22.

Vokinger, K. N., Feuerriegel, S., and Kesselheim, A. S. (2021). Mitigating bias in machine learning for medicine. *Nature Communications Medicine*, 1(1):25.

Wainschtein, P., Jain, D., Zheng, Z., Aslibekyan, S., Becker, D., Bi, W., Brody, J., Carlson, J. C., Correa, A., Du, M. M., Fernandez-Rhodes, L., Ferrier, K. R., Graff, M., Guo, X., He, J., Heard-Costa, N. L., Highland, H. M., Hirschhorn, J. N., Howard-Claudio, C. M., Isasi, C. R., Jackson, R., Jiang, J., Joehanes, R., Justice, A. E., Kalyani, R. R., Kardia, S., Lange, E., LeBoff, M., Lee, S., Li, X., Li, Z., Lim, E., Lin, D., Lin, X., Liu, S., Lu, Y., Manson, J., Martin, L., McHugh, C., Mikulla, J., Musani, S. K., Ng, M., Nickerson, D., Palmer, N., Perry, J., Peters, U., Preuss, M., Qi, Q., Raffield, L., Rasmussen-Torvik, L., Reiner, A., Russell, E. M., Sitlani, C., Smith, J., Spracklen, C. N., Wang, T., Wang, Z., Wessel, J., Xu, H., Yaser, M., Yoneyama, S., Young, K. A., Zhang, J., Zhang, X., Zhou, H., Zhu, X., Zoellner, S., Abe, N., Abecasis, G., Aguet, F., Almasy, L., Alonso, A., Ament, S., Anderson, P., Anugu, P., Applebaum-Bowden, D., Ardlie, K., Arking, D., Ashley-Koch, A., Assimes, T., Auer, P., Avramopoulos, D., Ayas, N., Balasubramanian, A., Barnard, J., Barnes, K., Barr, R. G., Barron-Casella, E., Barwick, L., Beaty, T., Beck, G., Becker, L., Beer, R., Beitelshees, A., Benjamin, E., Benos, T., Bezerra, M., Bielak, L., Bis, J., Blackwell, T., Blangero, J., Bowden, D. W., Bowler, R., Broeckel, U., Broome, J., Brown, D., Bunting, K., Burchard, E., Bustamante, C., Buth, E., Cade, B., Cardwell, J., Carey, V., Carrier, J., Carson, A., Carty, C., Casaburi, R., Romero, J. P. C., Casella, J., Castaldi, P., Chaffin, M., Chang, C., Chang, Y.-C., Chavan, S., Chen, B.-J., Chen, W.-M., Cho, M., Choi, S. H., Chuang, L.-M., Chung, R.-H., Clish, C., Comhair, S., Conomos, M., Cornell, E., Crandall, C., Crapo, J., Curran, J., Curtis, J., Custer, B., Damcott, C., Darbar, D., David, S., Davis, C., Daya, M., de las Fuentes, L., de Vries, P., DeBaun, M., Deka, R., DeMeo, D., Devine, S., Dinh, H., Doddapaneni, H., Duan, Q., Dugan-Perez, S., Duggirala, R., Durda, J. P., Dutcher, S. K., Eaton, C., Ekunwe, L., El Boueiz, A., Emery, L., Erzurum, S., Farber, C., Farek, J., Fingerlin, T., Flickinger, M., Franceschini, N., Frazar, C., Fu, M., Fullerton, S. M., Fulton, L., Gabriel, S., Gan, W., Gao, S., Gao, Y., Gass, M., Geiger, H., Gelb, B., Geraci, M., Germer, S., Gerszten, R., Ghosh, A., Gibbs, R., Gignoux, C., Gladwin, M., Glahn, D., Gogarten, S., Gong, D.-W., Goring, H., Graw, S., Gray, K. J., Grine, D., Gross, C., Gu, C. C., Guan, Y., Gupta, N., Haas, D. M., Haessler, J., Hall, M., Han, Y., Hanly, P., Harris, D., Hawley, N. L., Heavner, B., Herrington, D., Hersh, C., Hidalgo, B., Hixson, J., Hobbs, B., Hokanson, J., Hong, E., Hoth, K., Hsiung, C. A., Hu, J., Hung, Y.-J., Huston, H., Hwu, C. M., Irvin, M. R., Jaquish, C., Johnsen, J., Johnson, A., Johnson, C., Johnston, R., Jones, K., Kang, H. M., Kaplan, R., Kelly, S., Kenny, E., Kessler, M., Khan, A., Khan, Z., Kim, W., Kimoff, J., Kinney, G., Konkle, B., Kramer, H., Lange, C., Lee, J., Lee, S., Lee, W.-J., LeFaive, J., Levine, D., Levy, D., Lewis, J., Li, X., Li, Y., Lin, H., Lin, H., Liu, Y., Liu, Y., Lunetta, K., Luo, J., Magalang, U., Mahaney, M., Make, B., Manichaikul, A., Manning, A., Marton, M., Mathai, S., May, S., McArdle, P., McFarland, S., McGoldrick, D., McNeil, B., Mei, H., Meigs, J., Menon, V., Mestroni, L., Metcalf, G., Meyers, D. A., Mignot, E., Mikulla, J., Min, N., Minear, M., Minster, R. L., Moll, M., Momin, Z., Montasser, M. E., Montgomery, C., Muzny, D., Mychaleckyj, J. C., Nadkarni,

G., Naik, R., Naseri, T., Natarajan, P., Nekhai, S., Nelson, S. C., Neltner, B., Nessner, C., Nkechinyere, O., O'Connor, T., Ochs-Balcom, H., Okwuonu, G., Pack, A., Paik, D. T., Palmer, N., Pankow, J., Papanicolaou, G., Parker, C., Peloso, G., Peralta, J. M., Perez, M., Peyser, P., Phillips, L. S., Pleiness, J., Pollin, T., Post, W., Becker, J. P., Boorgula, M. P., Qasba, P., Qiao, D., Qin, Z., Rafaels, N., Rajendran, M., Rao, D. C., Ratan, A., Reed, R., Reeves, C., Reupena, M. S., Rice, K., Robillard, R., Robine, N., Roselli, C., Ruczinski, I., Runnels, A., Russell, P., Ruuska, S., Ryan, K., Sabino, E. C., Saleheen, D., Salimi, S., Salvi, S., Salzberg, S., Sandow, K., Sankaran, V. G., Santibanez, J., Schwander, K., Schwartz, D., Sciurba, F., Seidman, C., Seidman, J., Sheehan, V., Sherman, S. L., Shetty, A., Shetty, A., Sheu, W. H.-H., Silver, B., Silverman, E., Skomro, R., Smith, A. V., Smith, J., Smith, T., Smoller, S., Snively, B., Snyder, M., Sofer, T., Sotoodehnia, N., Stilp, A. M., Storm, G., Streeten, E., Su, J. L., Sung, Y. J., Sylvia, J., Szpiro, A., Taliun, D., Tang, H., Taub, M., Taylor, K. D., Taylor, M., Taylor, S., Telen, M., Thornton, T. A., Threlkeld, M., Tinker, L., Tirschwell, D., Tishkoff, S., Tiwari, H., Tong, C., Tracy, R., Tsai, M., Vaidya, D., Van Den Berg, D., VandeHaar, P., Vrieze, S., Walker, T., Wallace, R., Walts, A., Wang, F. F., Wang, H., Wang, J., Watson, K., Watt, J., Weeks, D. E., Weinstock, J., Weiss, S. T., Weng, L.-C., Willer, C., Williams, K., Williams, L. K., Wilson, C., Wilson, J., Winterkorn, L., Wong, Q., Wu, J., Xu, H., Yang, I., Yu, K., Zekavat, S. M., Zhang, Y., Zhao, S. X., Zhao, W., Zody, M., Cupples, L. A., Shadyab, A. H., McKnight, B., Shoemaker, B. M., Mitchell, B. D., Psaty, B. M., Kooperberg, C., Liu, C.-T., Albert, C. M., Roden, D., Chasman, D. I., Darbar, D., Lloyd-Jones, D. M., Arnett, D. K., Regan, E. A., Boerwinkle, E., Rotter, J. I., O'Connell, J. R., Yanek, L. R., de Andrade, M., Allison, M. A., McDonald, M.-L. N., Chung, M. K., Fornage, M., Chami, N., Smith, N. L., Ellinor, P. T., Vasan, R. S., Mathias, R. A., Loos, R. J. F., Rich, S. S., Lubitz, S. A., Heckbert, S. R., Redline, S., Guo, X., Chen, Y. D. I., Laurie, C. A., Hernandez, R. D., McGarvey, S. T., Goddard, M. E., Laurie, C. C., North, K. E., Lange, L. A., Weir, B. S., Yengo, L., Yang, J., and Visscher, P. M. (2022). Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nature Genetics*, 54(3):263–273.

Walsh, B. and Lynch, M. (1998). *Genetics and Analysis of Quantitative Traits*. Oxford Univeristy Press, Oxford.

Wei, W.-H., Hemani, G., and Haley, C. S. (2014). Detecting epistasis in human complex traits. *Nature Reviews Genetics*, 15(11):722–733.

Whalen, S., Schreiber, J., Noble, W. S., and Pollard, K. S. (2022). Navigating the pitfalls of applying machine learning in genomics. *Nature Reviews Genetics*, 23(3):169–181.

Wolff, R. F., Moons, K. G., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., and Mallett, S. (2019). PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Annals of Internal Medicine*, 170(1):51.

Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E., and Visscher, P. M. (2013). Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics*, 14(7):507–515.

Wright, S. (1934). An Analysis of Variability in Number of Digits in an Inbred Strain of Guinea Pigs. *Genetics*, 19(6):506–36.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., and Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565–569.

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics*, 88(1):76–82.

Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., and Price, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, 46(2):100–6.

Yengo, L., Vedantam, S., Marouli, E., Sidorenko, J., Bartell, E., Sakaue, S., Graff, M., Eliasen, A. U., Jiang, Y., Raghavan, S., Miao, J., Arias, J. D., Graham, S. E., Mukamel, R. E., Spracklen, C. N., Yin, X., Chen, S.-H., Ferreira, T., Highland, H. H., Ji, Y., Karaderi, T., Lin, K., Lüll, K., Malden, D. E., Medina-Gomez, C., Machado, M., Moore, A., Rüeger, S., Sim, X., Vrieze, S., Ahluwalia, T. S., Akiyama, M., Allison, M. A., Alvarez, M., Andersen, M. K., Ani, A., Appadurai, V., Arbeeva, L., Bhaskar, S., Bielak, L. F., Bollepalli, S., Bonnycastle, L. L., Bork-Jensen, J., Bradfield, J. P., Bradford, Y., Braund, P. S., Brody, J. A., Burgdorf, K. S., Cade, B. E., Cai, H., Cai, Q., Campbell, A., Cañadas-Garre, M., Catamo, E., Chai, J.-F., Chai, X., Chang, L.-C., Chang, Y.-C., Chen, C.-H., Chesi, A., Choi, S. H., Chung, R.-H., Cocca, M., Concas, M. P., Couture, C., Cuellar-Partida, G., Danning, R., Daw, E. W., Degenhard, F., Delgado, G. E., Delitala, A., Demirkan, A., Deng, X., Devineni, P., Dietl, A., Dimitriou, M., Dimitrov, L., Dorajoo, R., Ekici, A. B., Engmann, J. E., Fairhurst-Hunter, Z., Farmaki, A.-E., Faul, J. D., Fernandez-Lopez, J.-C., Forer, L., Francescatto, M., Freitag-Wolf, S., Fuchsberger, C., Galesloot, T. E., Gao, Y., Gao, Z., Geller, F., Giannakopoulou, O., Giulianini, F., Gjesing, A. P., Goel, A., Gordon, S. D., Gorski, M., Grove, J., Guo, X., Gustafsson, S., Haessler, J., Hansen, T. F., Havulinna, A. S., Haworth, S. J., He, J., Heard-Costa, N., Hebbar, P., Hindy, G., Ho, Y.-L. A., Hofer, E., Holliday, E., Horn, K., Hornsby, W. E., Hottenga, J.-J., Huang, H., Huang, J., Huerta-Chagoya, A., Huffman, J. E., Hung, Y.-J., Huo, S., Hwang, M. Y., Iha, H., Ikeda, D. D., Isono, M., Jackson, A. U., Jäger, S., Jansen, I. E., Johansson, I., Jonas, J. B., Jonsson, A., Jørgensen, T., Kalafati, I.-P., Kanai, M., Kanoni, S., Kårhus, L. L., Kasturiratne, A., Katsuya, T., Kawaguchi, T., Kember, R. L., Kentistou, K. A., Kim, H.-N., Kim, Y. J., Kleber, M. E., Knol, M. J., Kurbasic, A., Lauzon, M., Le, P., Lea, R., Lee, J.-Y., Leonard, H. L., Li, S. A., Li, X., Li, X., Liang, J., Lin, H., Lin, S.-Y., Liu, J., Liu, X., Lo, K. S., Long, J., Lores-Motta, L., Luan, J., Lyssenko, V., Lyytikäinen, L.-P., Mahajan, A., Mamakou, V., Mangino, M., Manichaikul, A., Marten, J., Mattheisen, M., Mavarani, L., McDaid, A. F., Meidtner, K., Melendez, T. L., Mercader, J. M., Milaneschi, Y., Miller, J. E., Millwood, I. Y., Mishra, P. P., Mitchell, R. E., Møllehave, L. T., Morgan, A., Mucha, S., Munz, M., Nakatochi, M., Nelson, C. P., Nethander, M., Nho, C. W., Nielsen, A. A., Nolte, I. M., Nongmaithem, S. S., Noordam, R., Ntalla, I., Nutile, T., Pandit, A., Christofidou, P., Pärna, K., Pauper, M., Petersen, E. R. B., Petersen, L. V., Pitkänen, N., Polašek, O., Poveda, A., Preuss, M. H., Pyarajan, S., Raffield, L. M., Rakugi, H., Ramirez, J., Rasheed, A., Raven, D., Rayner, N. W., Riveros, C., Rohde, R., Ruggiero, D., Ruotsalainen, S. E., Ryan, K. A., Sabater-Lleal, M., Saxena, R., Scholz, M., Sendamarai, A., Shen, B., Shi, J., Shin, J. H., Sidore, C., Sitlani, C. M., Slieker, R. C., Smit, R. A. J., Smith, A. V., Smith, J. A., Smyth, L. J., Southam, L., Steinthorsdottir, V., Sun, L., Takeuchi, F., Tallapragada, D. S. P., Taylor, K. D., Tayo, B. O., Tcheandjieu, C., Terzikhan, N., Tesolin, P., Teumer, A., Theusch,

E., Thompson, D. J., Thorleifsson, G., Timmers, P. R. H. J., Trompet, S., Turman, C., Vaccargiu, S., van der Laan, S. W., van der Most, P. J., van Klinken, J. B., van Setten, J., Verma, S. S., Verweij, N., Veturi, Y., Wang, C. A., Wang, C., Wang, L., Wang, Z., Warren, H. R., Bin Wei, W., Wickremasinghe, A. R., Wielscher, M., Wiggins, K. L., Winsvold, B. S., Wong, A., Wu, Y., Wuttke, M., Xia, R., Xie, T., Yamamoto, K., Yang, J., Yao, J., Young, H., Yousri, N. A., Yu, L., Zeng, L., Zhang, W., Zhang, X., Zhao, J.-H., Zhao, W., Zhou, W., Zimmermann, M. E., Zoledziewska, M., Adair, L. S., Adams, H. H. H., Aguilar-Salinas, C. A., Al-Mulla, F., Arnett, D. K., Asselbergs, F. W., Åsvold, B. O., Attia, J., Banas, B., Bandinelli, S., Bennett, D. A., Bergler, T., Bharadwaj, D., Biino, G., Bisgaard, H., Boerwinkle, E., Böger, C. A., Bønnelykke, K., Boomsma, D. I., Børglum, A. D., Borja, J. B., Bouchard, C., Bowden, D. W., Brandslund, I., Brumpton, B., Buring, J. E., Caulfield, M. J., Chambers, J. C., Chandak, G. R., Chanock, S. J., Chaturvedi, N., Chen, Y.-D. I., Chen, Z., Cheng, C.-Y., Christophersen, I. E., Ciullo, M., Cole, J. W., Collins, F. S., Cooper, R. S., Cruz, M., Cucca, F., Cupples, L. A., Cutler, M. J., Damrauer, S. M., Dantoft, T. M., de Borst, G. J., de Groot, L. C. P. G. M., De Jager, P. L., de Kleijn, D. P. V., Janaka de Silva, H., Dedoussis, G. V., den Hollander, A. I., Du, S., Easton, D. F., Elders, P. J. M., Eliassen, A. H., Ellinor, P. T., Elmståhl, S., Erdmann, J., Evans, M. K., Fatkin, D., Feenstra, B., Feitosa, M. F., Ferrucci, L., Ford, I., Fornage, M., Franke, A., Franks, P. W., Freedman, B. I., Gasparini, P., Gieger, C., Girotto, G., Goddard, M. E., Golightly, Y. M., Gonzalez-Villalpando, C., Gordon-Larsen, P., Grallert, H., Grant, S. F. A., Grarup, N., Griffiths, L., Gudnason, V., Haiman, C., Hakonarson, H., Hansen, T., Hartman, C. A., Hattersley, A. T., Hayward, C., Heckbert, S. R., Heng, C.-K., Hengstenberg, C., Hewitt, A. W., Hishigaki, H., Hoyng, C. B., Huang, P. L., Huang, W., Hunt, S. C., Hveem, K., Hyppönen, E., Iacono, W. G., Ichihara, S., Ikram, M. A., Isasi, C. R., Jackson, R. D., Jarvelin, M.-R., Jin, Z.-B., Jöckel, K.-H., Joshi, P. K., Jousilahti, P., Jukema, J. W., Kähönen, M., Kamatani, Y., Kang, K. D., Kaprio, J., Kardia, S. L. R., Karpe, F., Kato, N., Kee, F., Kessler, T., Khera, A. V., Khor, C. C., Kiemeney, L. A. L. M., Kim, B.-J., Kim, E. K., Kim, H.-L., Kirchhof, P., Kivimaki, M., Koh, W.-P., Koistinen, H. A., Kolovou, G. D., Kooner, J. S., Kooperberg, C., Köttgen, A., Kovacs, P., Kraaijeveld, A., Kraft, P., Krauss, R. M., Kumari, M., Kutalik, Z., Laakso, M., Lange, L. A., Langenberg, C., Launer, L. J., Le Marchand, L., Lee, H., Lee, N. R., Lehtimäki, T., Li, H., Li, L., Lieb, W., Lin, X., Lind, L., Linneberg, A., Liu, C.-T., Liu, J., Loeffler, M., London, B., Lubitz, S. A., Lye, S. J., Mackey, D. A., Mägi, R., Magnusson, P. K. E., Marcus, G. M., Vidal, P. M., Martin, N. G., März, W., Matsuda, F., McGarrah, R. W., McGue, M., McKnight, A. J., Medland, S. E., Mellström, D., Metspalu, A., Mitchell, B. D., Mitchell, P., Mook-Kanamori, D. O., Morris, A. D., Mucci, L. A., Munroe, P. B., Nalls, M. A., Nazarian, S., Nelson, A. E., Neville, M. J., Newton-Cheh, C., Nielsen, C. S., Nöthen, M. M., Ohlsson, C., Oldehinkel, A. J., Orozco, L., Pahkala, K., Pajukanta, P., Palmer, C. N. A., Parra, E. J., Pattaro, C., Pedersen, O., Pennell, C. E., Penninx, B. W. J. H., Perusse, L., Peters, A., Peyser, P. A., Porteous, D. J., Posthuma, D., Power, C., Pramstaller, P. P., Province, M. A., Qi, Q., Qu, J., Rader, D. J., Raitakari, O. T., Ralhan, S., Rallidis, L. S., Rao, D. C., Redline, S., Reilly, D. F., Reiner, A. P., Rhee, S. Y., Ridker, P. M., Rienstra, M., Ripatti, S., Ritchie, M. D., Roden, D. M., Rosendaal, F. R., Rotter, J. I., Rudan, I., Rutters, F., Sabanayagam, C., Saleheen, D., Salomaa, V., Samani, N. J., Sanghera, D. K., Sattar, N., Schmidt, B., Schmidt, H., Schmidt, R., Schulze, M. B., Schunkert, H., Scott, L. J., Scott, R. J., Sever, P., Shiroma, E. J., Shoemaker, M. B., Shu, X.-O., Simonsick, E. M., Sims, M., Singh, J. R., Singleton, A. B., Sinner, M. F., Smith, J. G., Snieder, H., Spector, T. D., Stampfer, M. J., Stark, K. J., Strachan, D. P., 't Hart, L. M., Tabara, Y., Tang, H., Tardif, J.-C., Thanaraj, T. A., Timpson, N. J., Tönjes,

A., Tremblay, A., Tuomi, T., Tuomilehto, J., Tusié-Luna, M.-T., Uitterlinden, A. G., van Dam, R. M., van der Harst, P., Van der Velde, N., van Duijn, C. M., van Schoor, N. M., Vitart, V., Völker, U., Vollenweider, P., Völzke, H., Wacher-Rodarte, N. H., Walker, M., Wang, Y. X., Wareham, N. J., Watanabe, R. M., Watkins, H., Weir, D. R., Werge, T. M., Widen, E., Wilkens, L. R., Willemsen, G., Willett, W. C., Wilson, J. F., Wong, T.-Y., Woo, J.-T., Wright, A. F., Wu, J.-Y., Xu, H., Yajnik, C. S., Yokota, M., Yuan, J.-M., Zeggini, E., Zemel, B. S., Zheng, W., Zhu, X., Zmuda, J. M., Zonderman, A. B., Zwart, J.-A., Partida, G. C., Sun, Y., Croteau-Chonka, D., Vonk, J. M., Chanock, S., Le Marchand, L., Chasman, D. I., Cho, Y. S., Heid, I. M., McCarthy, M. I., Ng, M. C. Y., O'Donnell, C. J., Rivadeneira, F., Thorsteinsdottir, U., Sun, Y. V., Tai, E. S., Boehnke, M., Deloukas, P., Justice, A. E., Lindgren, C. M., Loos, R. J. F., Mohlke, K. L., North, K. E., Stefansson, K., Walters, R. G., Winkler, T. W., Young, K. L., Loh, P.-R., Yang, J., Esko, T., Assimes, T. L., Auton, A., Abecasis, G. R., Willer, C. J., Locke, A. E., Berndt, S. I., Lettre, G., Frayling, T. M., Okada, Y., Wood, A. R., Visscher, P. M., and Hirschhorn, J. N. (2022). A saturated map of common genetic variants associated with human height. *Nature*.

Yin, B., Balvert, M., van der Spek, R. A. A., Dutilh, B. E., Bohté, S., Veldink, J., and Schönhuth, A. (2019). Using the structure of genome data in the design of deep neural networks for predicting amyotrophic lateral sclerosis from genotype. *Bioinformatics (Oxford, England)*, 35(14):i538–i547.

Zhu, J., Sun, S., and Zhou, X. (2021). SPARK-X: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome Biology*, 22(1):184.

Zhu, Z., Bakshi, A., Vinkhuyzen, A. A., Hemani, G., Lee, S. H., Nolte, I. M., van Vliet-Ostaptchouk, J. V., Snieder, H., Esko, T., Milani, L., Mägi, R., Metspalu, A., Hill, W. G., Weir, B. S., Goddard, M. E., Visscher, P. M., and Yang, J. (2015). Dominance Genetic Variation Contributes Little to the Missing Heritability for Human Complex Traits. *The American Journal of Human Genetics*, 96(3):377–385.

Zou, Z.-Y., Zhou, Z.-R., Che, C.-H., Liu, C.-Y., He, R.-L., and Huang, H.-P. (2017). Genetic epidemiology of amyotrophic lateral sclerosis: a systematic review and meta-analysis. *Journal of Neurology, Neurosurgery, and Psychiatry*, 88(7):540–549.

Appendix A

S1 Grid search

Hyperparameters

The following hyperparameters were optimized according to predictive performance and time taken during the *Arabidopsis thaliana* experiments described in Chapter 2:

Feed-forward Neural Network*: Number of Units, Learning Rate, Optimizer, Batch size, Network Shape, L1/L2 Regularization, Dropout rate, Number of Hidden Layers, Epochs, Kernel Initializer, Activation Function, Decay Rate, Maximum Norm Value, Learning Rate Decay*

Convolutional Neural Network*: Number of Units, Learning Rate, Optimizer, Batch Size, Network Shape, L1/L2 Regularization, Dropout Rate, Number of Hidden Layers, Epochs, Kernel Initializer, Activation Function, Filters, Kernel, Pool, Strides, Decay Rate, Maximum Norm Value, Learning Rate Decay*

Support Vector Machine (Linear)*: C, Epsilon, Loss*

Support Vector Machine (Non-Linear)*: C, Gamma, Epsilon, Loss, Kernel, Degree, Cache Size, Optimization Tolerance, Shrinking*

LASSO*: Alpha, Selection, Optimization Tolerance, Maximum Iterations*

Ridge Regression*: Alpha, Optimization Tolerance*

Random Forests*: Number of Estimators, Maximum Depth, Maximum Number of Features, Bootstrapping, Maximum Number of Samples, Minimum Number of Samples to Split, Minimum Leaf samples, Maximum Number of Leaf Nodes*

The following hyperparameters were optimized according to predictive performance and time taken during the ALS experiments described in Chapter 3:

Feed-forward Neural Network*: Number of Units, Learning Rate, Optimizer, Batch size, Network Shape, L1/L2 Regularization, Dropout rate, Number of Hidden Layers, Epochs, Kernel Initializer, Activation Function, Decay Rate, Maximum Norm Value, Learning Rate Decay*

Convolutional Neural Network*: Number of Units, Learning Rate, Optimizer, Batch Size, Network Shape, L1/L2 Regularization, Dropout Rate, Number of Hidden Layers, Epochs, Kernel Initializer, Activation Function, Filters, Kernel, Pool, Strides, Decay Rate, Maximum Norm Value, Learning Rate Decay*

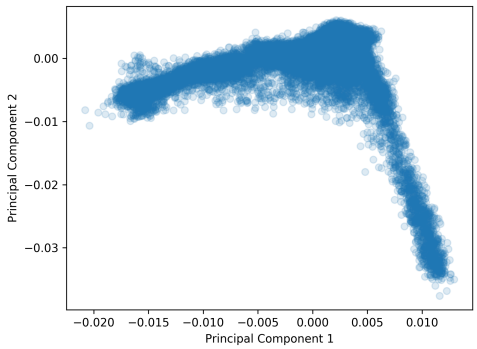
Support Vector Machine (Non-Linear)*: C, Gamma, Epsilon, Loss, Kernel, Degree, Cache Size, Optimization Tolerance, Shrinking*

LASSO*: Alpha, Selection, Optimization Tolerance, Maximum Iterations*

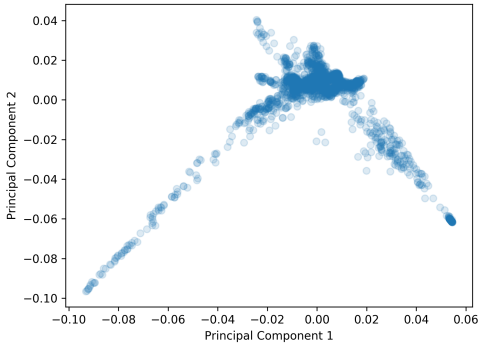
Ridge Regression*: Alpha, Optimization Tolerance*

Random Forests*: Number of Estimators, Maximum Depth, Maximum Number of Features, Bootstrapping, Maximum Number of Samples, Minimum Number of Samples to Split, Minimum Leaf samples, Maximum Number of Leaf Nodes*

S2 Principal Component Analysis



(a) **ALS Case/Control Data:** The first two principal components of variation in this data are plotted (calculated using 131,883 pruned SNPs as per Section 4.2.2).



(b) ***Arabidopsis thaliana*:** The first two principal components of variation in this data are plotted (calculated using the GRMs as described in Section 4.2.2).

Fig. A.1 **Principal Component Plots for ALS and *Arabidopsis Thaliana* Data**[16].