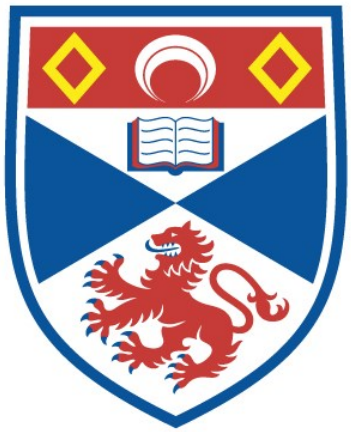
Biologically-informed interpretable deep learning techniques for BMI prediction and gene interaction detection

Chloe Hequet

A thesis submitted for the degree of PhD
at the
University of St Andrews



2024

Candidate's declaration

I, Chloe Catherine Hequet, do hereby certify that this thesis, submitted for the degree of PhD, which is approximately 39,051 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for any degree. I confirm that any appendices included in my thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

I was admitted as a research student at the University of St Andrews in January 2019.

I received funding from an organisation or institution and have acknowledged the funder(s) in the full text of my thesis.

Date Signature of candidate

Supervisor's declaration

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree. I confirm that any appendices included in the thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

Date Signature of supervisor

Permission for publication

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand, unless exempt by an award of an embargo as requested below, that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that this thesis will be electronically accessible for personal or research use and that the library has the right to migrate this thesis into new electronic forms as required to ensure continued access to the thesis.

I, Chloe Catherine Hequet, confirm that my thesis does not contain any third-party material that requires copyright clearance.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

Printed copy

No embargo on print copy.

Electronic copy

No embargo on electronic copy.

Date Signature of candidate

Date Signature of supervisor

Underpinning Research Data or Digital Outputs

Candidate's declaration

I, Chloe Catherine Hequet, understand that by declaring that I have original research data or digital outputs, I should make every effort in meeting the University's and research funders' requirements on the deposit and sharing of research data or research digital outputs.

Date Signature of candidate

Permission for publication of underpinning research data or digital outputs

We understand that for any original research data or digital outputs which are deposited, we are giving permission for them to be made available for use in accordance with the requirements of the University and research funders, for the time being in force.

We also understand that the title and the description will be published, and that the underpinning research data or digital outputs will be electronically accessible for use in accordance with the license specified at the point of deposit, unless exempt by award of an embargo as requested below.

The following is an agreed request by candidate and supervisor regarding the publication of underpinning research data or digital outputs:

No embargo on underpinning research data or digital outputs.

Date Signature of candidate

Date Signature of supervisor

Abstract

The analysis of genetic point mutations at the population level can offer insights into the genetic basis of human traits, which in turn could potentially lead to new diagnostic and treatment options for heritable diseases. However, existing genetic data analysis methods tend to rely on simplifying assumptions that ignore nonlinear interactions between variants. The ability to model and describe nonlinear genetic interactions could lead to both improved trait prediction and enhanced understanding of the underlying biology. Deep Learning models offer the possibility of automatically learning complex nonlinear genetic architectures, but it is currently unclear how best to optimise them for genetic data. It is also essential that any models be able to "explain" what they have learned in order for them to be used for genetic discovery or clinical applications, which can be difficult due to the black-box nature of DL predictors.

This thesis addresses a number of methodological gaps in applying explainable DL models end-to-end on variant-level genetic data. We propose novel methods for encoding genetic data for deep learning applications and show that feature encodings designed specifically for genetic variants offer the possibility of improved model efficiency and performance. We then benchmark a variety of models for the prediction of Body Mass Index using data from the UK Biobank, yielding insights into DL performance in this domain. We then propose a series of novel DL model interpretation methods with features optimised for biological insights. We first show how these can be used to validate that the network has automatically replicated existing knowledge, and then illustrate their ability to detect complex nonlinear genetic interactions that influence BMI in our cohort. Overall, we show that DL model training and interpretation procedures that have been optimised for genetic data can be used to yield new insights into disease aetiology.

Acknowledgements

I would like to take this opportunity to express my enormous gratitude to my supervisor Dr Juan Ye, whose never ending support and guidance have been invaluable through the ups and downs of this project. I feel so incredibly lucky to have had you as a supervisor. Thank you for always being a listening ear and for staying excited about my research every step of the way.

Thank you to Dr Silvia Paracchini for really sparking my interest in genetics and always giving me fascinating things to think about. Your guidance on the biological side of my research has been completely indispensable, and this would have been a very different project without your input. Thank you also to Dr John Thomson for conceiving the idea for this project with me and getting me underway on this exciting journey.

Thank you to Theo and Sam for always being willing to listen to an impromptu talk about genetics or deep learning (or both) over dinner. Thank you to my family for always believing in me and having my back, even when the going got tough. And finally, thank you to my dogs for their extensive barking during video meetings, which really contributed a lot.

This work was supported by funding from the School of Computer Science at the University of St Andrews. I am so grateful to the School of Computer Science for providing me with the opportunity to undertake this research project.

Acronyms

AI Artificial Intelligence.

BMI Body Mass Index.

BNN Bayesian Neural Network.

CNN Convolutional Neural Network.

DL Deep Learning.

DNN Deep Neural Network.

GSA Gene Set Analysis.

GWAS Genome Wide Association Study.

LD Linkage Disequilibrium.

LIME Local Interpretable Model-Agnostic Explanation.

LRP Layerwise Relevance Propagation.

MAE Mean Absolute Error.

MDR Multifactor Dimensionality Reduction.

ML Machine Learning.

MSE Mean Squared Error.

NGS Next Generation Sequencing.

NN Neural Network.

PRS Polygenic Risk Score.

SNP Single Nucleotide Polymorphism.

SVE Standard Variable Encoding.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

We are living in an era of unprecedented access to genetic data. Thanks to Next Generation Sequencing (NGS) methods, it is cheaper and easier than ever before to sequence the genome, and vast reserves of genetic data are becoming available for research [112]. Analysis of this new data has already led to scientific breakthroughs that have changed many of our existing ideas about the function of the genome and how it influences human traits and development [264]. We are also moving in the direction of a more precise clinical model of disease [12]. Datasets containing the genetic information of hundreds of thousands of individuals are now available to researchers [56]. This explosion of data means that it may now be possible to study phenomena that seemed completely impenetrable to generations of previous geneticists. However, many of the established genetic data analytic methods were not designed for this new era, and are not well-suited to very large datasets [4]. There is now a need to develop specialised tools to mine useful information from the new repositories of Genetic Big Data [293].

Despite recent advances, genetics as a scientific field is still relatively new, and a lot of fundamental questions remain unanswered [92]. Understanding how genetics contributes to phenotype and interacts with the environment could lead to crucial insights into disease aetiology, potentially informing the development of new treatments and clinical tools [116]. In spite of this, the genetic basis of many common heritable traits is very poorly understood [333]. For most diseases, genetic research has so far not been able to pinpoint a single causal gene of large effect size [60]. As a consequence the field has moved away from narrow-focus candidate gene studies,

with methods such as GWAS scrutinising the entire genome at once for association signals [306]. Tools such as Polygenic Risk Scores [152] and Gene Set Analysis [115] aim to leverage GWAS results to model phenotype in terms of genetics and gain new insights into underlying biological pathways. These methods have further extended the utility of the common, low penetrance disease-associated variants frequently discovered by GWAS [166]. Larger datasets have also increased the power of genome-wide methods to uncover smaller and more nuanced effects [130]. However, these approaches rely on a simplified model of the genome, in which variants are uncorrelated and interact only in a linear additive fashion [137]. This means they unavoidably fall short when phenotypic variance is influenced by complex nonlinear genetic effects, and thus will never be able to accurately capture the entire genetic architecture to a phenotype [221, 326]. It is clear that moving towards more finetuned genetic insights will require a sophisticated model, one that considers multiple loci at a time, and is capable of modelling interactions between features [282]. But there is a difficulty here: in order to design such a sophisticated model, we would ideally need access to the kind of detailed biological prior knowledge that we are hoping to learn from the model itself. In genetics, this kind of domain expertise does not yet exist. The ability to build a model without access to this information beforehand requires a nonparametric approach that can automatically extract useful information from the data without human input.

In recent years, Deep learning has been making major advances across a variety of disparate fields [315]. Deep learning builds on machine learning approaches, but differs greatly in the amount of domain expertise required to engineer a successful prediction model. Deep learning models are able to use representation learning to automatically learn intricate patterns without explicit parametrization and are particularly well-suited to handling big, complex, high-dimensional data [148]. They could thus offer new capabilities for learning complex patterns from the large genetic datasets generated by NGS methods, but the question of how to design deep neural networks to predict phenotype from genetic data remains open [17]. Furthermore, a major pitfall of DL models across different domains and applications is that their internal logic is neither visible nor reconstructible to the user, meaning they by default provide no explanation for their predictions and are effectively a "black box" [100]. Even if a successful DL predictor for genetic data was constructed and trained, this lack of explanation would both foil any attempts at using

it for genetic discovery and also greatly limit its clinical utility by making it impossible to audit. A number of different methods have been developed to try to extract explanatory information from trained DL models in an effort to mitigate the black box problem [156, 37]. However, most of these methods were developed specifically for more popular DL application domains such as image recognition, and they cannot necessarily be applied out-of-the-box to problems in other fields. This is especially the case when there are notable idiosyncracies in the nature of the model's input data, as is the case in genotype data: genetic point mutations resemble neither pixels in an image nor sequential time-series data like voice recordings. Research is thus needed not only into how to create and optimise DL prediction models for genotype-phenotype mapping, but also how to successfully interpret these models. It is not enough to build predictive genotype-phenotype models: only with successful model interpretation will we be able to get an "under-the-hood" view of the learned genotype-phenotype relationship, and potentially gain more insight into the genetic basis of disease.

The aim of the rest of this chapter is to discuss the various research questions investigated by this thesis, and summarise the main contributions made to the field by this PhD project. Finally, we will provide an outline of the structure of the rest of the thesis.

1.1 Research Questions and Main Contributions

The major aim of this doctoral research project is to investigate whether the training and interpretation of deep learning models to predict phenotype from genotype can increase the amount of useful biological information that may be gleaned from large genetic datasets. In order to assess this broad question, we constructed an end-to-end pipeline for training, evaluating and interpreting Body Mass Index prediction models on genetic data. The following research questions each contribute to different stages of the pipeline.

1. **RQ1: Can more specialised encoding techniques increase the amount of information that can be learned from genomic data by prediction models?**
   In previous studies, SNP-level genomic data derived from GWAS is typically encoded using the "standard variable encoding", where the number of alternate alleles is used

as a stand-in for the genotype at a certain locus. For a variety of reasons (which are covered in depth in chapter 5), this does not necessarily encode useful information and may even be misleading for a deep learning model. Some studies have instead treated SNP-level genotype data as categorical inputs to avoid these pitfalls, but this introduces its own problems and is much more computationally intensive. An accurate and descriptive encoding method for SNP data could greatly increase the amount of information that can be automatically learned by a deep neural network. To our knowledge, a purpose-designed encoding method for deep learning models training on SNP data has not been previously introduced.

Contribution:

We propose a variety of novel encoding methods for SNP-level genetic data intended to be used for training a deep learning model. The proposed encodings are designed to avoid many of the simplifying assumptions made by the standard variable encoding or basic categorical encoding, and follow a range of methods, leading to different trade-offs between fidelity of representation and efficiency of training. We tested these encodings on a variety of different deep learning and linear models and showed that they led to improved performance over the standard variable encoding while offering improved efficicency over the categorical variable encoding. For more detail, see chapter 5.

2. **RQ2: Can deep learning techniques improve prediction of Body Mass Index from genetic data in comparison to linear models?**

   BMI is a complex polygenic phenotype with an as yet unclear aetiology for which traditional statistical methods have had somewhat limited success. Though previous studies have developed deep learning models and tested them on BMI, many considered it as one among many phenotypes and did not optimise networks with BMI in mind specifically, and other performed analysis only on very small datasets. Understanding whether the automatic nonlinear modelling capabilities of deep neural networks can offer improved BMI prediction in comparison to simpler linear models could provide clues about the performance of neural networks on predicting other complex heritable phenotypes, as well as about the nature of the genetic basis of BMI itself. How to design and optimise networks for genetic data is an open question, with the choice of hyperparameters in particular still

proving challenging.

Contribution:

We trialled a number of different deep learning and linear models with different hyperparameters, the combinations of which were examined exhaustively with gridsearch. We trialled several different types of neural network architecture as well as different types of linear baseline, and we also investigated how the number of input SNPs affected model performance, as well as how different models worked with the different encoding methods described in RQ1. We found that deep learning models consistently outperform linear models for the prediction of BMI across all combinations of hyperparameters, and make some suggestions as to how to best prepare models for training on genetic data and datasets for use with deep learning models.

3. **RQ3: Can deep learning model interpretation techniques allow us to extract biological insights from models trained on genetic data? How can the use of biologically-informed interpretation feature design improve the utility of model insights?**

   While phenotype prediction in and of itself has some utility, all applications of phenotype predictor models require some level of interpretability in order to be deployable in any context (clinical or research). While existing deep learning model interpretation methods can to some extent be transferred to genetic data applications, attempting to use these techniques "out of the box" without adapting them for the specific problem context can greatly limit the amount of useful biological information that can be gleaned from a trained model. Conversely, finetuning the interpretation methods with features designed with specific biological questions in mind can lead to the generation of more comprehensible model interpretations that can be applied more readily to biological research.

   Contribution:

   We propose a novel framework for grouping input features into biologically-informed interpretable features in order to extract maximally informative explanations for model predictions, and also a method for aggregating these explanations across samples in order to gain insights into the entire dataset. We also propose a method for generating a null interpretable feature importance distribution in order to determine a threshold for genuine interaction values. We show that the results of our proposed method are robust and

reproducible across data partitions, and that they closely correspond to existing scientific knowledge of the genetic basis of BMI, validating that the model has automatically learned this information. We show that model explanations derived using this method are informative and useful for both model validation and model probing for biological research. Our approach is flexible and model-agnostic, meaning it could be applied to any trained model of sufficient prediction performance in order to extract useful genetic information.

4. **RQ4: Can perturbation-based model interpretation techniques with biologically-informed features be extended to examine interactions between multiple genes?**
   It is extremely difficult to detect gene-gene interactions from genetic data using traditional statistical methods, and there is not a clear consensus on how to extract this data from trained deep learning models. Many existing methods may produce results that are difficult to replicate or interpret, and analysis can rapidly become intractable due to combinatorial complexity. A computationally efficient method to detect gene-gene interactions from a trained deep learning model could be applied across a variety of contexts, leading to potential novel insights for different phenotypes and informing the direction of future research.

   Contribution:

   We propose a novel approach to detecting gene-gene interactions from any model trained on genetic data that is capable of learning complex nonlinear interactions between features. We also propose methods for further analysis of the results of the detected interacting pairs, allowing for the construction of a gene-gene interaction network. We show that the application of our approach on our best performing deep learning network for BMI results in the replication of several known gene-gene interactions, as well as the suggestion of possible new interactions that would have been difficult to predict using other methods. Our gene-gene interpretation pipeline is highly flexible and could be applied to a variety of phenotypes and models.

1.2 Structure of thesis

In chapter 2, we will survey the relevant genetic background to the research questions. We will give an in-depth overview of the current state of the art methods for detecting genetic influences on phenotype, and outline some of the research that led to the development of these methods, as well as the areas in which these methods may fall short. We will then cover background knowledge and state-of-the-art techniques for detecting gene-gene interactions. In chapter 3 we will provide an overview of the field of Deep Learning, outlining some of the theory governing DL models, as well as the different types of DL models considered in this project. We will then discuss the possibilities opened up by applying deep learning methods to genomic research, and explore the advances made by deep learning so far in a variety of genomic contexts. In chapter 4 we will provide an overview of the dataset used for this research, as well as the preprocessing steps that were used on the data before any training and prediction took place. In chapter 5, we will first explore the advantages and shortcomings of different methods of encoding genetic data for machine learning models, contrasting the more standard approaches with the four novel encoding methods I designed. Then we will discuss the various types of machine learning models trialled and factors that were considered in their design. Finally, we will present and discuss the results of the various models and architectures on the dataset. In chapter 6, I outline a variety of deep learning interpretation methods, including a novel interpretation technique with features designed around biological prior knowledge, and present and discuss the results of applying several of these techniques to my trained models. In chapter 7, I give an overview of gene-gene interaction detection techniques, outline my novel method, and explore and analyse the results of applying this method on the trained models, using the results from chapter 5. In chapter 8, we summarise the contributions of this thesis, evaluate their shortcomings, and make recommendations of directions for future research. This concludes the thesis.

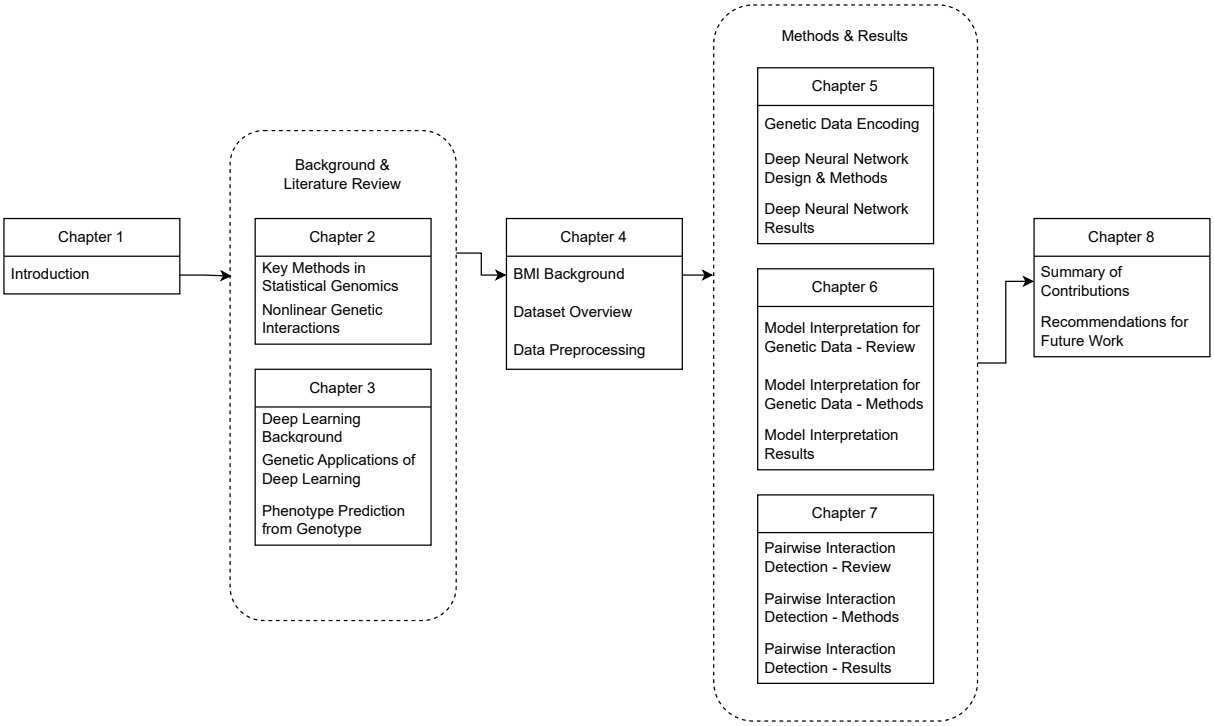


Figure 1.1: Diagram illustrating the structure of this thesis.

Chapter 2

Statistical Genomics

The aim of this chapter is to provide a detailed overview of the current status of research and methods in the area of statistical genomics. Recent developments in genomics have led researchers to theorise that the genetic basis of common complex traits may be more complicated than originally supposed, with potentially a large number of weakly significant variants contributing to a phenotype [30]. Techniques such as GWAS and PRS have been developed to identify these contributing variants, but these methods make a number of simplifying assumptions that may limit their ability to model or detect complexities in disease aetiology [326]. Initial thresholds for significance in GWAS results are being revised as modern PRS metrics with many millions of "insignificant" variants appear to outperform their predecessors [128], and as a result, new methods such as Gene Set/Pathway Enrichment analysis have been developed as a way to extract comprehensible biological information from these multi-million variant collections. However, these too have methodological limitations stemming from their simplistic modelling of the genome [193]. Furthermore, complex nonlinear interactions are known to exist in the human genome and may contribute a significant amount of "missing heritability" to traits but are confounding to all of the above methods, leading to the development of a variety of techniques to identify these interactions [187].

In this chapter, I will provide a broad overview of our scientific understanding of polygenic inheritance and how theories have changed with advancing research. First I will examine in detail the key scientific methods for the statistical analysis of the genetic basis of complex traits,

namely GWAS and PRS, investigating their theoretical basis, advantages, and limitations. Then I will discuss our current understanding of the phenomenon of nonlinear genetic interactions, and outline the basis, strengths and weaknesses of the methods that currently exist to detect them. Since the rest of this project attempts to expand upon the capabilities of current state-of-the-art methods, it is important to first investigate the main scientific goals for probing genetic data, the challenges for achieving these goals, and how the current SOTA methods either overcome or are limited by these challenges. Understanding the theoretical motivations behind different statistical techniques and how they may take advantage of or be confounded by certain features of the genome provides a basis from which to build more powerful or flexible tools.

2.1 Key Methods in Statistical Genomics

The question of how to relate genotype to phenotype is one of the oldest and most pertinent problems in genetics. Much early genetic research was inspired by plant breeding and hence focused on the idea of discrete monogenic phenotypes, with the idea that most phenotypic variability was governed by mutations in single genes [60]. However, this model fails to adequately explain traits that appear continuous in the population, such as height [332, 99]. In Fisher's seminal 1918 paper the infinitesimal model was proposed, which demonstrated that random allele sampling from a large number of genes, where the contribution each individual gene is small, would lead to a normally-distributed continuous phenotype [83]. In line with this, the *common disease common variant* hypothesis states that common diseases are most likely to be caused by variants that appear frequently in the population, which would mean that the effect size or penetrance of each causal variant must be small in comparison to the significant effects seen from rare variants, and thus that multiple such variants would need to work together in order to cause disorders with high levels of heritability [30, 14, 250]. In contrast to monogenic Mendelian disease, these diseases have been termed polygenic [96]. Despite the evidence that many complex traits appear to be polygenic, it is not immediately clear how many genes or loci are likely to be involved in such traits in humans, nor how to find them. For complex diseases where the pathophysiology is completely unknown (e.g. psychiatric disorders) identifying groups of candidate genes to analyse for significance in a disease was essentially

impossible until recently [26].

2.1.1 Genome-Wide Association Study

The advent of Next-Generation Sequencing methods made genetic data much cheaper to generate and therefore greatly increased the amount available for statistical analysis [242]. This allowed for the simultaneous analysis of a high number of commonly occurring individual point mutations known as Single Nucleotide Polymorphisms (or SNPs) across many different genes, which eliminated the need for a priori knowledge of which genes or pathways were likely to be involved in the mechanism of a disease [30]. Thanks to this newly available data, methods such as the Genome-Wide Association Study or GWAS were developed (sometimes referred to as Whole Genome Association Studies or WGAS) [307, 305]. In a GWAS, a population of individuals is observed to see if the expression of a trait is statistically associated with the presence of specific variants; higher prevalence of certain variants correlating with higher levels of the trait of interest suggests a possible link between the two [286]. While many novel genes and pathways have been identified through GWAS, shedding considerable light on the genetic bases of many diseases, for most common traits even the cumulative contribution of the SNPs that reach the significance threshold is small, a phenomenon known as the Missing Heritability Problem [180, 26]. More recent research has suggested that for many complex polygenic traits the aforementioned missing heritability may be explained by the combined effects of large numbers of SNPs with small enough effect sizes that they do not individually reach statistical significance [328, 270] and rare variants with large effect sizes which tend to be excluded from GWAS during the data quality control stage [185, 342].

A second departure from the assumptions made by the monogenic Mendelian paradigm is that many of the SNPs identified as influential by GWAS are not in protein-coding regions, but rather in areas associated with gene expression regulation such as promoters or enhancers, or in regions with unknown function [24]. So the discoveries of GWAS present us with a complex picture in which a high number of different regions all over the genome may interactively contribute to produce a phenotype. Exactly how this interactive contribution works, however, is not obvious to derive from GWAS results.

While GWAS has made huge strides in many areas of genetics, it also has a number of shortcomings that are becoming more apparent as datasets grow in size [221]. A major limitation of GWAS is simply that its utility is largely underpinned by the assumption that many diseases of interest will be governed by relatively few informative mutations, an assumption which does not appear to hold true for many phenotypes [24]. In cases where the assumption holds, the few important loci should implicate a handful of genetic areas to investigate for functional relationship with the phenotype. If, however, the genetic architecture of the trait involves a very large number of variants, it is not entirely clear what useful information GWAS can reveal; if the entire genome is implicated by GWAS, it is unclear what the next step is in terms of research or clinical utility [286]. Another major problem with GWAS is that the method assumes SNPs influence the phenotype independently and are uncorrelated with one another, which means that interaction effects confound the detection of effect variants [282]. This creates a number of issues; these are discussed in more detail in the following sections.

2.1.2 Polygenic Risk Score

As previously mentioned, most causal variance identified by GWAS have very small individual effects on the phenotype, which can limit their ability to shed light on disease mechanisms, quantify an individual's overall susceptibility, or identify potential drug targets. A method which aims to improve the power of GWAS findings by considering the cumulative contributions of many individually weakly significant variants is the Polygenic Risk Score or PRS, sometimes called a Genome-wide Polygenic Score or GPS [325]. In a PRS, a collection of variants known to be associated with the trait of interest (normally identified via GWAS) is combined in a linear weighted sum, where the weight of an individual variant in the sum is typically determined using the effect size of the variant in question, in conjunction with other factors [134, 49]. This sum can then be applied to the genotype of an individual to produce a score quantifying their relative genetic risk of developing the trait of interest in comparison to the rest of the population under study [80, 146]. Polygenic Risk Scores have shown promise in expanding the field of genomic medicine to account for polygenic phenotypes as well as Mendelian ones [144], with some PRS metrics able to detect individuals with a magnitude of risk equivalent to monogenic mutations

[127].

The success of a particular Polygenic Risk Score is evaluated by assessing whether it is able to divide a population into subgroups with sufficiently different degrees of risk, often considered in conjunction with other clinical risk factors [297]. Theoretically a PRS with high predictive power that is able to divide individuals into clinically meaningful subgroups could be useful as a way of flagging individuals with a high likelihood of developing a disease before the onset of symptoms, allowing for preventative measures to be instituted, which may be more effective or less invasive and expensive than those administered after disease onset. However, there are several limitations to the deployment of PRS in a clinical setting. Firstly, PRS is only able to capture a small fraction of the overall risk - it is unable to model environmental/lifestyle factors, gene-gene interactions (epistasis), or gene-environment interactions (epigenetic effects), and any resulting effects on the phenotype will not be accounted for [283, 151]. Poor understanding of the overall relationship between genetic susceptibility and likelihood of developing the disease also means there is immense difficulty in translating a relative risk profile within a population to a measure of absolute risk for the individual [152, 276]. This lack of interpretability in turn makes the score hard to explain, and means it may be difficult for clinicians to justify certain measures or quantify the certainty of the score for a particular individual [297].

Despite their lack of immediate clinical utility, Polygenic Risk Scores can serve another purpose: painting a broader picture of the genetic architecture of complex traits [152, 176]. PRS metrics utilising a different p-value threshold and therefore comprising of different numbers of SNPs are often trialled, and the number of SNPs in the best performing score give an indication as to the level of polygenicity of the trait. The specific variants that comprise the most successful metric can then give clues as to which genes and pathways are involved in the disease aetiology, and methods such as gene-set analysis (see section 2.1.3) can be used to try to extract this information [246, 267]. Several studies have shown that the use of a lower p-value threshold than typical in GWAS for selecting SNPs for PRS - leading to the inclusion of a larger number of more weakly significant SNPs - can improve predictive power of PRS [75, 40] and many newer scores use millions of weakly significant variants [128, 328]. While the relative success of massive PRS

scores further reinforces the validity of a more omnigenic model for these traits, it can also become a problem for interpretability, as such a large number of genes may be implicated that it no longer suggests any particular pathway or biological process [286]. Including millions of weakly or potentially non-contributing SNPs also muddies the signal of genuinely important genes or variants when methods like gene set analysis are used, raising concerns about whether the increased predictive power is worth the decrease in useful biological information which can be extracted [35]. On the other hand, the fact that prediction is improved by adding more and more "uninformative" variants throws into question the insights from the higher p-value threshold scores.

While PRS can be highly effective, there are a number of concerns with its methodology, which throw into question both its ability to accurately quantify individual risk and also any biological insights derived from its interpretation [119]. Perhaps the most glaring shortcoming is the assumption of additivity; there is concrete evidence for the non-additive interaction of genes and variants, and as such a model that is constrained to consider only additive genetic interactions will never be able to capture the full picture of the genetic aetiology of a phenotype [220, 199, 324]. The large number of non-coding SNPs detected by GWAS are less likely to directly influence the phenotype in a linear additive fashion than they are to be involved in regulatory pathways, within which the relationships are unlikely to be linear. Additionally, environmental and epigenetic effects can distort what would otherwise have been an additive relationship [102]. The PRS method also assumes that risk variants are independent from one another, which is untrue (see discussion of Linkage Disequilibrium in section 2.1.2.1) [26]. Linkage disequilibrium also means that the composite SNPs of the PRS may not actually be the causal SNPs for the disease, or the imprecision of their GWAS-derived effect size may mean their weighting in the sum is inaccurate [152]. To further complicate matters, the effects of some non-transmitted SNPs are accidentally captured in a PRS due to the phenomenon of "genetic nurture", meaning it is not possible to fully exclude environmental influences [133]. All of these factors make biological knowledge derived from analysis of PRS more dubious, and suggest that disease aetiologolical theories developed via PRS are unlikely to be comprehensive.

Differences among populations in factors such as allele frequency and linkage disequilibrium block size also mean that PRS metrics often don't transfer well to populations of different ancestry than the cohort that they were built for, and their construction is not particularly flexible [232, 317], and even show differences in performance among subgroups *within* ancestry groups [208]. Due to the fact that the majority of Biobanks and other genetic data-gathering initiatives are undertaken by countries with majority European-ancestry populations, there is a significant over-representation of this ancestry group in available genetic data, and it has been standard practice in the past to exclude individuals from minority backgrounds or whose ancestry is mixed from statistical genomic studies such as GWAS, as usually the number of these samples is too low for the model to be able to glean useful information from them. As a result there is significant risk that disease understanding will be accelerated for some groups while leaving others behind, potentially leading to clinical tools or treatments designed for only people of certain ancestries, with damaging and potentially dangerous effects for others [76].

Despite the evidence that the linear additive model used by PRS is overly simplistic and not true to the biological reality, for a sizeable subset complex traits meta-analysis of heritability (the proportion of phenotypic variance that is attributable to genetics) from twin studies shows that a simple additive model is sufficient to explain the majority of effects, so it often is still possible to get good prediction results without considering nonlinear interactions [107]. However, this required assumption of additivity will always be a major disadvantage when it comes to interpretation.

2.1.2.1 Linkage Disequilibrium

The theoretical basis of PRS depends on the assumption that the individual variants in the sum are not correlated with one another, an assumption which makes the selection of constituent variants somewhat challenging, as correlation between variants in sections of the genome is well-documented. The nonrandom association of alleles of two or more different loci is known as Linkage Disequilibrium (LD), and this is one way in which the assumption of SNP independence in a PRS is flawed [275]. Due to the manner in which DNA is passed from parent to offspring,

there are large regions in the genomes of populations where there is little evidence of historic genetic recombination, known as haplotype blocks. Within a haplotype block, which can span relatively large chromosomal regions in some populations, all the SNPs are highly associated with one another [90]. This means that all may show up as important GWAS hits despite only a small number being biologically associated with the disease, leading effectively to noise in the PRS sum and confounding interpretation as well as potentially reducing prediction accuracy [30]. The size and composition of haplotype blocks also may differ between populations, meaning that two SNPs may be in LD in some populations but not others [232].

To try to eliminate this issue, various methods are employed on GWAS-identified variant sets before the calculation of a PRS (or other further analysis). LD pruning is a simplistic method that calculates correlation between SNPs and removes one SNP of every highly correlated pair from further analysis, potentially resulting in a significant loss of data and with no guarantee that the removed SNP was not the one responsible for the biological effect [32]. Informed LD pruning, otherwise known as LD clumping, takes a slightly more sophisticated approach in that it ensures the most statistically significant SNP in a region of pre-defined size is retained, with the aim being to end up with one representative SNP for each region [238]. However this method may also remove a large number of potentially informative SNPs, and it is possible that the most statistically significant SNP is still not the one responsible for the effect. The most sophisticated tool currently in use is LDPred [304] which uses machine learning to calculate the posterior mean effects of SNPs using a genetic architecture prior and information about LD from a reference panel, and uses this information to re-weight the constituent SNPs of a PRS. As the prior for the effect sizes is a point-normal mixture distribution, there is no assumption of infinitesimal genetic architecture. LDPred has some limitations, such as poor performance on the Human Leukocyte Antigen region, which were improved upon by the later release of LDPred2 [239].

How much and how to account for the effects of LD is a major consideration for any method which aims to link genotype and phenotype. Use of sophisticated tools for SNP selection with consideration of LD has been shown to improve the efficacy of PRS [304]. However this consid-

eration may be redundant for more complex nonlinear models that make no prior assumptions about SNP independence. For a model that uses representation learning to automatically model patterns, it would be possible to automatically learn LD from the data itself and use this information in prediction; indeed, previous studies seem to show that attempting to parametrically account for LD in deep learning feature selection hinders network performance [17].

2.1.3 Gene Set Analysis

Another family of methods which aims to capture the cumulative effects of many SNPs is Gene Set Analysis (GSA). These methods may be performed on either the significant SNPs identified via GWAS or on the constituent SNPs of a PRS, which can be mapped to their closest associated genes. Genes are not believed to usually work in isolation in disease pathogenesis, instead forming complex networks involving many regions across the genome, and certain groups of genes have been associated with biological processes and functions by prior research [142]. GSA methods investigate whether there is a statistical relationship between these biologically informative gene groups and the variants associated with a phenotype, with a correlation suggesting that there may be a direct relationship between the biological function and the pathogensis of the phenotype [68, 312]. This can help move away from the granularity of individual variant effects and towards a more broad picture of the genetic basis of disease, as well as potentially allowing for the detection of more subtle combination effects that may be missed individually [86].

There are many different tools and approaches to perform GSA, which may use different significance tests or null hypotheses. These methods can broadly be separated into two categories: competitive and self-contained [95]. Competitive methods compare the proportion of significant SNPs contained in the gene set of interest to the significance proportion in other gene sets, so the null hypothesis is that the SNPs/gene in the gene set of interest are no more associated with the phenotype than those outside of the gene set of interest. A variety of different statistical tests may be used to determine significance of association between SNPs and the gene set of interest, depending on the method [224, 271]. This approach requires both the partitioning of variants into "significant" and "not significant" categories while ignoring the relative strengths of the association, and choosing a significance threshold for gene set association, which can prove

problematic. Some examples of competitive methods include functional class scoring (FCS) and parametric and non-parametric significance assessments, which assign scores to individual genes and gene sets, usually by using some kind of statistical test, and rank gene sets based on these [178].

Self-contained methods, on the other hand, do not take a comparative view of significance, instead defining the null hypothesis as no significant association between SNPs/genes in the gene set and the phenotype, versus the alternate hypothesis that the gene set is statistically associated with the phenotype. Instead of comparison with other gene sets, the deviation from the expected proportions of significant SNPs is assessed. This is also known as over-representation analysis [113].

Gene sets may be derived using information from a variety of sources, such as functional pathways or gene expression data, and are stored in public meta-databases such as GeneSetDB or the Kyoto Enclyclopedia of Genes and Genomes (KEGG) database [10, 124]. Some more sophisticated methods may take into account prior information about pathway topology to prioritise certain genes within the set [74].

GSA methods offer the opportunity to move from the large volume of association data derived from GWAS towards actual genetic discovery, providing future promise for identifying the functional genetic basis of phenotypes of interest. Unfortunately there is currently little consensus on the best practices or even the best data sets to use for benchmarking, and many results are not reproducible [178]. Even how to map SNPs to their nearest genes has proven contentious, with numerous different gene boundaries proposed by different studies [311, 233, 121, 45], and questions about whether to include known associations or SNPs in LD [313]. The most commonly used over-representation based methods rely on the assumption that genes are uncorrelated and do not interact with one another, which has been showed to be flawed [91]. While competitive GSA methods fix some of these issues, they do still suffer from a lack of reproducibility of results, possibly indicating either a lack of sensitivity or specificity in these methods. For all methods, gene set overlap (where some genes appear in many gene sets due to gene multifunctionality or

as an artefact of the parent-child structure of functional labelling systems like Gene Ontology) is a confounder for interpretation of results, as one gene having a high significance score may result in all the sets it is present in to be reported as enriched, even if the associations between other genes in the sets and the discovery SNP set is low [177, 289]. Additionally, many methods are confounded by highly polygenic phenotypes which are likely to have causal SNPs all across the genome, which means that many gene sets will capture some of the heritability by chance and the results will be less informative [68]. There are also limitations associated with relying on biological prior information, namely that the gathering of information for Gene Ontology is not yet finished, with for example only 5000 genes so far appearing in the KEGG database, meaning that there is no guarantee that gene sets derived from it are 100% correct and complete [313]. In sum, GSA is an exciting area of opportunity in statistical genetic analysis which still has many unanswered questions.

2.2 Nonlinear Genetic Interaction

2.2.1 Epistasis

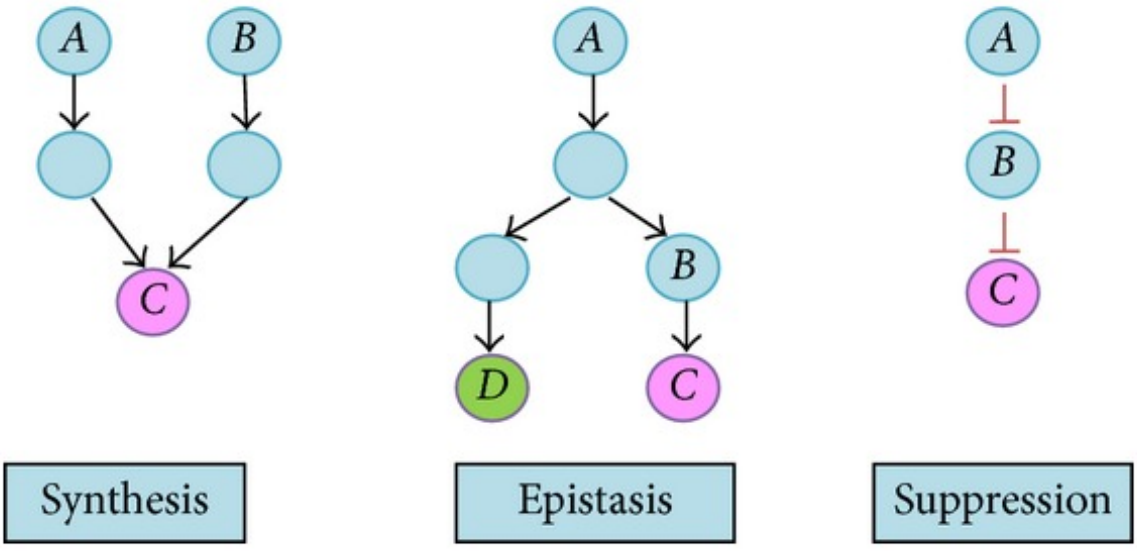


Figure 2.1: Diagram showing some different types of genetic interactions. [135].

The term epistasis is not always consistently defined, but it broadly refers to the interactions between genes. Epistatic effects pose a major problem for methods such as GWAS, as if the effects of an allele at one locus mask or augment that of those at another there is a high possibility at least one of the two important loci will be missed by GWAS, limiting its ability to shed light on genetic disease mechanisms; the problem becomes dramatically worse when more than two loci are involved [57]. On the other hand, if a complex interaction were to be detected and analysed, it could potentially shed significant light on the underlying functional pathways of a disease [235].

There are several ways in which genes may interact, which are illustrated in figure 2.1[135]. Synthetic interaction is when two genes independently produce the same phenotype, meaning that the phenotype may still be observed even if one of the genes is inactive, despite that gene being influential on the phenotype. Synthetic genes A and B in figure 2.1 are on different non-interacting pathways but both can result in the phenotype C, so C will result even if A or B is knocked out. Suppressive interaction occurs when the negative effects of one gene are compensated for by the positive effects in another, leading to the phenotypic influence of both to effectively cancel each other out, again confounding methods that attempt to measure the effect of one variant at a time. In the suppressive interaction in figure 2.1, knocking out gene A would result in no longer observing phenotype C, since gene B would start suppressing it once it iself was no longer suppressed by gene A. Finally, the two genes may interact epistatically to produce an entirely new phenotype [41]. In the wildtype epistatic interaction in figure 2.1 the observed phenotype is a mixture of D and C; knocking out gene B would change the observed phenotype somewhat because C would not be observed but D still could be, whereas knocking out A would change the phenotype more significantly because neither C nor D could be observed.

The phenomenon of genetic dominance refers to any epistatic mechanism in which only a single allelic variant is required to produce the full phenotypic effect; conversely, when a single allelic variant has no observable effect on the phenotype, this variant is deemed recessive [84, 20]. This can occur as a result of a variety of different epistatic interactions, which may also be mediated or modified by the presence of other variants. Alleles diaplaying dominance effects are believed to be more widespread than those with linear additive effects, though the exact distribu-

tion remains unknown [179, 323]. The term epistasis is also sometimes used more broadly to refer to any deviations from the linear additive combination of SNP effects, regardless of the underlying mechanism [235]. Sometimes this definition is referred to as "statistical epistasis", in contrast to "biological epistasis", which refers to the biomolecular interactions occurring at the individual level [234, 205].

On the whole it is largely unknown how much epistatic effects contribute to phenotypes in complex diseases, with research to date returning conflicting evidence, and the answer likely to differ between traits and populations [107, 47]. It has been suggested that the lack of reliable reproducibility of GWAS data across populations could be largely due to epistatic effecs that differ between populations [187]. With the presence of epistatic effects a certainty and the magnitude of their impact on a given phenotype for a given population an unknown, any phenotype model's inability to account for these complex effects will always be a shortcoming. As well as potentially improving phenotype prediction, the ability to model mechanisms of epistasis for given diseases could give key insight into disease aetiology overall, for example by implicating functions or pathways, as well as providing invaluable information for drug discovery by highlighting a greater number of potential drug targets [206].

2.2.2 Gene Interaction Detection

As previously discussed, it is highly likely that traits are governed by the interactions of multiple genes, and the effects of this may be confounding to traditional statistical analysis of gene-trait associations. Understanding the way genes interact to cause a phenotype may explain some of the missing heritability and shed light on the functional systems involved, but the detection of such interactions is challenging [34]. In simple organisms such as *e. coli* or yeast, it is possible to relatively straightfowardly detect and analyse pairwise interactions by manipulating the organism's genome in a lab and comparing the phenotypic result of a double gene knockout to the results of knocking out the single genes, a method which can then be expanded to entire interaction methods using high-throughput datasets [174, 202, 251]. These methods have gone a long way towards illuminating the types of interactions that take place even in the genes of simple organisms, and the prevalence of such interactions [59]. Unfortunately this approach does

not obviously transfer to more complex organisms with larger numbers of genes where genetic knockout studies are infeasible.

Machine learning models have been utilised to improve the detection of gene networks from multimodal data in simple prokaryotic organisms [125], but it is again unclear how well this would transfer to the genetic data of complex eukaryotic organisms. Some epistatic interactions have been discovered in humans via candidate gene studies, for example the interaction between an allele of the *ACE* gene and one of the *AGTR1* gene which greatly increate the risk of myocardial infarction [294]. However, candidate gene studies require an advance hypothesis of which genes to investigate; in most cases no such hypothesis exists, and so one of the greatest difficulties for finding new interactions in humans is the "search" phase, as the search space of possible loci involved in interaction effects is becoming extremely large for modern genetic studies [58].

Generally, in order to gain maximum insight into functional pathways, we are interested in the interactions between genes, but tend to have information about human genetics at the variant level. A gene may have a very large number of associated SNPs in a discovery set, so for linear statistical methods each gene is often modelled by the SNP most correlated to the phenotype of interest. Interactions between top SNPs are then investigated as a proxy for the genes, ignoring the effects of other SNPs, an approximation which may or may not be problematic, depending on phenotype and population studied [57].

Perhaps the simplest way of searching for pairs of SNPs which may be interacting in the aetiology of a given phenotype is the exhaustive "brute force approach", which just involves performing a linear pairwise association test between every possible pair of SNPs and the trait [319]. As GWA studies and other methods return larger and larger numbers of phenotype-relevant SNPs, the dimensionality of such exhaustive calculations increases exponentially, causing the traditional parametric statistical analysis to rapidly become highly inefficient or even intractable as studies get bigger [337]. The problem of dimensionality is significantly worse if any larger number of SNPs than 2 is considered for interaction, or if there are a higher number of polymorphisms at that locus in the population [41]. This is a major shortcoming, as research has indicated that

many important interactions are likely to be higher order [291]. It is thus essential to constrain the dimensionality of the search space if traditional statistical analysis is to be performed, and more sophisticated methods usually aim to both reduce the number of inputs and perform computations more efficiently.

There is currently no consensus on the best practice for identifying gene-gene interactions in humans at the population level, and results are often difficult to replicate [216]. The existing methods can broadly be divided into two categories: parametric and nonparametric. Parametric methods require a model designed to quantify the effect(s) of each variant on the phenotype, mathematical information from which can then be used to better understand the nature of the relationship [187]. A major drawback with parametric models is that the huge scale of modern datasets means that attempting to model interactions results in a high likelihood of ending up with too many independent predictor variables for the model [216]. Aside from this, parametric models require simplifying assumptions which may or may not be valid, and it is very difficult to design a model that adheres to the biological reality [226]. In nonparametric methods, there is no genetic model and no assumptions are made about the nature of interactions, and statistical methods are used to look for correlation between certain pairs or sets of SNPs/genes and the phenotype of interest instead [256]. This makes them more flexible and avoids issues arising from making simplifying assumptions about the underlying genetics. The downside of nonparametric methods is that they have less statistical power than parametric methods with valid simplifing assumptions and often require choosing an arbitrary significance threshold [216].

Of the parametric methods, the most popular tend to involve some kind of regression, usually logistic or linear. A regression model that contains multiplicative interaction terms may be designed for the SNPs or genes of interest. This can then be fitted to the data, and the presence of interaction between a SNP or gene pair will be indicated by a nonzero coefficient for the relevant interaction term [58]. A major limitation of this approach is that the nature of the functions governing linear regression models is such that they are unable to detect interaction effects between SNPs in the absence of main effects, which is an issue as there is no reason to believe that SNPs which are epistatically influential over the phenotype must necessarily be individually influential

as well, so many interaction effects could be missed this way [204, 64]. As mentioned previously, they also rapidly become computationally intractable when too many genes/variants or higher order interactions are included in the model, meaning that it is usually necessary to perform significant filtering on the input, potentially resulting in the discarding of useful data. They are also highly sensitive to missing or incomplete data. As a result, some methods have been developed to choose the right number and selection of features for input to a parametric model [106, 109, 136].

Another parametric method for detecting interactions between candidate genes in humans is the Focussed Interaction Testing Framework or FITF [192]. In this method, several stages of likelihood-ratio tests are performed, with each stage corresponding to a joint test of main effects and an order of interactions. Candidate genes can be pre-screened with goodness-of-fit tests to reduce computational complexity, and false discovery rates are controlled in order to limit type-I errors from multiple testing. This method shows some promise for detecting interactions in candidate genes, but its efficacy has not been thoroughly evaluated, and in particular little is known about its performance on population-level GWAS data [216].

Nonparametric methods offer an advantage over parametric methods in that they can detect interactions in the absence of significant main effects, however on the other hand they may be confounded by situations where there are simple main effects without major interactions [216]. The main goal of certain nonparametric methods is simply to reduce the dimensionality of the problem in order to make further analysis tractable. One method which aims to analyse the interactions of multiple SNPs at a time for binary phenotypes is Multifactor Dimensionality Reduction (MDR) [254, 212], which is a nonparametric approach to simplify multifactorial genotypes into a binary for further analysis. This means it can model higher order SNP interactions, gene-gene interactions, or interactions including multimodal data such as information about environmental factors [187]. This method has been used successfully to detect interactions in the absence of main effects [255]. MDR has been used to successfully detect interactions in numerous phenotypes such as breast cancer [255], hypertension [204], type 2 diabetes [48], atrial fibrillation [209], autism [171], myocardial infarction [54], and schizophrenia [243]. While versions of this method are very popular due to their ability to greatly reduce the computational

overhead required to perform pairwise SNP analysis, the method is still computationally intense if more than 10 variants are being considered, and may require useful SNPs to be discarded to reduce complexity. The resulting models can also be difficult to interpret, especially when the number of SNPs considered increases; this makes them better suited for suggesting SNPs involved in interactions than for rigorously probing the nature of these interactions [255]. The Combinational Partitioning Method (CPM) similarly aims to reduce dimensionality by creating informative variant groups, aiming to create groups of genotypes that predict similar phenotypes in order to predict as much of the phenotypic variance as possible, and then make inferences about the underlying genetic mechanisms from these genotype groups [219]. This method has been used with success to detect some gene-gene interactions [200]. The computational burden of this method is very high and it suffers from some methodological problems, so the Restricted Partitioning Method was developed to address these [63]. This uses an iterative procedure to create the genotype groups by sequentially merging genotypes with similar mean phenotype values until no genotypes remain. Both methods suffer from high computational burden and their efficacy is not thoroughly explored. While RPM is more interpretable than CPM, it also has methodological issues due to multiple testing [216]. Principal Component Analysis has also been suggested as a tool for investigating gene-gene interaction via dimensionality reduction, but the effectiveness of this approach is unclear [153].

Other types of nonparametric methods simply use linear statistical tests look for correlation between pairs of SNPs or genes and the phenotype of interest. Some methods make use of Linkage Disequilibrium to compare between cases and controls to see if there is increased association between a pair of markers in people with the trait of interest, suggesting a possible interaction [338]. Other methods utilise statistical metrics such as the Pearson χ^2 test or odds ratios to see if certain genotype combinations at the two loci are overrepresented in the case group vs control group [111]. A general limitation of these *a posteriori* association methods is that the null hypothesis relies on the assumption that the two variants are not correlated in the general population, which may not necessarily be true even if they are not in LD, leading to potential false positive results [58]. They also require the division of the population into case and control groups, making it less obvious to apply them to quantitative phenotypes.

Some new methods aim to combine parametric and nonparametric methods in order to exploit the advantages of both [203, 186]. While these may offer more coverage across different types of epistatic scenarios (in terms of the strength of main and interaction effects), there is still currently no single method that will work in every case.

Neural networks are a new class of nonparametric method for interaction detection. These have the advantage of being able to model complex epistatic situations without prior knowledge of the types of interactions taking place, and they are well suited to large datasets [149]. The drawbacks include the fact that they are especially prone to overfitting, their architecture can be very difficult to design, and they can be extremely difficult to interpret due to their black box nature. They also, despite on the whole handling large datasets efficiently, rely fairly heavily on cross-validation, which is likely to make them computationally intensive. Specific studies that have made use of deep learning models for interaction detection will be discussed in the next chapter. Some neural networks have also been used as automatic input selection tools for other methods [210].

There are a number of methodological problems that plague all of the aforementioned approaches. One is data sparsity, whereby there may be insufficient samples exhibiting all of the different allelic combinations of the variants of interest to compare them; while this problem is somewhat mediated by the large sample sets available from databases such as Biobanks, it remains an issue for trying to infer effects. This issue can lead to either model underfitting, as data is insufficient to illustrate a pattern, or overfitting, where the model fit is too specialised to the training data; or for correlation-based methods, it can simply make the study too underpowered to discover anything. Overfitting can be reduced by using cross-validation on a partitioned dataset; however doing so increases the computational burden. Another issue is that the large number of input variables generated by GWAS studies can elevate the risk of false positive results due to stochastic variations potentially also combined with multiple testing. One of the most major issues is still computational burden, which will always have the potential to be considerate for any method if a large number of variants are investigated [216]. Often there is a tradeoff between

being statistically rigorous (using methods such as cross-validation and permutation testing for establishing significance) and being computationally effective.

Chapter 3

Deep Learning and Genomics

The purpose of this chapter is to provide a general understanding of the field of deep learning, its previous applications, and how it has specifically been applied to statistical genetic data analytics to date. We will also discuss the increasingly crucial role in which model interpretation is playing in many modern deep learning applications. The aim is to illustrate how the properties of deep learning models make them uniquely well-suited to the new era of genetic big data, as well as highlighting the significant challenges which still obstruct their successful implementation in this domain.

This chapter is divided into four main sections. In the first section, we will provide a general background of deep learning, including discussing the broad applications of deep learning and how the field has advanced in recent years, and outlining the theoretical basis of neural networks. In the next section we will discuss neural networks in more detail, describing some different types and their applications as well as techniques and tools used for training. In the following section we will move on to discussing the motivations of the interpretation of deep learning models, and the specific challenges of interpreting models trained on genetic data. Finally, we will provide a survey of how deep learning techniques have been applied to the field of statistical genomics. This will all provide context for the deep learning and interpretation techniques which I implemented for this PhD project and which are described in chapters 5, 6 and 7.

3.1 Deep Learning Background

Deep learning is a specialised branch of machine learning that has been rapidly expanding in recent years, with methods being applied to a wider and wider range of problem domains, often with unprecedented success [273]. The term deep learning refers to the use of a deep learning model to detect patterns in data for the purpose of prediction via classification or regression, where a deep learning model is simply a neural network (sometimes also called a multi-layer perceptron [259]) with more than two hidden layers. Deep learning models excel at automatically learning patterns from large amounts of data without explicit parametric input, a form of representation learning [7]. While the theoretical basis for deep learning has been in place for many decades now, the recent explosion in "big data" availability coupled with advances in computing power has unlocked practical potentials for this technology [218].

Neural networks are constructed of layers of neurons or nodes interlinked by weighted connections [266]. Individual neurons transform the input they receive from a single previous neuron into an output $z = x.w + b$, where x is the input received from the previous neuron, w is the weight of the connection between the neuron and the previous neuron and b is a bias term used to adjust the intercept of the transformation to get the right outcome. The input layer contains one neuron for each input feature. There will then be a hidden layer where typically each neuron is connected to multiple input neurons and combines the effects of its j input connections using a linear weighted sum, producing an output $z = \sum_{i=1}^{j} w_i x_i + b$, which is then multiplied by a nonlinear activation function. The resulting output will be passed to any number of further hidden layers in the same way and further augmented with nonlinear activation functions, until it is transformed into an interpretable result for the output layer. The model "learns" to represent the data by changing the weights of the connections between neurons iteratively during the training process using back-propagation [149, 262]. In this way, a series of individually simplistic but nonlinear units can be combined to automatically learn progressively more complex representations of the input data, without requiring explicit engineering for this task by a human with domain expertise [148]. This means that deep learning models have the potential to learn representations of data that humans are unable to interpret, because it is too large, too complex, of a format that is not naturally comprehensible to humans, or all of the above.

The structure of nodes and connections in a neural network is known as its architecture, and different classification and prediction applications require different architectures with features designed to handle the unique attributes of the data. For image classification, Convolutional Neural Networks (CNNs) have proven highly successful, as they are able to use convolutional and pooling layers to exploit spatial relationships between features in two dimensional inputs; more recently, CNNs with three dimensional convolutions have been successfully used to interpret video data [158]. For language applications, various types of Recurrent Neural Network (RNN) have been used with success [225]. RNNs contain hidden units that retain a state vector, allowing them to "remember" information about the previous state of a system, which makes them well-suited to sequential data such as text. For applications where there is inherent stochasticity and the training labels are uncertain, Bayesian neural networks may be used [173]. There is a lot of structural diversity even within network subtypes, and on top of this, different architectural families may be hybridised to create an enormous array of different network architectures for different problem domains.

The types of networks described above all make use of supervised learning via backpropagation, in which the network is trained on data that has been labelled by a human. The network aims to minimise a loss metric that quantifies how far away its own predictions are from the human-derived labels in the training dataset, and it does this by adjusting its modifiable parameters, i.e. model weights, during training [65]. In order to minimise its objective function, thereby minimising the error metric, most networks utilise Stochastic Gradient Descent (SGD) to search for local minima in the loss function [8]. Some deep learning models, however, make use of unsupervised learning, in which the training data is not labelled and the model's task is to divide sufficiently samples into a set of categories of its own design [71]. Such models include autoencoders, restricted Boltzmann machines and deep belief networks [36]. This PhD only considers supervised machine learning models.

3.2 Neural Networks

Also known as multi-layer perceptrons, feedforward neural networks are a class of prediction models that can automatically learn to model a relationship between inputs and outputs, even when the nature of this relationship is not known in advance [148]. NNs are comprised of an input layer, an output layer, and at least one hidden layer; deep learning networks may have large numbers of hidden layers. The term feedforward refers to the fact that the network does not have any loops where the same neuron may see the same input twice in a single pass. For the purposes of this thesis, all feedforward networks are also fully connected, meaning that every neuron in a layer is in some way connected to every neuron in the preceding and subsequent layers. The input layer consists of a set of neurons that represent the input, in our case SNPs. The output, also known as the label or target value, is in out case the Body Mass Index of an individual. The input is transformed into the dimensions of each subsequent layer of neurons, using a weighted linear sum, until the output layer is reached and the BMI output is generated. In this way, a neural network can be considered to be a directed series of nonlinear classifiers. The output from the input nodes to the nodes of the first hidden layer can be represented as follows:

$$H_j^{(1)} = \sigma(\sum_i w_{ij}^{(0)} X_i + b_i) \tag{3.1}$$

where σ is the nonlinear activation function, $w_{ij}^{(0)}$ are the weights of the connections between the nodes in the first hidden layer and those in the input layer, X_i represents the nodes in the input layer, and b_i is the bias term. The output is thus a weighted sum of its inputs. The equation can be extended to the kth hidden layer in the form:

$$H_j^{(k)} = \sigma(\sum_i w_{ij}^{(k-1)} H_i^{(k-1)}) \tag{3.2}$$

The output layer is then simply the linear combination of all hidden layers:

$$O = \sum_i H_j^{(k)} \tag{3.3}$$

3.2.1 Training and Optimisation

Neural networks are not able to infer anything about the data based on their design alone; they must go through a phase of weight and bias optimisation in which they will tune their parameters to best match the data available. During this training phase, the network is repeatedly exposed to the training data and after each pass the weights and biases are adjusted using backpropagation. The goal is to adjust the weights in such a way that the loss or cost function of choice - which quantifies how far the network's predictions are from the true values of the data - is minimised. The most common loss function minimisation approach for most types of neural networks uses stochastic gradient descent (SGD) [8]. It builds on the basic method of gradient descent, which is an algorithmic tool used to find the minimum of any differentiable function by taking a "step" of a pre-determined size (decreed by the learning rate) in the direction of greatest negative gradient until all of the surrounding gradients are positive, i.e. a "valley" in the function has been reached [261]. SGD augments this by only using a random subset of observed local gradients in its decision making, thus potentially reducing computation time and also adding an element of randomness that may help with issues such as getting stuck in local minima instead of finding the global minimum.

Neural networks on the whole are much more flexible than many other machine learning models and can learn to model arbitrarily complex functions. However, the variety of combinations of hyperparameters – such as architecture, learning rate, and activation function – can be significant, and choosing a good combination for the task at hand is nontrivial, especially when domain knowledge about the prediction problem is limited [274]. This is where various hyperparameter optimisation techniques may be employed, where the simplest approach is the exhaustive or semi-exhaustive exploration of hyperparameter combinations via gridsearch [159].

3.2.2 Convolutional Neural Networks

Convolutional neural networks were originally developed to exploit multi-dimensional spatial patterns for image recognition tasks [98]. They have since been successfully used for a variety of other pattern-finding applications, including classification problems involving DNA sequence

data [245, 341]. The application of CNNs on SNP data is less obvious, as it is not possible to infer local DNA pattern information from the presence of a few sparse variants that may be hundreds of kilobases apart on the genome. Furthermore, most DNA-related data is decidedly one-dimensional, which means that CNNs' abilities to locate higher-dimensional patterns are useless in this domain.

In line with previous research [17], we attempted to trial a 1-dimensional convolutional neural network alongside the feedforward networks. However, we encountered significant GPU memory issues in implementing the convolutional layers, even when using shallow networks with very small kernels, on any of the datasets larger than 100 SNPs, leading to us abandoning this line of enquiry.

3.2.3 Bayesian Neural Networks

Accounting for and quantifying uncertainty in predictions is a major obstacle for classical neural networks. Bayesian Neural Networks are a class of neural network that is uncertainty-aware [123]. The goal of neural networks as a whole is to learn to represent an arbitrary function; while typically this is done by minimising a cost function to find a single value for each parameter (hereafter referred to as the "point estimate" method), Bayesian neural networks explicitly model uncertainty by incorporating stochastic components into the network, either in the form of a stochastic activation function or stochastic weights [94]. This means that the final trained network is nondeterministic, in effect simulating multiple possible models in one; for this reason, Bayesian NNs can be considered a special type of ensemble classifier. As with other types of ensemble learning, this exploits the fact that aggregating predictions over multiple average-performing predictors tends to produce better results than a single well-performing predictor can [173]. There are also several other advantages to this method: the inclusion of stochastic components reduces overfitting, potentially allowing for smaller datasets to be utilised, and the resulting network is able to quantify the uncertainty of a given prediction, reducing prediction overconfidence and the risk of resulting silent failures. Bayesian NNs can concretely quantify the uncertainty of a given prediction by analysing the spread of prediction outcomes for the same input; low spread corresponds to low uncertainty in the prediction, whereas high spread indicates

high uncertainty. This means the trained network can then be used for prediction on samples that are very different than the ones it trained on without risk of a confident incorrect prediction, as it will make a prediction with high uncertainty if it has not learned enough from the training set to be able to generalise to this sort of sample. This also provides insight into the learning process.

While many of the above features (such as accurate confidence estimation and improved performance on smaller datasets) make Bayesian Neural Networks highly recommended for biological datasets and/or clinical usage, they are not frequently used for phenotype prediction from genomic data. A challenge of phenotype prediction is that there is unavoidably a large amount of both epistemic and aleatoric uncertainty involved. While information about factors such as diet may be useful for predicting BMI outcomes, the nature of the influence lifestyle and environmental factors is complex enough that is will never be possible to gather data of sufficient detail to make deterministic predictions using these factors. Additionally, there is an element of randomness to genetic expression that adds further unpredictability to the phenotype data. As a result, this entire prediction problem is unavoidably noisy, which increases the risk of overfitting by complex networks. Hence Bayesian Neural Networks could potentially offer a good approach due to their built-in overfitting protection and ability to learn a representation of noisy data.

3.2.4 Loss Functions

Loss functions in machine learning are a sytematic way of comparing a model's prediction with the ground truth value for a given sample, allowing a model's prediction performance to be assessed and compared across the dataset [314]. An optimisation function is an algorithm that seeks to maximise the model's performance by minimising the loss. Different loss functions may put more or less emphasis on different sorts of prediction errors the model makes while learning depending on how they quantify error. Since deep learning models update their parameters in real time while training, the choice of loss function can have a meaningful impact on the ultimate structure and performance of the model, and can make a model much more or less well suited to the type of data and task it is designed for. In the following subsections, some different loss functions are explored.

3.2.4.1 Mean Squared Error

The mean squared error (MSE) is a simple loss metric commonly used in machine learning [321]. It is simply the square of the difference between prediction and ground truth value, averaged across the training set or a single batch, defined as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{3.4}$$

where N is the number of data points in the set or batch being used for comparison, y is the ground truth value and $\hat{y}$ is the prediction being compared. It can be proven that MSE is the sum of the variance of the estimator and the square of the bias [73]. Thus when the model has no bias, the MSE corresponds exactly to the variance of the estimator.

Due to the squared property of MSE, larger errors quickly get amplified, meaning a model trained with MSE loss may heavily penalise outlier predictions during training, leading to a trained model which is better at predicting the majority class than samples which deviate from the mean. For fairly homogeneous datasets with few outliers, or applications for which a model that occasionally makes very erroneous predictions would be highly problematic, this can be a positive attribute. For more diverse datasets in which the model is likely to make many predictions with large errors during training, this can lead to underfitting of the dataset due to an inability to learn to predict outliers.

3.2.4.2 Mean Absolute Error

While being quite mathematically similar to MSE, the mean absolute error (MAE) loss has very different behaviour when it comes to outliers. Defined as the absolute value of the difference between prediction and ground truth, averaged across the batch or dataset, it is represented as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{3.5}$$

where N, y and $\hat{y}$ are defined as above. MAE weights all errors the same and hence provides a generic loss function without the pitfalls of MSE when it comes to outliers. However, in situations where we would prefer large error predictions to be more heavily penalised, a model trained with MAE may not be ideal, as it is more likely to occasionally make very erroneous predictions.

3.2.4.3 Huber Loss

The Huber loss function was more recently developed as part of "robust regression", a system that aims to overcome many of the common problems that occur in regression problems, and was designed to capitalise on the strengths of MSE and MAE while avoiding the pitfalls of both [114]. Huber loss achieves this by having different behaviour for predictions nearer and further from the mean, and is defined by a piecewise functions as follows:

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \textit{when } |y - f(x)| \leq \delta \\ \delta|y - f(x)| - \frac{1}{2}\delta^2 & \textit{otherwise} \end{cases} \tag{3.6}$$

This essentially means the huber loss can operate as MSE for small loss values and MAE for larger loss values, subject to the choice of an appropriate δ. This means that as smaller loss values are downweighted compared to larger ones, allowing training to focus more on learning to predicted outliers and other samples that the model is not yet performing well on.

3.2.4.4 Estimating Loss for Bayesian Neural Networks

The loss functions outlined above are all designed for point-estimate models. However, as discussed above, Bayesian neural networks are estimating the distribution of weights, so a different loss function is required. One of the challenges with this is that conventional SGD methods assume that weights and biases have discrete, knowable values, which is not the case in BNNs, so we need a method to convert the stochasticity of the forward pass into something more parametric in order for backpropagation to work. To do this, we define a cost function that is a combination of measures of goodness of fit and model complexity, as proposed in the original paper [173]. The complexity cost is the sum of the probability density function of the

sampled weights, in relation to a predefined simple probability density function. This means that minimising this complexity cost leads to a simpler weight distribution which will itself contribute to lower variance in the model distribution. The cost function can thus be roughly defined for dataset θ and weight ditribution w^n as:

$$C^n(w^n, \theta) = log(Q(w^n \mid \theta)) - log(P(w^n)) \tag{3.7}$$

where Q is the fitting-to-data cost function and P is the complexity cost function.

3.3 Interpretation of Deep Learning Models

A major problem with the application of deep learning models in many domains is that the automatic nature of their representation learning means it is impossible to trace the series of steps from input to prediction, rendering them essentially "black box" predictors [100]. This means that while the models themselves may show unprecedented successes at prediction and classification tasks, it is extremely difficult for a human to gain any insight to what is happening "under the hood". While performance metrics go some way towards quantifying the certainty of a given model or prediction, generating a human-comprehensible representation of the model's internal logic has a number of utilities, from allowing auditing of results for trustworthiness to permitting the extraction of useful domain knowledge [37]. For many applications, the lack of opportunity for humans to assess the trustworthiness of a model as a whole or a particular prediction of a given model is a major obstacle for their deployment. For example, in a clinical setting it is of paramount importance to patient wellbeing that clinicians are not required to blindly trust the diagnostic or treatment recommendations of a model that cannot be 100% infallible, where consequences of a faulty prediction could be catastrophic. Equally, in a scientific setting, it may not be enough to simply make a prediction - understanding the model's reasoning for predicting a certain outcome may be crucial for understanding the phenomenon that is being studied.

There is also the issue that many deep learning models do not generalise well to new data, meaning that a model may perform very well and be highly trustworthy when working with some types of data, but fail completely on a new sample that is very different than anything it has

seen before. Although training data sets are getting larger and larger, it will never be possible for a model to have seen everything it will ever encounter during training; additionally, it is often difficult, if not impossible, for a human who is using the model to identify whether a given sample is similar or dissimilar to samples it has seen before. Large training sets also bring their own issues, namely that the likelihood of spurious correlation between variables is increased as you increase the number of input features, making it more likely that the model will learn patterns that are not actually meaningful for the prediction task. Thus a way of making the model "explain its reasoning", such that the human can assess whether the basis of the prediction is sound, is also important for not falling foul of bad individual predictions, even if function of the model as a whole is sound. To this end, a number of deep learning interpretation algorithms have been developing alongside the advancing field of deep learning [156].

Applying interpretation methods on deep learning models trained on genetic data provides a unique set of challenges. In genetics, the high dimensionality and complex nature of SNP data, as well as the huge amount of variation observed in the phenotypes of individual humans, means that it is highly likely that even a generally good model will not be equally well performing across all samples. This means that, regardless of the intended purpose of the model, the prediction on a given unseen sample may not be trustworthy or meaningful and it will likely be necessary for an expert to be able to check the model's reasoning. Many large genetic datasets draw samples from a fairly homogenous group of people; for example, the participants of the UK Biobank are all between 40 and 69 years old and predominantly of European ancestry [2], meaning that a model trained using this dataset would most likely produce much more unreliable predictions on samples of different ages and/or ancestries. Additionally, since the complex genetic architectures of most common traits are not well understood, the basis of the model's decision is also usually of equal, if not greater, interest to researchers than the decision itself. However, genetic data in its usual format is extremely incomprehensible to humans; its large size and limited value range for each locus mean it is not generally possible to visualise and, unlike with an easily human-parsed format like image data, even an expert is unable to glean anything from genotype data at a glance. Thus designing a meaningful, biologically-informed method of interpretation for networks trained on genetic information is of utmost importance for their success.

3.4 Deep Learning in Genetic Data Analytics

The field of genetics is broad, with a huge variety of different types of data and analytic aims. However, these disparate sub-fields have two important things in common: our understanding of most genetic concepts and mechanisms is extremely limited, and, thanks to NGS methods, in many domains there is an unprecedented abundance of data. As we have discussed, deep learning offers several unique potentials for applications involving genetic data. Firstly, as previously mentioned, the scale of genomic datasets has increased exponentially in recent years and there are now many very large datasets available to researchers thanks to the establishment of large repositories like Biobanks. This poses many challenges for statistical genomic methods designed in the era of smaller datasets [163]. Deep learning models, on the other hand, are uniquely well-suited to handling the challenges of big data analytics and have made many breakthroughs in this field [336].

Genetics is also quite a new field, and our understanding of many concepts in this domain is very limited [53]. As the amount of genetic data available for analysis increases it becomes more apparent that many questions are too complex to be handled by the limited tools designed based on simpler ideas of genetics [293]. Methods to find and describe complex patterns in large genetic datasets are becoming increasingly important in genomic research [161]. Our lack of detailed understanding of many mechanisms makes it extremely difficult to design biologically accurate parametric models, and what's more, most types of genetic data are not obviously parseable by humans, meaning it is not obvious for even experts to spot patterns in the data. For these reasons, deep learning models, which are nonparametric pattern recognition tools which do not need input from humans in order to learn from data, are beginning to step in in multiple domains where more traditional statistical methods have reached their limits.

3.4.1 Previous Applications of Deep Learning in Genomics

One of the most successful areas of application for deep learning in genomics is in elucidating the function of regulatory motifs in DNA and RNA targets [227]. Deep learning has predicted the sequence specificity of DNA and RNA-binding proteins, detected enhancers and other regulatory regions, and learned to automatically model DNA methylation and gene expression levels, to name just a few [126, 245, 6, 147, 334, 162, 132, 157]. These models start from a variety of different types of data, including but not limited to that produced by DNase-I sequencing, chromatin immunoprecipitation (ChIP) sequencing, RNA immunoprecipitation sequencing, and transcription-factor datasets, sometimes with the incorporation of additional information from sources such as the Encyclopedia of DNA Elements (ENCODE) project. Deep learning has also been able to successfully model DNA methylation, an important predictor of gene expression, from 3D genome structure and DNA sequence [316].

For applications that use DNA or RNA sequence fragments as all or part of their input, techniques borrowed from deep learning for computer vision and natural language processing can exploit the local pattern information in adjacent base pairs or "remember" motifs that occurred earlier in the sequence. For these reasons, many of the above applications have found success using Convolutional Neural Networks (CNNs) or Recurrant Neural Networks (RNNs). This property does not translate directly to genomic prediction tasks where the start point is no longer a sequence fragment - for example, in a collection of individual SNPs identified by GWAS the local neighbourhood information is destroyed, rendering the local pattern recognition abilities of these methods effectively useless. CNNs are broadly designed for dense data [120], so the sparsity of SNP datasets (which may ignore tens of thousands of bases between each "subsequent" SNP) makes it difficult to design architecture that works for them.

Analysis of gene expression data is another growing area of success for deep learning techniques. In these applications the goal is to estimate how much RNA is produced from the DNA template of interest in a given set of conditions [46]. Many such studies make use of unsupervised methods such as autoencoders to overcome the significant size and noise of gene expression data [287, 43, 61, 327]. While dimensionality reduction is a major necessity in most problem

domains involving genomic data, the amount of noise in expression data is higher than most other types of genomic data once they have gone through appropriate quality control measures, meaning that successful DL methods must account for this. Autoencoders work well on large datasets as they are able to automatically reduce the dimensionality of the data, and they also do not require detailed labelling of a training set by an expert, both beneficial attributes for lots of genomic applications. However, the lack of known parameters adds more complexity to prediction tasks, and hence is not as well suited for clearly defined prediction on pre-determined labels, such as classification and regression on phenotype data.

Another genomic puzzle for which DL methods have been proposed is for the analysis of single variants, for example analysing the pathogenicity of coding variants [240, 138, 244], or predicting the functional consequences of noncoding variants [340]. This task is challenging to do with traditional statistical methods, and while non-deep machine learning classifiers such as support vector machine have been successful in the past [131], deep learning models offer the opportunity to exploit complex nonlinear relationships between features and therefore potentially improve performance. Variant annotation methods tend to incorporate various types of inputs, including sequence fragments and multimodal data such as histone modification or DNase-I activity, to make predictions about single SNPs [105]. A variety of different architectures are used, including basic feedforward neural nets, CNNs, and RNNs, with no obvious consensus on the best architecture. Comparison of network performance across different studies is particularly challenging when multimodal data is used as the different types of data have different properties and may suit different architectural features. Overall, predicting the properties of single variants across multiple samples, where the variant is the target of prediction, is quite a different classification problem than predicting the properties of a single sample based on the contributions of multiple variants, which considers variants as input features. So even aside from the methodological variation introduced by the multimodal data, it is not obvious if any knowledge gleaned on these sorts of tasks can be used to inform the design of multi-SNP prediction models.

A more obviously related field is the prediction of phenotype from multimodal input data, which may or may not include genotypes. These methods mimic the way diagnosis tends to be

performed by doctors, taking a holistic view of the patient that may include clinical data such as test results or images, health and lifestyle information, family history, tissue- or cell-level expression data, and genetic data such as SNPs. Deep learning models trained on multimodal data have been used successfully to predict pancancer prognosis [42], detect Alzheimer's disease stage [303], classify cases of schizophrenia [39], predict the survival of lung cancer patients [110], and predict survival times of glioblastoma patients [288]. Since it's often very unclear how to combine multimodal data from disparate sources, a major challenge is knowing which data to include; adding more types of data to a network may improve its performance, but this performance improvement is accompanied by increased risk of overfitting, and the results will become harder to interpret as the inputs become more bloated. The more data types are included, the more difficult it is to validate the results on different cohorts, as only datasets which include all the relevant data types can be used, and this kind of environmental data is particularly scarce. When studies try to design interpretability into their models in advance by integrating known information about interactions they necessarily limit what the network can discover, and information about multimodal interactions for most phenotypes is extremely limited. In all cases where a multimodal model performs well on a classification or prediction task, it is necessary to ask how models trained on the individual data types would perform, in order to quantify the relative contribution of each type of data to the prediction and therefore better understand the aetiology and risk profile of the phenotype.

While introducing additional data types can contribute to interpretation challenges, other such challenges arise from the complexity of the phenotype itself. Often a phenotype of interest may actually be comprised of several different sub-phenotypes with distinct aetiologies, which can prove confounding to a prediction model. A way of circumventing this issue is to scale down the complexity of the phenotype under investigation, for example by considering phenotypes at the cell level. This may be done for drug discovery, where the cell phenotypes of interest correlate to different drug responses in cells [141], or the goal may simply be to understand processes better at the cellular level [93]. These studies may also incorporate multi-omics data. While considering a more "intermediate" phenotype does potentially simplify the prediction problem, it can be more difficult to get datasets containing this phenotype in comparison to easy-to-measure

phenotypes like height or blood pressure, and the end goal is still generally to make predictions on or learn about the underlying features of the more complex phenotype, often it is still unclear how to make the final step from knowledge of the intermediate phenotype to biologically or clinically applicable results on the final phenotype. Since the conditions of cultured cells can be much more carefully controlled than those of humans participating in an ethical study, it is possible to have more precise information about the phenotype and hence train with much more specific labels, which can improve performance but also fundamentally changes the nature of the prediction problem. For this reason methods that work for cellular or tissue phenotypes may not transfer to less precise phenotypes.

Finally, a number of studies have been conducted on predicting phenotype from genotype in non-human organisms, such as plants [285, 172, 164], farm animals [3], or mice [335]. Often these sorts of studies have much higher rates of success with simpler linear machine learning models than is typically seen when attempting the same prediction tasks in human populations. This is in large part due to the very limited genetic diversity found in species such as crops, lab mice and agricultural livestock which have been selectively bred with a high degree of control over their phenotypic presentation. The low diversity of SNPs in selectively bred organisms means that the prediction of phenotype from genotype is a much more constrained problem space, and the methods developed for this are unlikely to be at all applicable to wildtype organisms with a high degree of genetic and phenotypic diversity.

In sum, predicting human phenotype from genotype is quite a unique problem, even within the general domain of predictive DL models for genetic data. It is also of great clinical and research interest, and as such many different research groups have attempted it using different methods, phenotypes, and approaches. These will be discussed in the following section.

3.4.2 Human Phenotype Prediction with Machine Learning

An early study by Curtis, North & Sham published in 2001 aimed to use a neural network to automatically infer haplotype and linkage information from SNP data, which led to increased statistical power over single-SNP analysis on a variety of simulated data [66]. It is worth noting

that this research pre-dates the advent of GWAS as well as many other modern statistical methods to amalgamate information about multiple disease-causing SNPs; in fact, even the investigation of SNPs as disease mechanisms was a relatively recent suggestion and there was not a clear consensus on how this analysis should be done [27]. Much research at the time was focussed on the idea of reconstructing ancestral haplotypes in the hopes of using these to infer information about the aetiology of certain phenotypes [55, 69], but at the time, (and is still the case today to a more limited extent, depending on the population) accurately mapping haplotype blocks proved challenging. Curtis, North & Sham chose to move away from a parametric approach to haplotype mapping and exploit a neural network's automatic pattern recognition capabilities to investigate the connection between a multi-SNP genotype and phenotype of interest instead. Four SNPs were used as input and the network had two hidden layers of three units each. The 10000 simulated datasets imitated a variety of haplotype models and each contained ~ 100 samples. Individual SNPs were also tested for over-representation in the cases vs controls in a manner similar to GWAS. The results were quite variable across similar datasets, indicating that neither method was especially robust, but the neural network generally showed greater power to discover associations, a trend which became more marked as the number of input variants was increased, even if some of them were not actually associated to the phenotype, and persisted whether the linkage of the datasets was weak or strong. While the authors acknowledge that their simulation will not be entirely accurate to real genetic data and an input of four SNPs is very small by the standards of modern studies, these early results on simulated data, in a study conducted even before the genesis of GWAS, already indicate the potential of neural networks to detect patterns in genotype data. The authors make many recommendations for the direction of future work in this area, suggesting that more complex networks with more SNP inputs could successfully model complex polygenic traits, and recommending the use of a quantitative phenotpye value as the target for training a neural network. They also highlight the potential of neural networks to detect SNP interactions, but acknowledge the difficulty with extracting this information from a trained network.

Another early study that used synthetic data to test neural networks for the identification of associated SNPs for complex traits was conducted by Bhatt, Lucek, & Ott in 1999 [19]. Inputs to the neural network were in the form of alleles rather than SNPs, and 1893 alleles were included.

The networks were tested on 10 simulated datasets with $\sim$ 10000 individuals per dataset. The network performed multiclass classification, dividing individuals into unaffected, mildly affected, and severely affected by the simulated phenotype, and it contained a single hidden layer with fifteen nodes. Information about how the architecture was chosen was not included in the study. After training, the network had a mean squared error of < 0.005 across all datasets. Once the network was trained, the weights of the trained network were interpreted to indicate which input SNPs were most important. The performance of the network was mixed, with it successfully detecting a number of true positive loci but also a greater number of false positives. The results of this study indicate the potentials of neural networks to overcome some issues associated with traditional statistical methods (for example, issues arising from multiple testing) but also highlight how accurately extracting information from a trained network can pose real difficulties even if the classification performance of the network is very high.

In a 2020 study by Behravan et al., a machine learning approach was used to classify Breast Cancer, using a combination of interacting SNPs and demographic risk factors as input features [15]. In an earlier study, they had used an gradient tree boosting method (XGBoost) followed by adaptive iterative SNP search to capture groups of SNPs that interact in a complex nonlinear way to increase breast cancer risk, which when combined with a Support Vector Machine outperformed the state of the art approach of GWAS and PRS for classifying breast cancer cases[16]. In this paper, they reused their existing SNP selection tool to identify SNPs to be used as features for a deep learning model. These were then combined with different types of demographic factors (those related to family history and those related to oestrogen metabolism) to create multimodal datasets for prediction. The optimal number of SNPs for classification was found to be 18 (where the smallest number trialled was 9 and the largest was 37). Each group of features (the SNP features and the two sets of demographic features) was evaluated separately as a baseline, and then combinations of feature groups were evaluated together; the best combination of groups was the SNP features with demographic feature group 2, which led to a mean average precision of 78.00. The authors argue that this demonstrates that the aetiology of breast cancer is governed by interacting genetic and demographic risk factors. Feature importance was later calculated by iteratively leaving out features during training and observing how this affected classification

performance, indicating the relative importance of each feature to the overall prediction and highlighting a number of interesting SNPs and lifestyle factors. Feature importance was then validated using SHAP values, which use a game theory-inspired approach to estimate the contribution of input features to a prediction, which can then be used to compare the importance of individual features against each other [260, 183]. Analysis was then performed to determine if any of the identified SNP and lifestyle features had interactions, where interactions are defined as the features appearing together during the iterative feature importance process. While this study was more focussed on the interaction between genetic and lifestyle factors than the genetic interactions alone, and it did not include a very large number of SNPs, it does demonstrate the potential utility of perturbation-related analysis for the interpretation of neural networks trained on genetic data. The authors also showed that their own SNP selection method provided a more successful feature set than SNPs taken from literature (most of which would have been identified by GWAS), and that all of their different deep learning models outperformed a PRS metric developed with the 82 literature-identified SNPs, suggesting that novel methods offer the potential to improve upon state-of-the-art protocols.

Badre et al. also examined Breast Cancer as a phenotype, but utilised deep learning instead of a classical ML method such as XGBoost [13]. They found that in comparison to a series of classical ML models and PRS, the deep learning model offered the best prediction performance. They then used DeepLift and LIME to extract important SNP features from the network. The input for the PRS, ML and DL models was sets of breast cancer-associated SNPs derived by GWAS from the DRIVE breast cancer project. To evaluate the effectiveness of different models Receiver Operating Characteristic curves were plotted, and the Area Under Curve (AUC) metric was calculated from these, with greater AUC indicating a more optimal tradeoff between specificity and sensitivity and hence corresponding to better model performance [25]. Different sized input sets, produced by using different p-value cutoffs for filtering as well as using no p-value filtering at all, were tested, and they found that using an intermediate amount of filtering (cutoff of $10^{-}3$) gave the best results, with a more strict cutoff losing important information and a less strict cutoff leading to the network overfitting; however, the fact that the differences in performance between the highly filtered and completely unfiltered SNP sets were not large

($\sim$ 1% AUC) suggests that the network was easily able to identify relevant features without p-value filtering. They also tried using an autoencoder instead of p-value cutoff for reducing the input dimensions, but found this did not improve performance. The best performing DNN had three hidden layers of 1000, 250 and 50 neurons respectively, outperforming a shallower NN with only one hidden layer and a one-dimensional convolutional neural network, as well as a variety of linear ML models and several different PRS metrics derived from different SNP sets using different software, to produce a 67.1% AUC metric. They found that when operating at the kind of precision required for clinical applications (90%) the recall of the DNN was significantly higher than any of the PRS metrics. While the DNN performance was better when a relatively large number of SNPs were included as input, Badre et al. hypothesised that the majority of these SNPs were not actually important for prediction, and so used LIME and LRP to extract the top-100 most important SNPs for identifying high risk individuals. They found that a minority of significant SNPs identified by these methods had p-values below the cutoff threshold, which suggests that they may have nonlinear relationships to disease outcome, which they confirmed to be the case with a subset of these SNPs. While the two interpretation methods did not yield much overlap in their important SNP sets, and many SNPs previously identified as important to breast cancer were missed by both interpretation methods, several known breast cancer associated SNPs that were missed by GWAS were identified by at least one of the methods. This suggests that deep learning and interpretation of resulting networks offers the possibility of uncovering associations missed by current methods, but that this particular approach to interpretation is highly limited.

Tomita *et al.* trained several neural networks to classify cases of childhood allergic asthma in a Japanese cohort, using a selection of 25 SNPs located in 17 genes that had been previously associated with asthma in other studies [296]. This study built on a previous work classifying allergic diseases from SNP data with neural networks [295]. The networks had one hidden layer and a variety of input widths depending on the SNPs used as features; no information is given in the study about how the network architecture was chosen. All of the neural networks outperformed a set of logistic regression models which were trained as a baseline; however, it is worth noting that the training and test sets were very small ($\sim$ 100 samples), meaning

that the neural networks would be very prone to overfitting to the dataset, and validation on an independent dataset was not performed. A variety of parameter decreasing methods were used to reduce the input SNPs from the original 25 without compromising the model performance; this eventually led to the smallest high-performance input set of 10 SNPs. Combinations of SNPs within the set were then analysed using linear statistical methods for over-representation in cases as compared to controls, while controlling for likelihood of co-occurrence of alleles. A number of suggestive 2- and 3-SNP interaction combinations were identified, though the authors acknowledge the difficulty of choosing a significance cutoff and of evaluating the prevalence of these combinations in such a small dataset. This study provides a possible pipeline from disease classification to interaction analysis and possible biological insight; however, it is difficult to know how this would scale to larger datasets and feature sets containing more SNPs, both of which would be necessary to increase the predictive power of the method.

Mieth et al. developed DeepCOMBI to directly utilise explainable AI methods for the purpose of finding phenotype-associated variants [190]. Specifically, they use Layerwise Relevance Propagation to extract a vector of the most significant SNPs for each sample, based on predictions by a deep neural network trained to predict phenotypes from SNP data. They then analysed which SNPs tended to be considered most explanatory across the whole dataset, and performed statistical association tests on these to determine p-value, with the aim of producing a list of ranked associated SNPs as would be produced by a GWAS, but generated in such a way that there is no assumption of linearity or inter-SNP independence. The deep neural network, the architecture of which was inspired by previous studies [198] had two hidden layers with 64 neurons each and made use of dropout to prevent overfitting. SNP selection performance was evaluated on synthetic data and compared against a previous SVM-based method [189]; they found that DeepCOMBI outperformed the SVM and was better able to detect less influential SNPs and avoid noise. It also outperformed a GWAS p-value thresholding method. In order to limit the computational complexity of the model, only SNPs that passed a predetermined p-value threshold were used as inputs for the Deep COMBI, with the rest being discarded. This does limit the model's ability to discover truly novel variants, as the interpretation step is likely to closely reproduce the input SNPs and important SNPs may have already been excluded by

the filtering step. It is possible that DeepCOMBI would not have outperformed GWAS if the GWAS-based filtering step hadn't been performed first, meaning that it doesn't straightforwardly offer a replacement for GWAS at the moment.

A common limitation of all the studies discussed so far is that, when interpretation methods have been used, they have only been able to uncover individually significant SNPs rather than important interactions, genes or pathways, limiting how much the research is able to reveal beyond results that would be within the scope of a typical GWAS. To overcome this, Van Hilten et al. developed GenNet, a network whose architecture is designed using biological priors to aid interpretation by backpropagation-based methods [301]. Because biological knowledge (from public databases) is used to create a network with only biologically plausible connections, the resulting network is much sparser than typical fully-connected networks, making it more efficient to run and therefore able to use larger amounts of input data, eliminating the need for the p-value threshold filtering that most of the other methods require. The architectural framework of GenNet involves first clustering SNPs according to associated genes using gene annotations, such that the millions of nodes corresponding to SNPs in the input layer may be connected to a single node representing a gene in the first hidden layer. From here many gene nodes may connect to a single local pathway node, and so on, with various different options for progression depending on which kind of annotations are being used. After a prototype network was tested on simulated data and determined to be interpretable under the right conditions, the method was applied to population-based data for a variety of phenotypes. The models performed well across a variety of phenotypes, matching or outperforming the baseline LASSO logistic regression models in all cases. Predictive performance was better for less polygenic phenotypes such as eye and hair colour, but the model for schizophrenia, a highly complex phenotype for which the genetic basis is not well understood, also performed well with a test AUC of 0.74. However, in a follow-up experiment designed to test whether it was the biological prior information or merely the generalised network architecture leading to the performance improvement, a randomised version of GenNet in which the greater network structure was the same but the connections had no biological meaning vastly outperformed the network with embedded priors for schizophrenia, suggesting that the network was better able to learn a representation of the underlying genetics of

schizophrenia automatically, but destroying any possibility of interpretation. This result suggests that designing networks with biological prior information can improve prediction performance across the board, but risks the possibility of misleading interpretation results when the model is not actually making use of the biological elements of the architecture. This is especially likely for complex polygenic phenotypes such as BMI, for which the biological priors may be limited or unreliable due to lack of understanding of the genetic basis of the disease, meaning they are the hardest to construct biologically sensible networks for, and thus the least likely to be interpretable via this method.

Another study which aims to make use of prior knowledge about genetic data to construct an appropriate network architecture is that of Yin et al., who designed a network to predict cases of amyelotrophic lateral sclerosis (ALS) from phenotype [217]. ALS was chosen as a phenotype as it is believed to have a complex genetic aetiology with many nonlinear interactions between causal SNPs, which makes it particularly difficult to interrogate using the state of the art linear methods, and the majority of the disease's heritability remains unexplained [302]. To avoid the limitation of numerous studies in which important variants are preemptively discarded by using GWAS as a filtering step, the authors chose a different filtering method of only considering SNPs in regions of the genome that have been identified as regulatory, making use of prior biological knowledge that shows that pathogenic variants are over-represented in regulatory regions. However, this filtering method only considers SNPs found in promoter regions from several chromosomes believed to be more associated than most with ALS, so it has the potential to run into many of the same issues with GWAS-dependence/excluding causative SNPs as other studies, as well as risking missing ALS associated genes on other chromosomes [290]. A two-step approach is then employed, where first individuals are classified using only SNPs from individual promoter regions, and then the promoter regions which correspond to the highest prediction accuracy are combined for classification by a later network. Like GenNet, this results a pipeline which is designed for interpretation, meaning that methods identifying the features most important for classification will be automatically biologically meaningful. The models were trained on data from the Dutch cohort of project MinE, which has $\sim$ 10000 samples. Both classification models are Convolutional Neural Networks (CNNs), and the reasoning for this

is that CNNs have achieved the best prediction results in image-based applications, though as the authors note genetic data is very structurally different than image data and it is difficult to know how to design biologically informed CNNs. Encoding is based on minor allele frequency, which is an interesting choice in contrast to other studies; sadly, the authors do not discuss the rationale behind this or use any other encoding methods for comparison. The study found that only a small subset of promoters appeared to be important for classification, and some of these were associated with genes previously associated with ALS but many were not. Different machine learning methods appeared to reveal entirely different sets of important promoters, suggesting that these findings may not be very robust. ALSNet outperformed linear methods in combination with the promoter selection CNN, but there was no comparison with how ALSNet would have performed on SNPs from randomly chosen promoters, and in fact the promoters from chromosome 22 appeared to yield an equivalent result to those on the other chromosomes, despite the fact that chromosome 22 has not been identified as important for ALS and was included as a control. Additionally, performance was generally improved by combining promoters across all chromosomes rather than training on separate chromosomes, further suggesting that the initial selection step may not have improved prediction and in fact may have excluded useful information.

Muneeb & Henschel applied a variety of machine learning and statistical methods to predict eye colour (between blue and brown) and type 2 diabetes status from genotype. All of the methods tested for eye colour had high classification accuracy of $> 90\%$ and there was not much difference between the various methods, though an ensemble of LSTM models had the best performance. Information about the model architectures or how these were designed was not included in the paper. Only random forest was tested for type-2 diabetes, resulting in a test accuracy of 97%. A relatively small dataset of ~ 100 people was used for eye colour and an even smaller dataset of ~ 10 people was used for type-2 diabetes, making it likely that complex deep learning models would have significant overfitting issues on this data. A variety of different numbers of SNPs, ranging from 3 to 1560, were used as inputs, and there was a general performance improvement across all models with the 1560 input set as opposed to the smaller sets. This suggests that the inclusion of more insignificant SNPs may improve classification performance. However, it is dif-

ficult to draw concrete conclusions from the results of this study as the sample sizes were so small.

A study which attempted to take a systematic approach to phenotype prediction from genotype was by Bellot *et al.* [17]. They explored a variety of phenotypes that are known to be polygenic, including height and body mass index, and used a genetic algorithm to try to automatically determine an optimal deep neural network architecture for this type of regression. They used both convolutional and basic feedforward neural networks, trained on data from the UK Biobank, and compared results to those from a variety of Bayesian linear models. They found that the performance of different models was very varied, but on the whole the deep learning models did not significantly outperform the linear models, and the performance of each architecture varied greatly across the different phenotypes, suggesting that the underlying genetic mechanisms of different phenotypes may look extremely different from one another and hence there might not be a network architecture that performs uniformly well at predicting different phenotypes from genetic data. However, the linear models used (BayesB and Bayesian Ridge Regression) were more stable across phenotypes, suggesting they might be more robust. The authors postulate that traits with more additive SNP contributions, such as height, are better fits for linear models, whereas deep learning models were more competitive on traits with more complex genetic interactions such as BMI. A GWAS was performed as an initial SNP-filtering step and the authors did find that including more variants as input improved the performance of both the deep learning and linear models, even though for the larger input sets the vast majority of included SNPs did not come near the significance threshold of the GWAS, which suggests that there is likely a genetic background effect from the combinations of weakly significant SNPs which is not captured by GWAS. They also trialled different selection methods for filtering SNPs; as well as just choosing the top N hits from GWAS, they implemented a "uniform" selection method whereby the genome was divided into N equal segments and the top GWAS hit was selected from within each segment. This was an attempt to account for the effects of linkage disequilibrium, though determining the best segment size to mitigate LD effects while not losing useful information is nontrivial, and in the end the more straightforward top N selection method outperformed the uniform selection method across the board, suggesting that autocorrelation among input SNPs is not a major issue for deep learning models. It was also hoped that the uniform selection method

would help CNNs exploit spatial information between local variants, but again this was not successful, and the authors concluded that more work is needed to figure out how to prepare SNP data for CNNs. Bellot *et al.* also tried both the standard variable encoding and one-hot encoding, finding that one-hot encoding did not improve network performance for any of the traits. The authors conclude that more work is needed to determine both the optimal encoding and the best network architecture for genotype-phenotype regression problems.

Montanez *et al.* also investigated predicting phenotype from SNP data pre-filtered for BMI association, but to avoid discarding useful SNPs they used logistic regression instead of GWAS as an initial filtering step [198]. They then modelled BMI as a binary classification problem with the dataset divided into obese/healthy weight classes, and trained a deep learning model to predict obese individuals. In an earlier study that compared different ML methods with neural networks for the same task using 13 associated SNPs as input features, Montanez *et al.* found that support vector machine performed best with 0.905 AUC [197]. However they then went on to test a neural network with larger groups of SNPs (resulting from less stringent p-value thresholding) and found that including more SNPs improved network performance (where the SNP sets ranged in size from 5 to $\sim$ 2000 SNPs). It is worth noting that the sample set used for this study (derived from personal health and genetic data from patients at a clinic in Pennsylvania) is quite small ($\sim$ 1000), which means there is a high likelihood of neural networks trained on this data alone to overfit and generalise poorly to new data. Overfitting is potentially indicated by the very high ROC AUC scores of the best performing network ($\sim$ 0.99). Montanez *et al.* then went on to utilise Stacked Autoencoders (SAE) and Association Rule Mining (ARM) to evaluate a set of filtered SNPs for epistatic interaction in cases of extreme obesity and in normal samples[195]. The SAE is used to perform automatic dimensionality reduction on the input SNPs and ARM is used in conjunction with frequent pattern matching to try to detect groups of SNPs that frequently appear together and create rules to govern their interactions. The resulting identified patterns were then further tested for interestingness and pruned to account for redundancy, before being used to design a neural network with a reduced input feature set. This approach appeared to lead to worse classification performance than simply using the neural network alone with all the features, and the gene associations generated by ARM did not conform well to prior biological

knowledge, as no well-known obesity-causing genes such as FTO make an appearance in any of the rules. The small dataset may again be to blame for this, which the authors acknowledge - however it is difficult to benchmark the effectiveness of this approach due to this. As well as being hard to verify, the detected epistatic networks are also difficult to interpret at the SNP level.

Muneeb, Feng and Henschel used transfer learning to predict phenotype from genotype with the aim of improving the prediction of disease risk in small populations which have insufficient data for model training [215]. They point out that PRS are often not transferrable to different human populations than the ones they are designed with, but that a model where important SNPs could be extracted from data from a large population could then potentially be used for risk prediction on a small population. DL models have more potential predictive power than classical ML methods but tend to not work well on small datasets, so by training on a large dataset and then finetuning the model to work better on the small dataset with transfer learning, they were able to gain greater predictive power on the small population. Only synthetic data was used for both training and validation, as the real data was incomplete and had issues making it difficult to train with; however, this raises the question of how well this kind of method would function with real data, which is often messy. A relatively small number of causative SNPs were used to train the DL models, again suggesting that the method would not generalise well to real phenotypes which have large numbers of significant SNPs; additionally, a large number of causative SNPs would likely further amplify differences between populations, making transfer learning less effective. The study also used a variety of recurrent neural network architectures, but didn't clearly explain why these were used instead of feedforward nets or CNNs.

3.4.3 Gene-Gene Interaction with Deep Learning

Machine Learning methods have also been exploited to model complex nonlinear gene-gene interactions. These have the advantage of not necessarily needing an underlying prior model for the data in order to make predictions, as well as being better able to detect complex nonlinear interactions between many SNPs. Random Forests have been adapted for the structure of GWAS data to detect SNP-SNP interaction and use it for prediction, leading to improved prediction

results over standard linear models which indicates that the interaction effects are important for prediction [23]. Further iterations of random forest models for SNP-SNP interaction also included methods to extract the SNP interaction pattern detected by the random forest [331, 122]. Support vector machines have also been trialled with some success, though they are prone to type-I error and are computationally expensive [44]. In general, ML methods still encounter issues with excluding potentially important SNPs (since models can't have too many parameters to limit computational complexity, making some feature selection necessary), and also tend to suffer from an inability to generalise to new datasets and phenotypes.

Deep Learning models can handle more parameters and model interactions of greater complexity, but also have even more potential to end up overly specialised to the dataset and unable to make predictions on new data. They also have the major pitfall of being "black boxes", making it highly nontrivial to extract the information about exactly what interactions are being modelled, effectively rendering the modelling useless. Fortunately, deep neural network interpretation methods offer a way to determine interactions after the fact.

Cui *et al.* designed a sparse neural network architecture that groups input SNPs into relevant genes, and then trained this network to learn the relationship between genotype and phenotype data (using several phenotypes related to cholesterol, but the method is intended to be generalisable) and used Shapley scores to quantify interactions between genes in the trained network [62]. They also provide a customised permutation test to assess the significance of detected interactions, involving a separate comparison neural network that makes the assumption that genes do not interact and contains only a linear layer after the SNP-to-gene mapping. The method was initially developed using synthetic data and benchmarked against a variety of linear methods including linear regression, lasso and XGBoost. They also tried a variety of SNP representations and permutation tests, and compared the area under the receiver operator characteristic curve across methods, and found that their method significantly outperforms existing methods at detecting interactions. They found that the neural networks significantly outperformed the linear models, and that the neural networks with the permutation test they devised showed greater specificity than the same networks using existing permutation tests. Benchmarking of different

gene representation methods also indicated that the networks were able to learn more from using all available SNPs for each gene than they were from either using only the most significant SNP or from using phenotype-agnostic reduction methods like PCA; however, on synthetic datasets where there are only a few causal SNPs or the interactions between genes are very simple the top-SNP approach led to good prediction performance, suggesting that the relative effectiveness of neural networks vs linear models for modelling gene interaction is probably phenotype specific and dependent on the complexity of the genotype-phenotype relationship. In real data, using cholesterol as a target phenotype, the proposed network and permutation test approach identified 19 candidate gene-gene interactions for further research, 9 of which were replicated in an independent dataset. The successful networks did require significant feature filtering in that only genes previously associated with the target phenotype in the literature were included as inputs, due to concerns about computational complexity.

Another study by Lee *et al.* focussed not only learning on gene-gene interactions for BMI but also gene-environment interactions as well [150]. They used the Generalised Multidifactor Dimensionality Reduction (GMDR) method to select SNPs, DNA methylation sites, and dietary and lifestyle factors that passed statistical significance in BMI association in their small ~ 1000 cohort from the Framingham Offspring Study. They found that gradient boosting algorithms performed the best out of the ML models trialled (DL models were not tested). BMI was modelled as a multiclass classification problem with obese, overweight and healthy weight classes. GMDR uses permutation tests to look for significant combinations of features, and the study then used pathway enrichment analysis to link identified genes and DNA methylation sites to functional genetic information. Out of the individual SNPs and DNA methylation sites identified by GMDR, only a minority were in the vicinity of genes associated with BMI in previous studies. However, they detected several interactions which had been identified in previous studies. They also found that DNA methylation sites were generally more useful predictors than SNPs. Synthetic data was used to test and quantify the contributions of different factors, with a particular focus on lifestyle factors. While these results are promising for machine learning applications for gene-gene interaction, the small cohort and absence of interactions testing on real data makes it hard to compare these results with those of other studies.

As discussed, one of the major challenges of applying deep learning to previously unexplored problem domains lies in the appropriate choice of hyperparameters. The goal of the creators of GPNN was to design a pipeline to perform hyperparameter selection for deep neural networks to model gene-gene interactions [257]. They utilised a genetic algorithm [140] to perform architecture optimisation, and tests on simulated data containing some epistatic effects revealed an improved performance over a neural network designed without the optimisation step of the genetic algorithm (where the architecture design simply followed a trial-and-error approach). The differences in prediction accuracy were relatively low between GPNN and the network designed without a genetic algorithm, but it did show markedly improved power to detect epistatic effects, especially when effects including more SNPs were simulated or when the overall heritability was lower, and it showed less propensity to overfit to the data. Overall, the results suggest that using a genetic algorithm to design a neural network for gene-gene interaction detection can improve the performance of the resulting network over other design approaches, which could be an important contribution since designing network architectures for genetic applications is currently very challenging. The authors do acknowledge that the resources required to do this are far more significant than those needed for the trial-and-error design approach. The authors did not include information about what tools were utilised for simulating epistatic effects, and no other epistasis detection methods were used as a baseline comparison on the simulated data, which makes it hard to fully assess the success of the approach. The simulated data did only contain a maximum of ten SNPs, however, which is certainly on the smaller side for a complex polygenic phenotype; this makes this method less useful as a SNP interaction discovery tool and more useful to verify previously suspected interactions. A later study evaluated the performance of GPNN on real data, investigating gene-gene interactions in Parkinson's disease on real data (the study does not indicate what the source of the real Parkinson's disease data was or how many SNPs/individuals were used) [211]. In this study, a multi-SNP stepwise logistic regression model was used as a baseline. Both the LR model and GPNN replicated earlier results of a Parkinson's disease epistatic interaction involving the *DLST-234* gene, and otherwise the results were generally similar, aside from the LR model identifying more SNPs than GPNN. These results seem to indicate that GPNN performs well on real data, though it is difficult to determine

whether it outperforms existing methods. The authors also acknowledge that it will be important to try to replicate these findings in datasets with a larger number of samples and input variants.

An attempt to create an end-to-end framework for first identifying plausibly interacting SNPs and then detecting epistatic interactions between them has been proposed by Moore *et al.* [203]. The goal is to integrate the successes of Multifactor Dimensionality Reduction with the strengths of more sophisticated predictors such as machine learning models. First a filter step is used to identify SNPs that are more likely to be involved in interactions, after which MDR is used to create new multi-SNP features that are indicative of possible interactions effects. Several different statistical filter steps were trialled [117, 258]. Then the possible interactive combinations identified by MDR can be used to inform the design of a predictor model; the final stage will then be to use interpretation methods on the trained neural network to extract information about the detected interactions. The study found that using MDR as a step to design features for a classifier improved the classifier's performance at classifying phenotype and detecting interactions on both simulated and real data. It is worth mentioning that the predictive model trialled was a naive Bayes classifier, which can be much more easily interpreted than a black box model such as a neural network, meaning that the fourth step in the pipeline would be significantly more challenging in this case. It is also worth noting that even the first filtering step was performed on only the SNPs in a single gene that had already been linked to the phenotype by a previous study, so the method actually relies on biological priors to perform a pre-filtering step, and it's not clear how well it would scale to the full genome.

Another study which proposes an end-to-end pipeline for designing neural networks for gene-gene interaction detection makes use of a grammatical evolution algorithm to optimise network architecture, resulting in a Grammatical Evolution Neural Network (GENN) [213]. GENN is shown to significantly outperform GPNN on simulated data with varing allele frequency, heritability, numbers of polymorphisms, and numbers of interactions. Simulated datasets included as many as half a million SNPs to assess scalability (though SNP sets larger than 1000 were not evaluated for model power due to complexity constraints), and contained no marginal contribution in the effort to assess detection of interactions in the absence of main effects. The

authors acknowledge that the accuracy and relevance of simulated datasets in representing biological data is somewhat untested, but argue that improved performance on simulated data is nonetheless likely to translate to improvements in real datasets. A later study used simulated data that deliberately included noise in the form of missing data, error and heterogeneity and found that the performance of GENN was generally robust to this noise, being at worst comparable to MDR [214]. Several tests indicated statistically significant difference between the powers of the two networks. The performance of both GENN and GPNN was also compared against MDR in an HIV immunigenomics dataset, to see if the discovered interactions would agree between the disparate methods, and it was found that only GENN was able to replicate the findings of MDR, with GPNN missing most of the major epistatic interactions. A decision tree was extracted from the trained neural network to interpret the findings of GENN and display them in a way that was human-parseable (this step is not discussed in detail in this study but it is worth noting that it can be nontrivial and computational expensive to do this, including potentially involving another genetic algorithm [143]). The results of this study indicate that GENN is highly powered to detect two-locus interactions, with power decreasing significantly for higher order interactions, and that it is able to do this in reasonable computational time, even when heritability is low. The interpretation method (extracting a decision tree) poses some concerns for hypothesis generation in larger datasets or for higher order interactions as the resulting trees can become difficult to parse. The method also applies only to classification problems and cannot currently be extended to regression.

Another proposed end-to-end pipeline for designing neural networks for non-additive gene-gene interaction detection using evolutionary algorithms is ATHENA [298]. This method makes several modifications to the genetic algorithm used to produce GENN with the goal of including some of the benefits of GPNN as well. The authors also investigated whether incorporating domain knowledge into the network design evolution by using the Biofilter tool to mine data from publically available databases further improved performance [29]. The advantage of this approach is that it allows domain-specific knowledge to influence the final network without limiting it, which avoids the usual pitfalls of deploying such methods for phenotypes where domain knowledge is very limited. Analysis was performed using simulated data of 500 SNPs

where two had a small main effect and a larger interaction effect on the phenotype and the rest were unassociated "noise SNPs". They found that using the updated genetic algorithm incorporating tree-based crossover and backpropagation improved the prediction performance of the resulting neural network, and that incorporating domain knowledge into the genetic algorithm further improved the resulting network's performance. This suggests that the careful and limited incorporation of domain knowledge can improve the design of gene-gene interaction detection networks without limiting them in the cases of limited or incorrect domain knowledge. As usual there is an open question of how this method would transfer from synthetic data to real data, and the authors did not offer any suggestions for probing the resulting neural network to interpret the detected interactions.

Gunther *et al.* also performed a study on simulated data to assess the feasibility of detecting two-locus epistatic effects with neural networks [101]. Logistic regression and MDR were used as baselines. In the majority of cases, the neural networks better captured the underlying disease architecture than either the LR models or the MDR approach across a variety of simulated datasets containing different epistatic mechanisms; the major exception to this rule was for multiplicative interaction models, in which the logistic regression models performed better. The NNs under investigation had at most one hidden layer and were feedforward and fully-connected; no information is given about how their architectures were chosen. Model accuracy was evaluated by estimating the penetrance using the models and comparing results to the real penetrance matrices of the simulated data. Additionally, attempts to interpret the fitted models indicated that the results of MDR and the interaction terms of the more complex LR models did not yield correct results about the interactions in the data. The results of this study indicate that even relatively simple NN models can more effectively capture the underlying structure of genetic interactions that logistic regression or MDR, making them a good tool for further investigation. The authors of the study indicate that the major challenge remains in effectively interpreting the neural networks due to their black box nature, and recommend that future work focuses on interpreting neural networks for genetic interaction. They also postulate that the improved performance of the LR model for the multiplicative case is due to the fact that the multiplicative interaction model adheres exactly to the structure of an LR model with

interaction terms, a situation unlikely to occur in real data.

Overall, the problem space of neural networks for gene-gene interaction detection still retains a lot of unexplored ground, particularly when it comes to larger studies on real data.

Chapter 4

Body Mass Index and The UK Biobank

The purpose of this chapter is to discuss the phenotype, Body Mass Index, which I have selected for further analysis with deep learning, to give the reader a better understanding of its attributes and why it was chosen for this PhD project. Different phenotypes have different properties in terms of factors such as population distribution, heritability, and risk-SNP distribution on the genome, which means that predicting phenotype from genotype can quite a different task depending on the phenotype under consideration. The choice of phenotype and model both may inform the selection of an appropriate dataset, and attributes of the dataset in turn may have a large influence on the success of a method. To this end, this section also discusses in detail the source of the data used in this PhD project, the UK Biobank. Together, these will inform the research outlined in the subsequent chapters.

This chapter is divided into two main sections. In the first section, we will investigate the phenotype of Body Mass Index (BMI). We will provide an overview of the current understanding of the genetic basis of BMI and discuss the questions that remain unanswered. We will also consider what aspects of BMI make it a good candidate phenotype for deep learning. In the second section we will talk about the UK Biobank, which provided the data for the dataset used in this study. We will outline the various preprocessing steps that were performed before this data was used for training or testing models. The aim is to provide a detailed understanding of the data used in this PhD project so as to give a better understanding to the results described in the next chapter.

4.1 Genetics of Body Mass Index

Obesity is a serious public health issue that is on the increase globally [194]. Obese individuals experience a reduced quality of life and also put significant strain on health services, as obesity increases susceptibility to a number of debilitating health conditions, including but not limited to diabetes, hypertension, and cancer [236]. Research such as twin studies has shown that BMI, a metric commonly used to quantify obesity, is highly heritable, and that 40-70% of the inter-individual variability in BMI can be attributed to genetic factors [78, 175]. While the increase in obesity prevalence is undoubtedly linked with environmental changes, marked differences have been observed in the phenotypic response of individuals to obesogenic environments [169, 5, 81]. Despite this, relatively little is understood about the genetic processes of obesity susceptibility, and only a limited number of causal genes have been identified; BMI appears to be a highly polygenic phenotype with complex and heterogeneous mechanisms underlying its risk and development [128, 87].

While there are monogenic forms of obesity, where a deletion or single-gene defect leads to severe early-onset obesity [21, 230, 310, 330], the underlying mutations are rare and thus not likely to constitute the sole cause of the more commonly observed development pattern of obesity [169]. The more common form, whose inheritance follows patterns seen in many other complex traits, is generally believed to be polygenic and governed by a large number of polymorphisms, predominantly with low pentrance and small effect sizes [108, 50, 127]. GWA studies were not able to reveal any confirmed BMI-associated common SNPs until 2007, when a number of important variants in the FTO gene region were identified [82], a finding which has been replicated numerous times [269]. Evidence of the importance of this region for obesity has been replicated by later studies as well as observed in related phenomena such as type-II diabetes risk, but the penetrance of the most significant variants is low and the phenotype contribution is relatively modest, with a 3kg weight increase being associated with homozygosity for the risk allele [85, 232]. As only 16% of adults are homozygous for this allele, and given the prevalence of obesity in most populations, it is clear the FTO variant only goes a small way towards explaining the aetiology of obesity.

A GWAS meta-analysis performed by Locke *et al.* in 2015 analysed the genotypes of >100000 individuals and uncovered 97 loci associated with BMI, 56 of which were novel [165]. Many of the novel loci had lower minor allele frequency and smaller effect sizes than those identified in previous studies, suggesting that the larger number of samples was required to reveal the full complexity of obesity aetiology. Pathway analysis performed on the most significant SNPs revealed strong evidence of central nervous system involvement in the development of obesity. Together, the 97 associated loci account for 2.7% of BMI heritability, which still leaves the vast majority of the heritability of this phenotype unexplained. Since then a further meta-analysis containing more than 700000 European-ancestry individuals has discovered 750 loci, cumulatively explaining 6% of the heritability of BMI [329].

While there is evidence that the contributions of rare variants may explain some of the unexplained phenotypic variation in BMI [322, 21], it is not currently clear how to integrate these into prediction models, especially because there is less of this data available and rare variant association studies require even larger datasets [139]. PRS metrics that include rare variants do not seem to lead to markedly improved obesity prediction over PRS with only common variants [318]. It appears that even for individuals with known forms of monogenic obesity, the genetic background of common variants still influences phenotypic expression, suggesting a possibly continuum of obesity from monogenic to polygenic [38, 50]. Interactions between environmental factors and BMI risk variants are difficult to measure, but robust reproducible interactions have been found for several variants including FTO [241, 129].

The vast majority of BMI-associated variants discovered by GWAS map to non-coding regions of the genome [169]. This makes it difficult to translate the findings of GWAS into biological insights about the nature of obesity, as it is less obvious to connect SNPs to relevant genes and functional pathways when they are located in non-coding regions [33]. However, through gene set analysis it has been discovered that the central nervous system plays a major role in the genetic basis of BMI via the regulation of appetite [165, 277, 191], with similar neurobehavioural mechanisms also being observed in some forms of severe obestiy caused by rare variants [309, 72]. Despite the clear significance of several variants within the FTO

gene, the actual mechanism by which the FTO gene influences BMI has yet to be discovered, and the fact that a number of studies have uncovered evidence of different functional bases for FTO involvement suggests that the variants governing BMI may have a significant degree of pleiotropy [339, 168]. Several epistatic effects have been discovered between BMI-significant SNPS [77] and it is theorised that further epistatic interactions may account for some of the missing heritability in this phenotype, but these effects are difficult and time consuming to identify and validate with current methods.

As discussed, BMI appears to be a phenotype with a large number of contributing common variants, and polygenic scores that incorporate larger numbers of variants than are initially identified as significant by GWAS have been shown to be more effective at stratifying individuals into risk categories and predicting BMI [128, 329]. While larger datasets offer new possibilities for high performance PRS scores, with Yengo *et al.* finding a correlation of 0.22 with actual BMI for their PRS, they also create new challenges for quality control methods such as population stratification, which may overcorrect and remove too much of the signal [329]. Khera *et al.* trialled several polygenic scores using LDPred [239] and found that the one including all 2 million variants outperformed the various subsets that reached genome-wide significance to achieve a correlation of 0.292 with BMI [128]. This high perfoming polygenic score was then analysed for effect on BMI and other phenotypes over the individuals' lifetimes, yielding new insights into developmental pathways of obesity and its relation to other traits. The success of this score as a predictor suggests that a large number of SNPs are involved in the genetic influence of BMI and that a large proportion of interactions between those SNPs are linear. It is difficult, however, to analyse a 2 million variant score for information about relevant genes or pathways.

Overall, the genetic basis of BMI is a highly complex polygenic architecture which is currently not well understood, and much of the disease's heritability is still "missing". Though tools such as GWAS and PRS have identified many associated variants and gone some way towards explaining the phenotypic variability, mapping these variants to functional biological knowledge that can be utilised for clinical tools or drug discovery remains a significant challenge, as is

modelling the potentially complex interactions between effect SNPs. All of this makes BMI a challenging phenotype to develop precision medical tools for using existing methods [167], but novel techniques may offer new opportunities.

4.2 The UK Biobank

The UK Biobank is a biometric database containing detailed health and genomic data for half a million UK-based adults aged between 40 and 69 years when recruited in 2006-2010. The Biobank holds extensive and detailed genotype and phenotypic information about its participants, with the goal of improving the understanding of the basis of common diseases, which are often complex and multimodal in nature. Data from the UK Biobank is available for all bona fide researchers with no requirement to collaborate, as long as the research in question is for the benefit of the general public [284].

4.2.1 Ethics

This project has been granted approval by the UK Biobank under application ID 52480 and has been conducted in accordance with UK Biobank Ethics and Governance Framework and under the policies of the supplied Material Transfer Agreement for application 52480. The UK Biobank itself has been granted ethical approval from the North West Multi-centre Research Ethics Committee (MREC) as a Research Tissue Bank and is also licensed by the Human Tissue Authority [2, 1]. The genotypes and associated information of withdrawn participants were removed from the dataset before the training and test sets were generated. The University Teaching and Research Ethics Committee (UTREC) for the University of St Andrews approved this project with approval code CS14307 on 28th May 2019 and the research has been conducted in accordance with UTREC policies and guidelines. Data has also been stored in accordance to UTREC and UK Biobank policies and copies of the data will be destroyed upon completion of the PhD.

4.2.2 Data Availability

This project has been conducted using data from the UK Biobank under accession number 52480. This data is available for all *bona fide* researchers and access can be applied for at http://www.ukbiobank.ac.uk/register-apply/.

4.3 Sample Inclusion Criteria

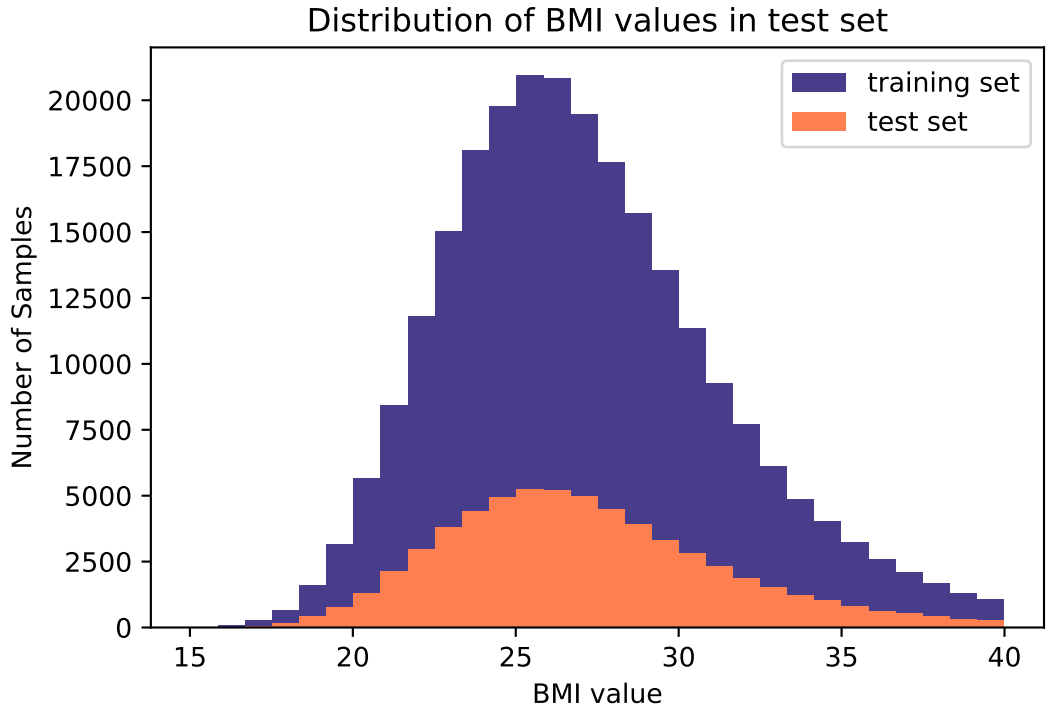


Figure 4.1: Histogram showing the distribution of Body Mass Index scores for individuals in the training set and test sets of my data.

For the purposes of this research, I followed the sample quality control measures of Bellot et al. [17], who themselves replicated the work of Kim et al. [130], and excluded individuals of non-european ancestry and individuals whose reported sex did not match their biological sex from further analysis. Ancestry was determined by self-report of Biobank participants. I also used the UK Biobank's exclusion recommendations to exclude samples with unusually high heterozygosity, samples with unusually high genetic relatedness to other samples (which could also include duplicate samples), and samples exhibiting sex chromosome aneuploidy. I also

excluded individuals who were not used in the Biobank's PCA calculations for any number of reasons, individuals with NaN BMI values, and individuals who had withdrawn their consent for participation in research. Lower sample call rates across the whole genome may indicate a low

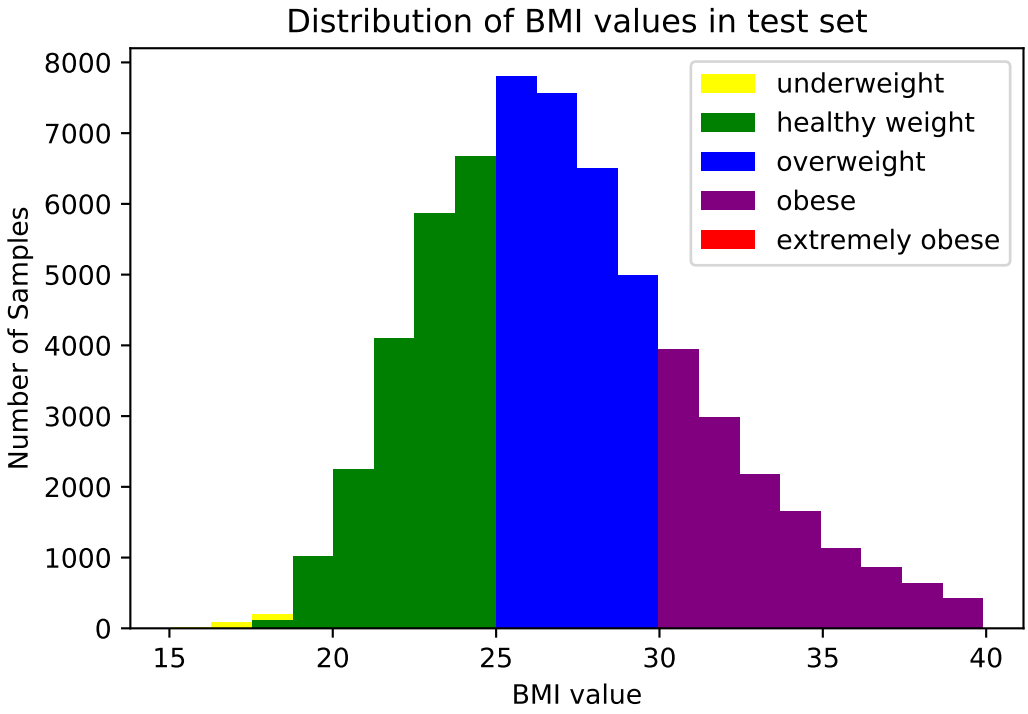


Figure 4.2: Histogram showing the distribution of BMI scores for individuals in the test set of my data, with colours representing different BMI categories.

quality DNA sample that is more likely to contain genotyping errors. For this reason I filtered samples with a call rate of less than 95%, as recommended in [249].

The final sample set thus consisted of 310471 distantly-related individuals of European ancestry (based on self-report by participants in the Biobank), which I further subdivided into 248376 training samples and 62095 test samples. Sample filtering was performed using Hail 0.2 [292]. As you can see in figure 4.1, the BMI distributions of the training and test sets are extremely similar. The distribution is approximately normal with some skew towards the higher BMI end of the spectrum. If we look at the BMI distribution in more detail in figure 4.2, it is apparent that there are very few underweight and no extremely obese samples in my data, so we would expect that it is likely the models will not be able to effectively learn these categories.

4.4 SNP Inclusion Criteria

SNP data in the UK Biobank was obtained using a custom Affymetrix Axiom array which directly measured 850,000 variants. From this, imputation was performed using the Haplotype Reference Consortium and UK10K + 1000 Genomes research panel, yielding in excess of 90 million SNPs [31]. For this investigation I decided to use only directly genotyped SNPs as input, as input size was already an issue and imputed SNPs are by nature highly correlated in a way that may potentially introduce issues when training a machine learning model. Variants that were not SNPs (e.g. copy number variants) were removed. As in Bellot [17] and Kim [130] I removed rare variants by filtering alleles with Minor Allele Frequency (MAF) <0.1% and missingness >3%. After filtering, 735221 SNPs remained for analysis. The statistical calculations and filtering for SNP quality control were performed with Hail 0.2 [292].

In an infinitely large population without selection pressures or migration, it can be assumed that the distribution of genotypes and alleles is constant over generations. This is known as the Hardy-Weinberg Equilibrium (HWE) law [184]. Generally, deviations from this equilibrium are considered to be indicative of poor genotyping quality and removed, especially from the control set. However, since deviation from this equilibrium can also signal genuine significance, pruning HWE deviating variants can potentially result in the removal of informative disease-causing variants [9]. Additionally, quantitative phenotypes such as BMI do not have a control set, which further increases the risk of removing important variants. For these reasons we chose to follow the example set by Kim *et al.* and Bellot *et al.* and not risk the removal of influential SNPs via HWE pruning [130, 17].

Additionally, after GWAS but before later analysis such as Polygenic Risk Score calculation, SNPs in Linkage Disequilibrium are often removed from the dataset. The rationale behind this is that SNPs in linkage disequilibrium will be highly correlated with one another and thus may register as significant during GWAS when they are not actually causally related to the phenotype [9]. However, removing SNPs in LD is risky as the LD architecture of the genome is complex and irregular, and also different for different populations, meaning it can be extremely difficult to determine which SNPs are the causal SNPs. Since pruning may result in removal of the causal

SNP by accident, more sophisticated tools such as LDPred automatically assign weights to SNPs in LD to try to reduce the effects of non-causal SNPs [304] (see chapter 2 for a more detailed explanation of the motivations for the editing of SNPs based on LD). For a linear additive model, such as a PRS or a linear regression model, it is also important that the inputs can be reasonably assumed to be uncorrelated with one another, so this down-weighting step is likely to improve performance. However DNNs do not assume the inputs are uncorrelated, and should be able to automatically learn to down-weight unimportant SNPs that only made it into the feature set due to correlation during the training phase. Additionally, prior research shows that attempting to account for LD a priori when working with DNNs can lead to reduced performance [17], so I have chosen not to perform any LD pruning or clumping as part of my pre-processing.

4.5 P-Value Filtered SNP Sets

To create input datasets for a deep learning model it was necessary to further decrease the number of input features in order to reduce computational complexity. I chose to follow the examples of Bellot et al. [17], Mieth et al. [190], and Hassan et al. [217] and perform a GWAS in order to rank SNPs and thus allow me to select subsets featuring the most relevant N SNPs, where N was set as 100, 500, 1000, 5000, 10000 and 50000. This approach has limitations, as discussed in the cited studies, the main one being that there is a possibility that SNPs that are important as part of interactions may not be individually significant and thus will be excluded, to the potential detriment of the network performance. Ultimately I decided that this was still the best approach as it was necessary to do some kind of dimensionality reduction and the top-SNP method appears to perform best in the Bellot study [17].

4.5.1 GWAS results

As shown in the Manhattan plot in figure 4.3, the GWAS identified numerous significant peaks spanning the entire genome, but the most significant loci were located in chromosome 16. This is as to be expected in accordance with the literature [165] which identifies BMI as being highly polygenic but with the most significant association being found with the FTO gene on chromosome 16 [82].

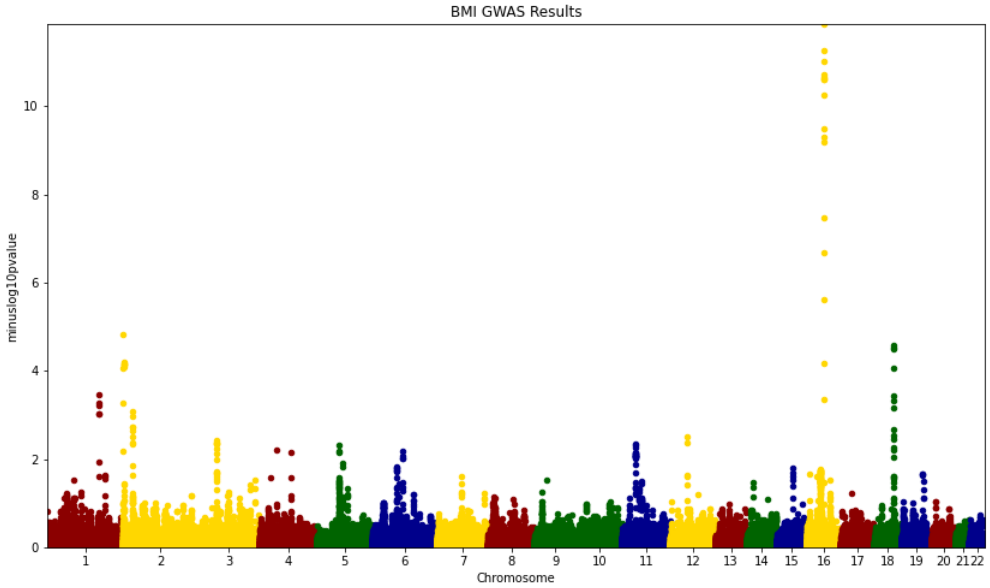


Figure 4.3: Manhattan plot showing the most significant SNPs discovered by our BMI GWAS. The y-axis shows the negative base ten logarithm of the p-value for SNPs in each chromosome.

Chapter 5

Deep Neural Networks for BMI Prediction

The purpose of this chapter is to outline the techniques used to analyse the performance of different linear and deep learning models in predicting BMI, as well as the results of these techniques. The goal is to address the research question of whether deep learning models can outperform linear models at predicting complex traits from genomic data, and to further investigate what factors in network and feature design may further enhance prediction performance.

This chapter is divided into three main sections. The first section focusses on the question of how to most effectively encode genetic data for deep learning models, summarising a variety of existing encoding techniques as well as several novel encoding methods designed for this project. The second section focusses on the methodology of this phase of the research, covering topics such as how neural networks were designed, how linear baseline models were selected, and how the performances of both types of model were evaluated and compared. The final section discusses the results of the performed analysis and draws conclusions about which models performed best and which model design features did and did not have an effect on performance.

5.1 Encoding of Genetic Data

The question of how to encode SNP data from genotype arrays in a way that is comprehensible and meaningful for machine learning models is a significant problem that does not receive as much attention as it should. All methods currently in widespread use rely on human genome

assemblies compiled from multiple sources to be used as a representative example of the human genome, known as reference genomes. These reference genomes (of which there are multiple versions, known as builds) do not correspond to the genes of any single living human, instead being a haploid mosaic curated from multiple individuals [51]. At points where variance occurs (SNPs), alleles can be encoded as either reference (REF), meaning corresponding to the allele of the reference genome, or alternate (ALT), corresponding to any other allele. It should be emphasised that there is nothing biologically meaningful about the "reference" label; this indicates only that the allele in question happened to be the one present at this site in the segment of genome used, not that it is more common, part of an older genetic lineage, or in any other way more "default" than any of the alternate alleles. Thus every encoding method that builds upon this protocol has at its heart a slightly arbitrary delineation of states.

5.1.1 Standard Variable Encoding

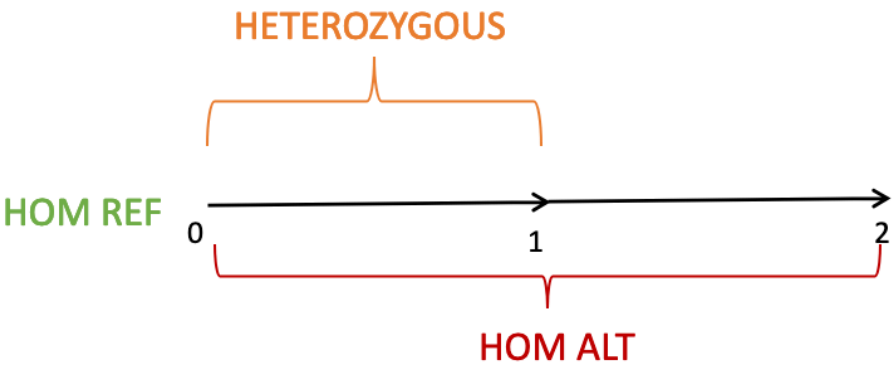


Figure 5.1: A graphical representation of the vector space of the Standard Variable Encoding.

The so-called Standard Variable Encoding (SVE) method uses the number of alternate alleles as a single digit encoding for the genotype at a given locus, with 0 corresponding to homozygosity for the reference allele, 2 corresponding to homozygosity of the alternate allele, and 1 corresponding to heterozygosity [104]. This creates a one dimensional vector space in which all genotypes are projected along the same axis and thus assumes a linear additive relationship between alleles, as illustrated by figure 5.1. Since the magnitude of the genotype vectors is determined by the number of ALT alleles, the homozygous reference vector has a magnitude of zero. Whilst this encoding method is efficient and computationally inexpensive,

and thus useful for reducing the complexity of models, it was derived for linear applications such as Polygenic Risk Scores and it contains an implicit assumption of linear additivity; as far as a machine learning model is concerned, the real effect size of two alternate alleles should be double the effect size of one, which biases against the potential of predicting epistasis or dominance effects. There is also the implication that 2 is of higher value than 0, despite the fact that the alternate allele may not be the effect allele.

5.1.2 One-Hot Variable Encoding

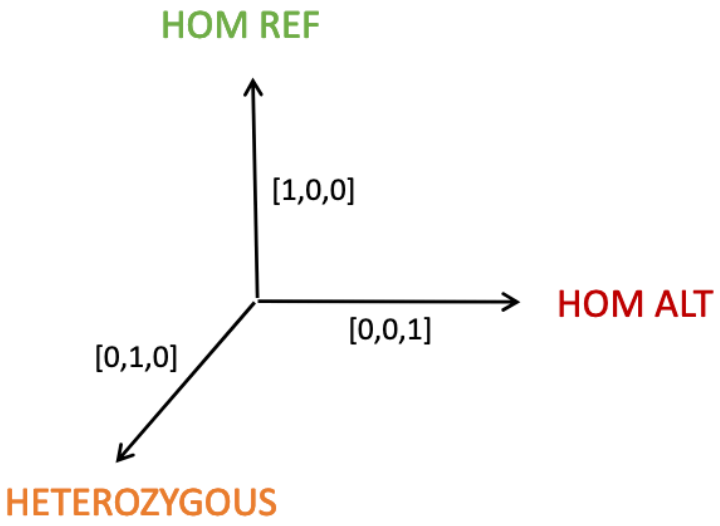


Figure 5.2: Graphical representaion of the three dimensional vector space of a one-hot representation of REF/ALT genotypes.

A way around the problem of assumed allele additivity is to treat the different genotype options as categorical variables wholly unrelated to one another. A common approach for encoding categoric variables in machine learning is one-hot encoding, in which a series of N zeroes and ones is used to label a feature in such a way that no ordering or mathematical relationship is implied, where N is the number of categories. The variable x_i is encoded with a vector of length N where every element is zero aside from the i_{th}, which is one [103]. For REF-ALT gentoypes, this creates a three dimensional vector space in which each genotype is projected along its own axis, which means that there is no assumption of linear additivity, and it may be easier for models

to learn a nonlinear relationship between phenotypes. This relationship is illustrated graphically in figure 5.2. The benefit of this approach is that it would allow for dominance and epistatic effect modelling, since no relationship between the heterozygous case and either of the homozygous cases is presupposed. It also implies no ranking of the alternate alleles over reference alleles.

However, there are a number of issues with one-hot encoding. In the first case, it increases the dimensionality of the input by a factor of three over the standard variable encoding, which can have big effects on computational complexity when it comes to deep learning models. It also loses some of the implicit priors of the standard variable encoding that were actually useful, namely the ordering of the three genotypes: despite the fact that the alternate allele is not necessarily the effect allele and hence should not be implicitly more important than the reference (as is the case in the SVE), the heterozygous case will always be somewhere between the two homozygous cases in terms of effect. In a one-hot representation, the projection of each genotype vector along any of the other genotype dimensions is zero, but in reality the heterozygous case does contain an element (a single allele) of each of the homozygous cases, and the three cases are not actually independent. This useful information is destroyed by the one-hot encoding approach.

5.1.3 Reference-Alternate Projection Encoding

I have developed an alternative encoding method that aims to capture the useful information from SVE without the less useful implications of additivity and value ordering. My method is inspired by the word vector projection methods employed in natural language processing, where words are represented as real-valued vectors in a multidimensional vector space [52].

As described, the standard variable encoding generates a one-dimensional vector space, with homozygosity for the reference allele (HOM-REF), homozygosity for the alternate allele (HOM-ALT), and heterozygosity (HET) cases considered vectors of differing magnitudes but the same direction (hence the inherent assumption of linear additivity); and in the one-hot encoding, HOM-REF, HOM-ALT and HET all have equal magnitudes and different directions. I wanted to preserve the assumption of equal magnitudes and differing directions for the two homozygous cases, while acknowledging that the heterozygous case shares some effect with each of the

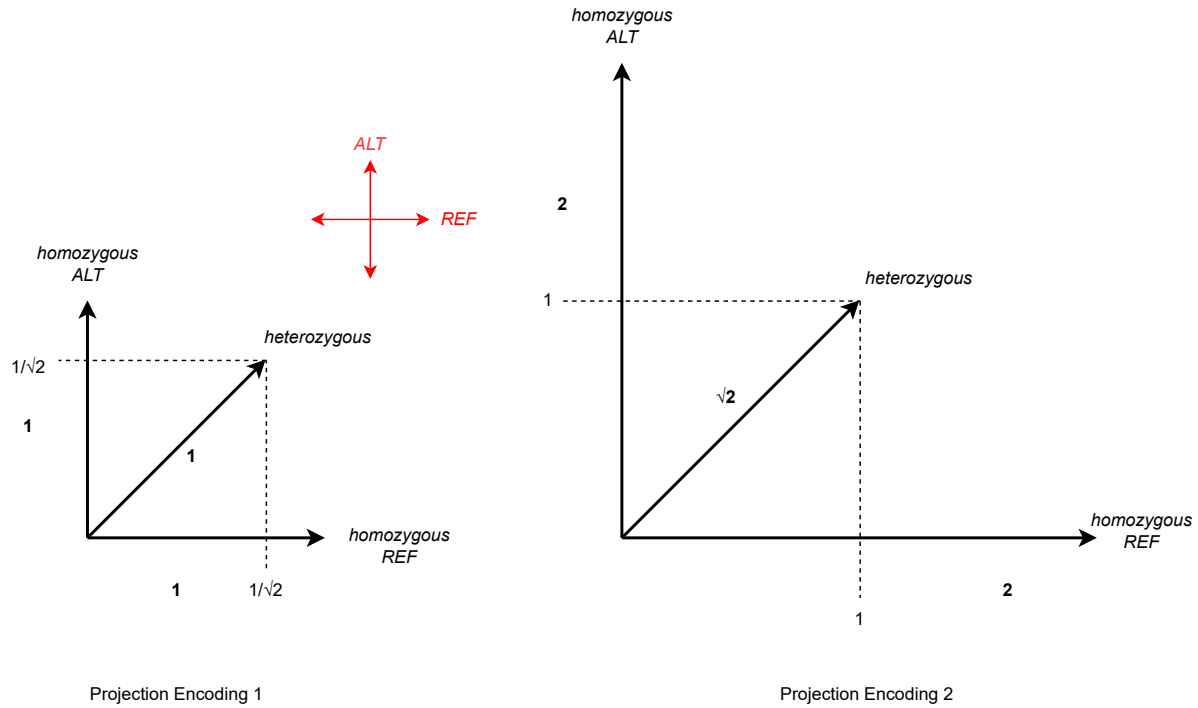


Figure 5.3: Graphical representation of the two dimensional vector space for REF/ALT projection encoding, in which the heterozygous genotype projects equally onto each axis.

homozygous cases and therefore should not be orthogonal to them. To achieve this, I propose a two-dimensional vector space where the two axes are REF and ALT, and the HOM-REF and HOM-ALT cases are vectors projecting along each axis respectively. The basic HET case assumes equal projection along both HOM and REF axes, reflecting the fact that the individual has one of each allele. The question of what magnitude the HET case vector should have was slightly unclear, and I tried two options. If we constrain all three vectors to have unit magnitude so that no genotype is of implicitly more importance, then the heterozygous case has a projection of $1/\sqrt{2}$ along both axis. Alternatively if we use the number of each allele literally for creating the encoding and don't worry about the resulting magnitude, we end up with a situation in which each homozygous case has a projection of 2 along its own axis, and the heterozygous case projects 1 along each axis. Both encodings are illustrated in figure 5.3).

5.1.4 Effect Direction Projection Encoding

A major shortcoming of all three methods listed above is their reliance on REF/ALT as the basic method of distinguishing between alleles, which as we have discussed is an arbitrary delineation

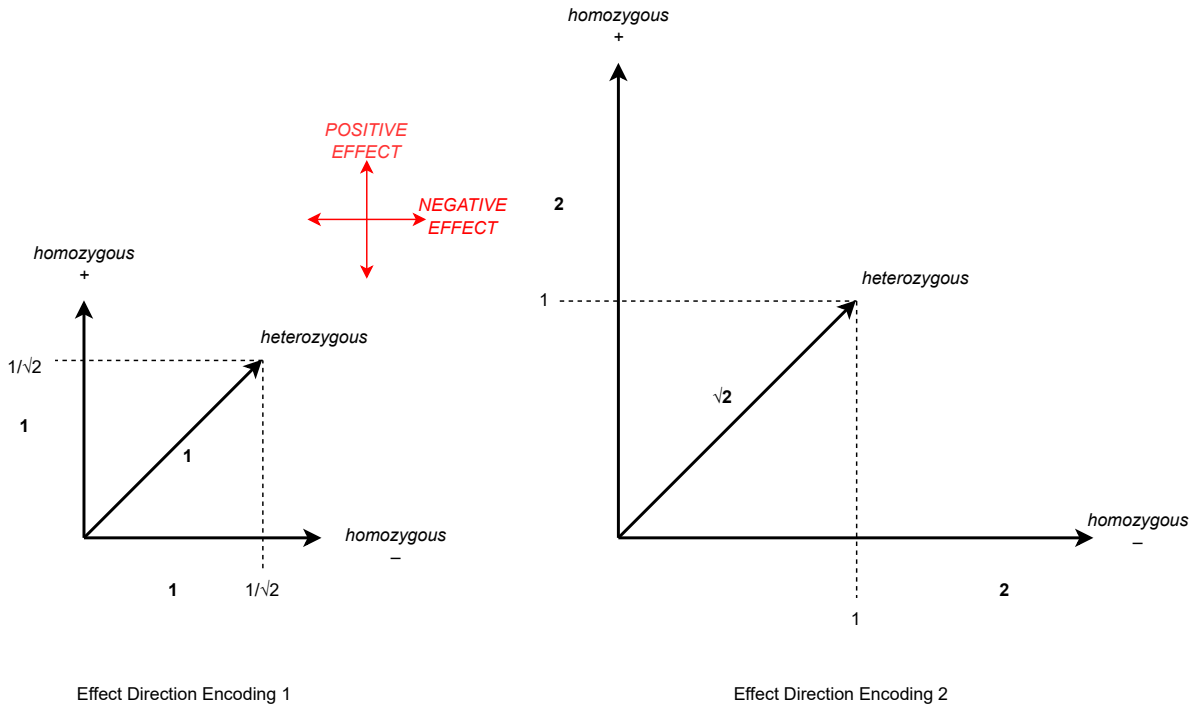


Figure 5.4: Graphical representation of the two dimensional vector space for effect direction projection encoding, in which once again the heterozygous genotype projects equally onto each axis, but the axes now correspond to the direction of influence of the allele.

that does not convey any biologically useful information. To overcome this, I have designed a further encoding method that aims to make use of more biologically informative data. In a GWAS the effect direction of all variants at a given locus is derived, which indicates whether each allele is associated with an increase or a decrease in a quantitative phenotype, information which could potentially be useful to the network (as, for example, a collection of alleles largely associated with phenotype increase would be more likely to be associated with a net phenotype increase, although the possibility of nonlinear interactions means this is not a given). The effect direction encoding is similar to the REF/ALT projection encoding in that it utilises projections along two axes, with the two homozygous cases being unit vectors along each axis and the heterozygous case having a projection on both, but in this case the two dimensions correspond to positive and negative direction of phenotype association, instead of REF and ALT alleles. This is illustrated in figure 5.4.

Encoding #	Encoding Name	Number of Dimensions	Homozygous Representation	Heterozygous Representation
1	Standard Variable Encoding (SVE)	1	[0], [2]	[1]
2	One-Hot Encoding	3	[1, 0, 0], [0, 0, 1]	[0, 1, 0]
3	Projection Encoding 1	2	[1, 0], [0, 1]	[0.7, 0.7]
4	Projection Encoding 2	2	[2, 0], [0, 2]	[1, 1]
5	Effect Direction Encoding 1	2	[1, 0], [0, 1]	[0.7, 0.7]
6	Effect Direction Encoding 2	2	[2, 0], [0, 2]	[1, 1]

Table 5.1: Table showing information about the various encodings trialled, labelled encodings 1-6 in the results section. Note that the positive effect direction allele is not necessarily the same as the reference allele, so for example the genotype that is encoded as [2, 0] in encoding 4 may be [0, 2] in encoding 6.

5.2 Methodology

5.2.1 Data Preprocessing

As can be seen in figure 4.1, the phenotype data closely follows a normal distribution. I performed Z-score standardisation on the phenotype data so that the mean BMI was centered at zero and the standard deviation was 1 as I found this marginally improved prediction performance. Since the data closely adheres to a normal distribution without many significant and obvious outliers, I did not remove outliers as part of data preprocessing. The deep learning model hypothetically has the capacity to learn to predict BMI values all the way along the spectrum of the data, and I did not want to remove potentially informative samples. However, during the interpretation phase I did constrain the analysis to samples that were more towards the middle of the data distribution, as the model did not learn to predict outlier values well.

5.2.2 Software

Pre-processing and quality control of genetic data, as well as GWAS, were performed with Hail 0.2 [292]. Linear baseline models were fitted with scikit-learn using the SGDRegressor functionality. The sklearn GridSearchCV function was used to explore a range of combinations of different parameters for these models [231]. Feedforward Deep Learning models were constructed and trained with gpu-enabled Pytorch [229]. Grid search for hyperparameter tuning was performed with Ray Tune [160]. Bayesian neural networks were also implemented with Pytorch as well as the Bayesian Neural Network library BLitZ [79].

5.2.3 Evaluation Metrics

For evaluation metrics, I followed the example set by Bellot *et al.* and used the Pearson R correlation between predictions and ground truth values of the test set to evaluate model performance [17]. As well as this I used the average Mean Squared Error (MSE), and plotted the prediction vs ground truth values to look for signs of visual correlation. I also trialled using the coefficient of determination, R^2; however, due to reasons discussed in subsection 5.2.3.3, these values became difficult to interpret and were not overly useful for evaluation. For the Bayesian Neural Networks the loss can be very high due to multiple sampling and stochasticity, so I used the Confidence Interval Accuracy as a measure of performance, as well as Pearson R.

5.2.3.1 Mean Squared Error

As discussed in the loss functions section, Mean Squared Error is the square of the difference between the prediction and ground truth value, and is commonly used as a loss function. It also can be used as a metric of model quality and a way of comparing the success of different models trained for the same task [308]. A limitation of MSE as a model benchmarking tool is that since it is not constrained to a set range, it can be hard to determine the MSE threshold of a "good" predictor.

5.2.3.2 Pearson R

Pearson product-moment correlation, often denoted as R, is a measure of association between continuous variables that is most recommended for data where both variables being measured follow a normal distribution, and in which there are not significant outliers. It ranges from -1 to 1, with 0 corresponding to no correlation and -1 and 1 indicating perfect inverse and linear correlation, respectively [268]. It is a good choice for this application as we know that the phenotype is normally distributed and that models trained on it are likely to produce normally distributed predictions, there are few outliers, and the sample size is large enough that it counterbalances the effect of outliers [11]. Unlike MSE it has a pre-defined range, meaning that it is easier to interpret the success of a predictor.

5.2.3.3 R^2

The coefficient of determination, usually written as R^2, is a measure of the proportion of variation in outcomes that a model is able to predict from the independent variables [300]. A perfect predictor would be able to capture the full relationship between input and output and therefore predict all variation in the data from the independent variables, leading to an R^2 score of 1. A model that predicts none of the variation would predict the mean output value regardless of input and would theoretically have an R^2 score of 0 [28]. In some controlled circumstances, R^2 is equal to the square of the Pearson R coefficient.

The R^2 value calculation depends on the assumption that the total sum of squares is equal to the explained sum of squares plus the residual sum of squares, i.e.

$$SS_{total} = SS_{expl} + SS_{res} \tag{5.1}$$

where SS_{total} represents the total variation in the data by the sum of squared differences between expected and actual values, SS_{expl} represents the subset of these squared differences that the model has successfully explained, and SS_{res} represents the rest of the squared differences, i.e. those the model is unable to explain. R^2 is thus loosely defined as

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}} \tag{5.2}$$

While the theoretical lower limit of the coefficient of determination is 0, there are a number of situations in which the value can be negative, which can make results difficult to interpret. One of the ways this can arise is when equation 4.4 does not appear to hold true because SS_{res} and SS_{total} have not been calculated using the same dataset. In machine learning applications it is standard practice to evaluate model performance on a separate test dataset, which should be similar but will not be identical to the data that the model was trained on. When R^2 is calculated using only data from the training set, it gives comprehensible values between 0 and 1; however, since the model is predicting on data it has seen before, in almost all cases the performance will

be better than on unseen data, leading to an artificially inflated R^2 value and overestimation of SS_{expl}. However, when there is an attempt at quantifying R^2 on the unseen test set, it can result in uninformative negative values that arise from differences in the training and test set [320]. For this reason I chose not to use this metric in evaluating my models, and used Pearson R and MSE instead.

5.2.4 Evaluation Protocols

I subdivided my data samples into training and test sets by randomly selecting batches from throughout the dataset. There were 60 batches in the training set, totalling 245760 samples, and 16 batches in the test set, totalling 65536 samples, or about 20% of the total samples available after data preprocessing. During training, the training data was automatically subdivided into training and validation sets, where the validation data was not passed to the regressor during the training phase when weights were being adjusted or used in backpropagation, but was instead used as a benchmark for prediction performance while the model was training. The validation sets comprised 20% of the training samples. During the initial search phase the training and validation set partition was only performed once for efficiency, since a large number of models were being trained at once. Then, once the initial best model for the given SNP set had been identified, I re-trained the successful model with five-fold cross-validation. This consisted of the randomly partitioning the training set into five groups, and using each group in turn as a validation set while the other four groups would be used for training. By checking that the model performance was similar across all training-validation partitions, as well as being similar to the model trained with the original randomly chosen partition during the search phase, I could eliminate the possibility that a performance improvement was due to a randomly fortuitous choice of partition.

The test data, on the other hand, was held out until the final best models were chosen to eliminate any possibility of data leakage across the whole model search phase. This dataset was used to provide a benchmark of the prediction performance on truly unseen data. As you can see in figure 4.1, the distributions of the training and test sets are nearly identical, meaning that the network should theoretically be able to generalise to the test data.

Since it is not definitively known how much of the genome is involved in regulating BMI expression, I trained all my models on a variety of datasets containing different numbers of BMI-associated SNPs. The smallest dataset contained only the top 100 SNPs based on significance scores from the GWAS, with the following datasets containing the top 500, 1000, 5000, 10000 and 50000 GWAS hits, respectively. Unfortunately due to model complexity it was necessary to perform some kind of input feature selection and so we could not train a network on the full 735221 SNPs that remained after quality control, and even the 50000-SNP set ran into some computational problems for some of the models. If the phenotype is only governed by a few genes, and the mechanism by which the genotype is related to the phenotype is of the linear sort that is easy to detect with GWAS, then the 100 most significant SNPs should contain most of the information needed for the network to make predictions, and the rest of the data should essentially be noise. If including more (less significant) SNPs improves performance then this suggests that the phenotype is highly polygenic and the most significant SNPs identified by GWAS provide an incomplete picture of the genetic basis of the disease.

5.2.5 Baseline Selection

A challenge with using linear models on data this size is memory; many linear models are limited to training on data that can be loaded into memory, which in my case is a small fraction ($\sim 10\%$) of the total training data available. So I chose to use a Stochastic Gradient Descent (SGD) Regressor as a linear baseline models as it is capable of iterative partial fits and thus I can use the entire training data set and have a fairer performance comparison to the deep learning methods.

5.2.5.1 Stochastic Gradient Descent (SGD) Regressor

A Stochastic Gradient Descent (SGD) regressor is a linear regression model that fits to data by using the stochastic gradient descent algorithm to minimise a regularised empirical loss function [253]. SGD is a good method for efficiently training models on large datasets as it works incrementally over the data, updating the gradient as it goes, and it has been successfully applied to machine learning problems with high-dimensional, sparse datasets in fields such as natural language processing. Stochastic gradient descent is merely a type of optimisation

N SNPs	Encoding	Penalty	Loss Function	R	MSE
100	4	L1	huber	0.089474	17.37
500	1	L1	huber	0.123288739	17.78
1000	3	L2	huber	0.147232211	17.59
5000	2	L1	huber	0.17301131	21.53
10000	1	L1	huber	0.217992522	18.44
50000					

Table 5.2: Parameters and performance metric results of best performing linear baseline models for each SNP set.

rather than a type of model in its own right, so depending on the choice of regularisation penalty and loss function an SGD Regressor will have an analogous linear model that solves the same problem in a slightly different way. For example, an SGD Regressor with squared loss and L2 regularisation is analagous to Ridge Regression, and the same loss with L1 regularisation would be equivalent to Lasso Regression.

There is not a clear consensus in previous work about which models should be used as linear baselines, with choices varying greatly between studies. However, linear regression and lasso are common choices in ML, and are also used as a baseline in [62], so I included these. As the SGD Regressor can act as many different linear classifiers, and it was not clear which type of linear model would be best for the data, I performed a gridsearch across available parameters, the results of which are shown in table 5.2. On the whole the SGD regressor was quite unstable and prone to very high loss values, meaning that the best performing model selected was usually just the one option that did not lead to exploding gradients. The huber loss function improved upon the standard ordinary least squares fit by penalising outliers less heavily and preventing the loss from exploding. The loss function used in Support Vector Regressors, epsilon-insensitive, was also trialled and also led to exploding loss, as did squared epsilon sensitive. The penalties trialled were L1, L2 and elasticnet, with L1 generally performing best across the majority of SNP sets and encodings, though it seems that penalties were not a deciding factor for linear model performance. There was not a clear trend among the different encoding options for the linear models, with all encodings performing comparably, and an even spread of encoding options across the best performing regressors. There was a clear improvement in Pearson R as the SNP sets got larger, suggesting that having more information available did improve performance even

when the models were only able to model linear additive interactions between SNPs. However, the 50000 SNP set was too large for the linear models to train on and could not be benchmarked. The mean squared loss stayed fixed around the standard deviation for all the models even as the R and R^2 metric, suggesting that the fit of the models was limited even as the prediction performance on the hold-out set improved.

5.2.6 Design of Neural Networks and Hyperparameter Optimisation

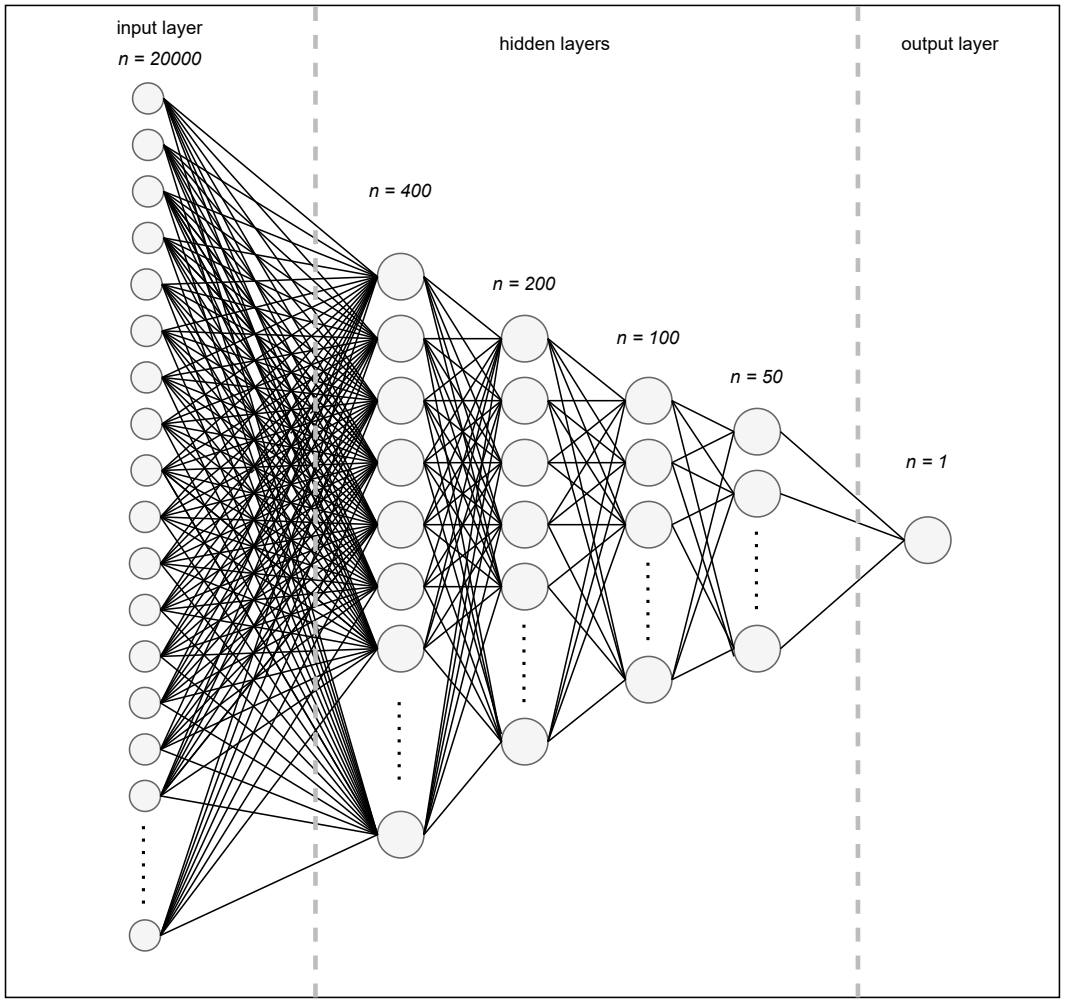


Figure 5.5: Representation of the architecture of the best performing model.

Choosing a neural network architecture for a problem where the relationship of input and output is largely unknown and the problem space is not well explored is challenging. Deeper and wider networks tend to lead to improved prediction performance, but also rapidly increase the size of the hyperparameter search space with respect to network architecture, making training

different models sequentially prohibitively time consuming. Since finding the best combination of hyperparameters is so difficult, I performed grid searches using Ray Tune across a variety of options to try to determine the best combination. The options available for each hyperparameter can be seen in table 5.3, while the network architectures tested (written in terms of reduction factors) can be seen in table 5.4. Since different encodings resulted in different input sizes, architectures were defined in terms of reduction factors, representing to the factor by which the input was reduced for each layer; this allows the architectures to be directly compared even for different encodings. As the number of SNPs in the input set increased the reduction factors had to increase too to prevent GPU memory errors due to too many parameters. Thus a separate grid search was performed for each encoding, for each SNP set, for each loss function, and the best hyperparameters for each of these categories is listed in table 5.5.

To further increase the efficiency of the hyperparameter search, I used the Asynchronous Successive Halving Algorithm (ASHA) to schedule trials with adaptive resource allocation and early stopping. ASHA builds upon the success of the Hyperband parallelisation algorithm, which optimises random search for neural network hyperparameters by dividing trials into brackets and periodically killing the lowest performing trials in each bracket [154]. ASHA scheduling follows a similar approach and achieves similar results to Hyperband but also avoids some issues with straggler trials [155].

The best hyperparameter combinations were then tested again using 5-fold cross validation, to ensure that results were robust across training/validation data partitions and to mitigate the possibility of overfitting. Finally, the best overall network for each SNP set was validated on the test data, and these results are seen in table 5.5.

Activation Functions	ELU	ReLU	LeakyReLU	
Dropout	0	0.1	0.2	0.3
Optimisers	radam	adam	adamw	adamax
Loss Functions	huber	MSE		

Table 5.3: Different options used in grid search for each type of hyperparameter excluding architectures.

Overall, for many hyperparameters the various options did not seem to make much difference to

network performance. For example, none of the tested activation functions greatly outperformed any of the others, though ELU was most common activation in the best performing networks. Networks trained using the Huber loss function generally performed better than those trained with Mean Squared Error, but the difference was small. It was unclear whether nonzero dropout improved performance noticeably, but training and validation loss trends showed the networks were much more prone to underfitting than overfitting, so it makes sense that many of the best performing networks had zero dropout as there wasn't a clear indication that dropout was required. There was again no clear distinction between any of the optimisers, though radam was the most common in the best performing networks. It also wasn't clear whether any particular architecture offered much advantage over the others, with the results being fairly similar between them, though across the three smallest SNP sets the [2,2,5,5] reduction architecture did perform the best.

Training and evaluating the Bayesian neural networks required significantly more resources in comparison to the feedforward ones, as the evaluation metrics for Bayesian neural networks require multiple sampling to compensate for the nondeterministic outcome of the network, so doing a gridsearch was not feasible. Hence I adapted the best performing network architecture and parameters for each SNP set into a Bayesian neural network and observed whether this resulted in an improved prediction performance.

Number of SNPs	**Architecture Options (Reduction Factors)**
100	[2,2,2], [1,10], [2,1,2,1], [2,2,5,5]
500	[2,2,2], [1,10], [2,2,5,5], [5,2,5,2]
1000	[10,2,2,2], [10,1,10], [2,2,5,5], [5,2,5,2]
5000	[10,2,2,2], [10,1,10], [10,2,2,5,5], [10,5,2,5,2]
10000	[50,2,2,2], [50,10,10], [50,2,2,5,5], [50,5,2,5,2]
50000	[100,2,2,2], [100,10,10], [100,2,2,5,5], [100,5,2,5,2]

Table 5.4: Different architecture options for each SNP set, where reduction factors correspond to the input/output ratio for each hidden layer. Larger SNP sets have larger initial reduction factors to avoid networks exceeding the GPU memory.

N SNPs	Architecture	Loss	Enc.	Activation	Dropout	Optimiser	LR	R	MSE
100	[2,2,5,5]	MSE	4	ELU	0.1	radam	1e-3	0.09	17.21
500	[2,2,5,5]	Huber	3	ReLU	0	adamax	1e-3	0.13	17.20
1000	[2,2,5,5]	Huber	6	ELU	0.2	radam	1e-4	0.15	17.10
5000	[10,10,10]	MSE	4	ELU	0.3	radam	1e-3	0.22	16.96
10000	[50,2,2,2]	Huber	4	ELU	0.2	radam	1e-4	0.25	17.16
50000	[100,2,2,2]	Huber	4	ELU	0.2	radam	1e-4	0.27	16.96

Table 5.5: Hyperparameters and prediction performance metrics of best performing neural networks (as discovered through grid search and verified through 5-fold cross validation) for each SNP set.

5.3 Results and Discussion

5.3.1 Deep Learning vs Linear Models

For all combinations of parameters, using all SNP sets, the deep learning models provided the best prediction performance, though for the smaller SNP sets in particular the differences were quite small. A comparison between the linear and deep learning model Pearson R scores for each SNP set can be seen in figure 5.6, and a comparison of the best linear and deep learning model performances on the test set can be seen in figure 5.7. The best Pearson R scores for the unseen test set are slightly lower than they are on the validation sets, which is typical. As the number of SNPs increases, the differences in performance between the linear baselines and the deep neural network become more pronounced, particularly in the 5000 and 10000 SNP sets, suggesting that including more SNPs may increase the occurrence of complex nonlinear interactions that could be picked up by the deep learning models and not the linear ones. The Bayesian neural networks also begin to fail to converge in these larger SNP sets. Thus we can conclude that the feedforward neural network outperforms the linear baseline models across all SNP sets and offers the best opportunity for high predictive performance in general. The MSE measures of all the models stay fairly fixed even as the Pearson R scores improve, suggesting that the fit of even the best models is limited. For wider context, a 2.1 million variant PRS metric achieved a Pearson R score of 0.292 [128], which is a slightly better prediction performance than any model tested in this study; however, we can see that the 10000 and 50000 SNP deep neural networks came quite close to this using only a small fraction of the variants. Unfortunately it was not possible to test any models with $\sim$1 million inputs due to computational constraints. The comparable performance of our models with linear additive models of BMI could suggest that

nonlinear interactions between variants do not play a huge role in determining an individual's role. It could also be possible that the network does not have enough data to fully model the interactions, either because crucial SNPs were excluded during filtering, or because interactive effects involve rarer variants that were filtered out or not included in the panel in the first place.

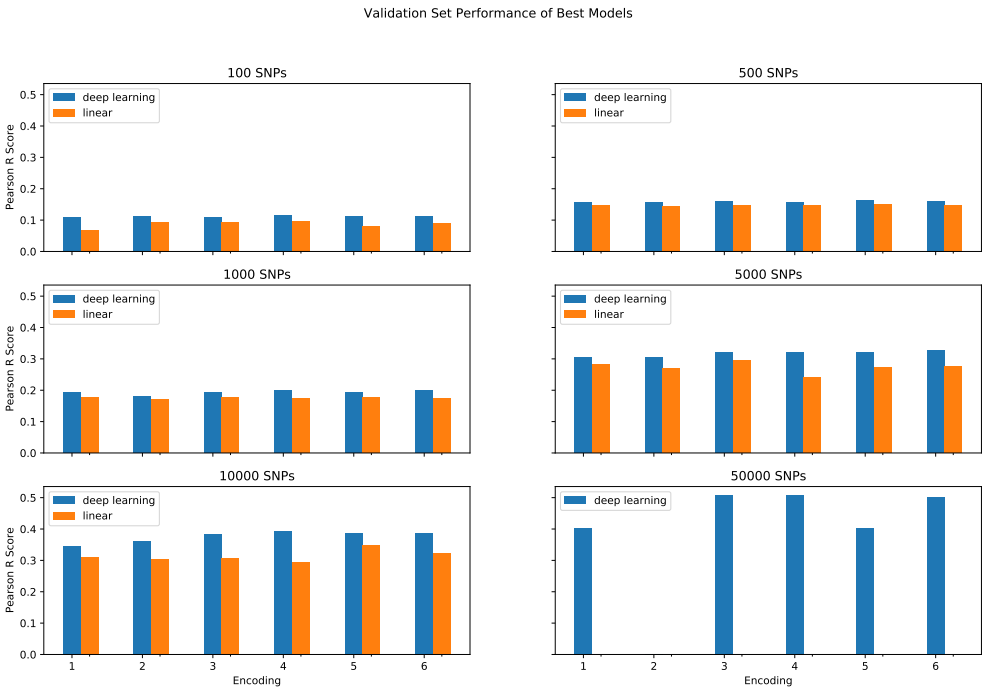


Figure 5.6: Graph comparing the validation Pearson R score for the best configuration/architecture of the best regressors (linear baseline and neural network), for each encoding.

If we look more closely at the model performances in figures 5.8 and 5.9, we can see that the fit of the models is indeed limited, with both models making far too many predictions around the mean without capturing the true distribution of the data. From these graphs the poor fits of the linear and deep learning models appear similar, but if we look more closely at the prediction distribution of both as compared with the ground truth distribution in figure 5.10, we can see that the deep learning model does have more variance in its predictions and comes slightly closer to approximating the underlying distribution, where the linear model falls far short, making far too many predictions around the mean. Both models struggle with samples around the ends of the distribution, though the deep learning model did make more overweight predictions than the linear model. This performance variation on the ends of the phenotype distribution is further

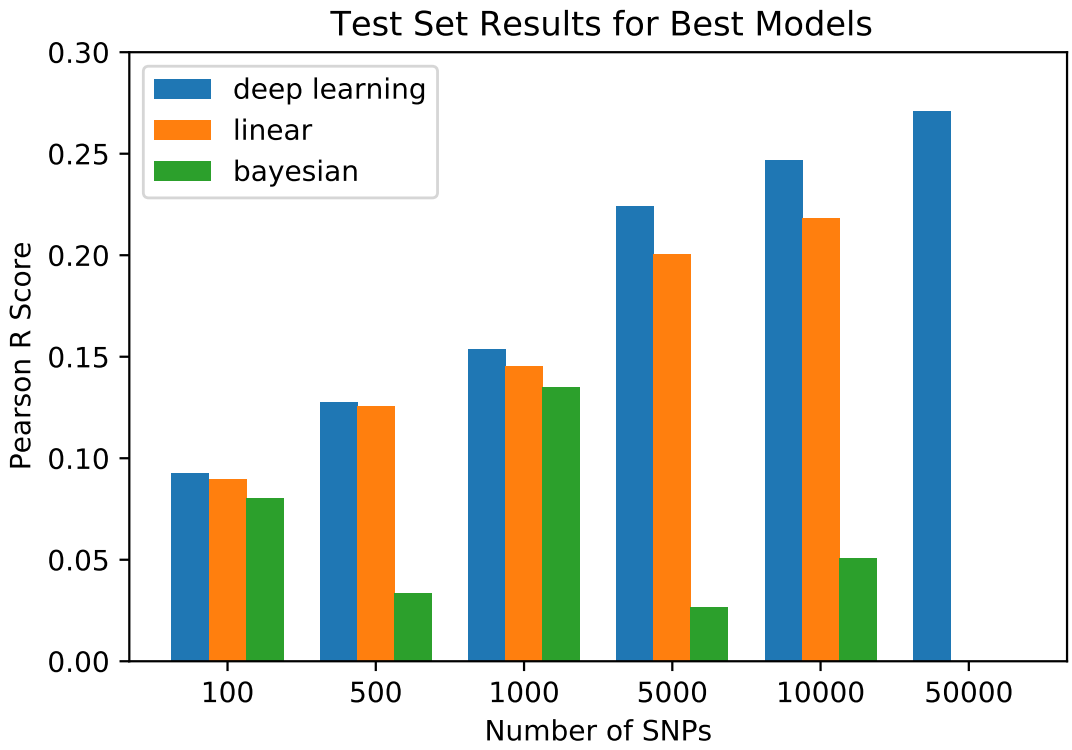


Figure 5.7: Graph comparing the test set Pearson R score for the best configuration/architecture of the overall best regressors (linear baseline and neural network), for each SNP set.

illustrated by figure 5.11. The MSE distribution of the best linear model is shown in figure 5.12; it is apparent that the BMI cutoff for poor performance is lower for this model. If we look at figure 5.13 it is apparent that the MSEs did not improve very noticeably as the SNP sets increased in size, though the deep learning models tended to have slightly lower MSEs than the linear models, and the linear models' MSE scores got worse for the larger models. The MSE hovered at roughly the square of the standard deviation of the dataset, suggesting again that neither linear nor deep learning models were able to get a very good fit of the data.

5.3.2 Effect of Number of Input SNPs

From figure 5.6 it is evident that there is a clear trend across multiple model types that shows that including more SNPs of lower statistical significance improves prediction performance. The exception of this is for the Bayesian neural networks, which perform very poorly on the larger SNP sets and peak on the 500 SNP set. However, for the feedforward neural networks and the linear baseline models, we can see that the Pearson R score rapidly increases with increasing number

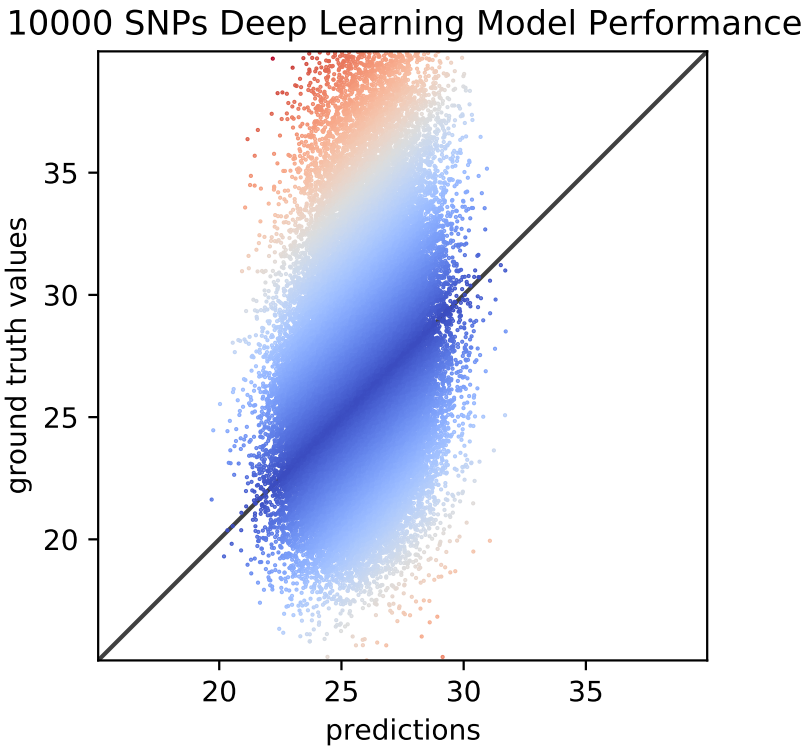


Figure 5.8: Prediction vs ground truth graph for the best 10K SNP deep learning model on the test data.

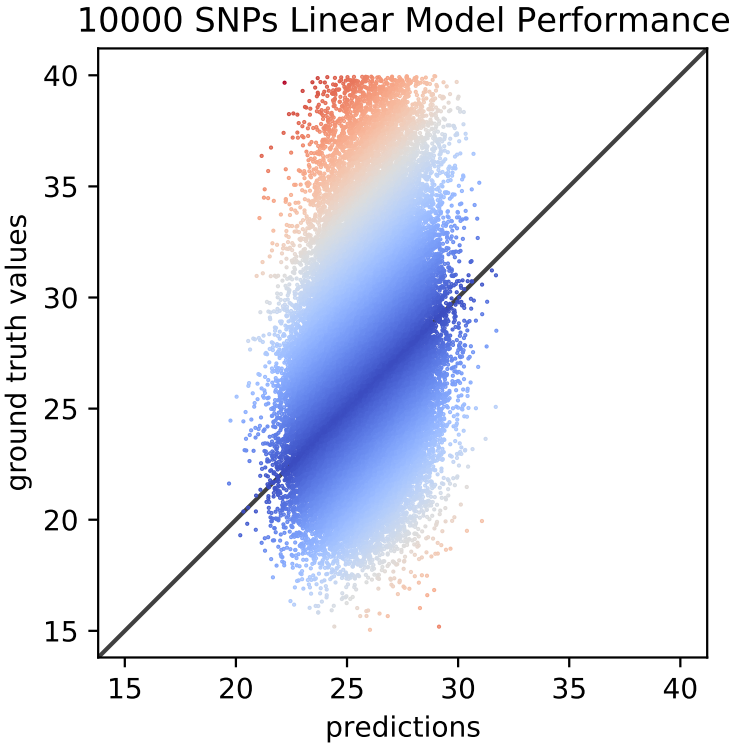


Figure 5.9: Prediction vs ground truth graph for the best 10K SNP linear model on the test data.

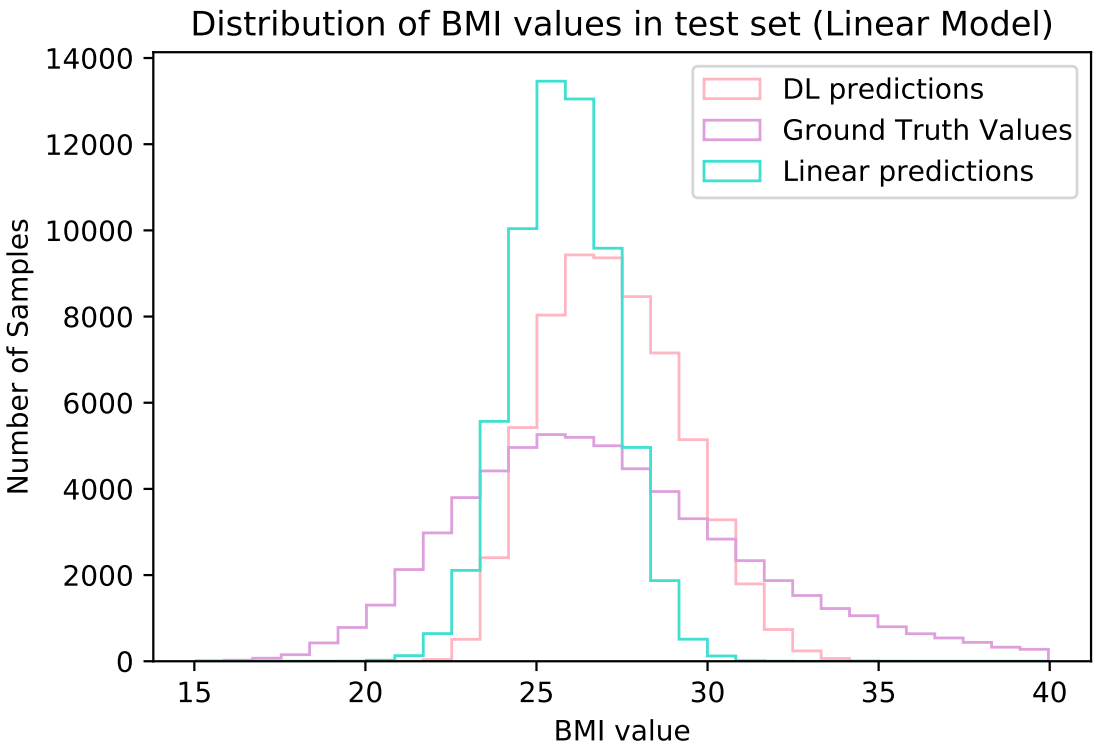


Figure 5.10: The distributions of predictions of the best linear model and best deep learning model, in comparison to the actual distribution of the data.

of SNPs, with particularly marked improvements in the 5000 and 10000 SNP sets; indeed, the 10000 SNP set more than doubles the R score of the 100 SNP set. This indicates that the SNPs reaching statistical significance in the GWAS capture only a small fraction of the genetic basis of BMI, and confirms that the aetiology of this phenotype is highly polygenic, suggesting that a large number of individually weakly contributing "background" SNPs are being used to inform predictions. These results also confirm that the deep neural network is better able to capture the polygenic nature of the trait, as its performance improves relative to the linear baselines as more background SNPs with more possibility of nonlinear relationships to the phenotype and complex interaction with one another are included. On the other hand, the fact that the performance of the linear models is not too much worse than that of the DL models indicates that the majority of genetic interactions for BMI are linear and additive, which is unsurprising given the success of the 2 million variant PRS measure by Khera *et al.* [128]. However, the widening performance gap between linear and DL models as the p-value threshold for SNP selection is relaxed suggests that more significant nonlinear interactions may be occurring between SNPs identified as less

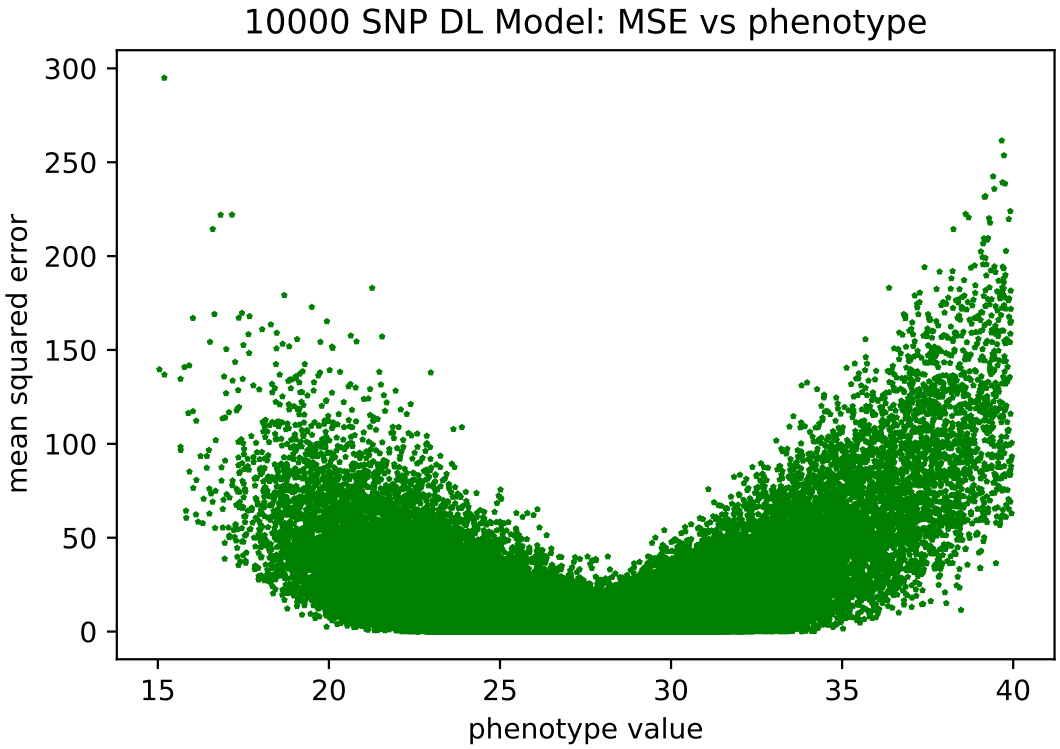


Figure 5.11: Mean squared prediction error versus phenotype for the best deep learning model on the test set.

significant by GWAS, and that SNPs that contribute to BMI via interactions with other SNPs may not be picked up as highly significant by GWAS.

5.3.3 Effect of Encodings

I tested all six encoding options for each model (linear and DL) and SNP set, and no single encoding was a standout in terms of performance, but the projection and effect direction based encodings (encodings 3-6 inclusive) did collectively appear to offer a modest performance increase over the more traditional SVE and one-hot options for the deep learning models trained on the larger (1000+ SNPs) input datasets (see figure 5.6). Encoding 1, which was the standard [0,1,2] variable encoding, underperformed across all models, though the difference was typically small, particularly in the smaller SNP sets, with as little as 0.005 difference in Pearson R in the 100 set. Encoding 2 (one hot encoding) generally improved upon encoding 1 but not significantly, and it also used significantly more memory than the other encodings, leading to memory capacity

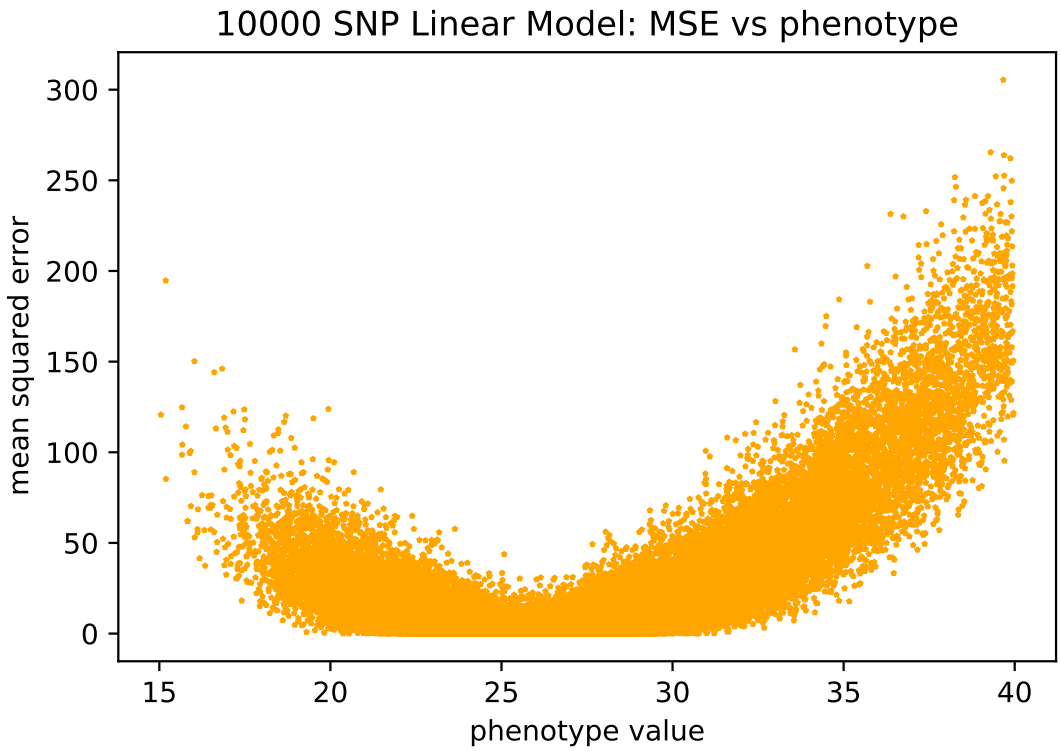


Figure 5.12: Mean squared prediction error versus phenotype for the best linear model on the test set.

problems for the 50k set and to slower running times across the board. There was never much difference between the four projection encoding options, with encoding 4 (REF/ALT projection with homozygosity encoded as [1,1]) most frequently performing the best, but usually by extremely narrow margins. The effect direction encodings (5 & 6) typically performed about the same as the other projection encodings (3 & 4), suggesting that the network was not able to glean any additional useful information from knowing which variants had positive or negative individual effects, potentially indicating that their combined effects differed from the individual ones in a significant way. Encodings 4 & 6, which encoded heterozygosity with a slightly smaller effect magnitude than than the homozygous genotypes, tended to marginally outperform encodings 3 & 4, which encoded all three genotypes with the same magnitude. This suggests that treating the heterozygous case as of weaker importance than either homozygous case may have provided some useful information for prediction for the models. For the 50K set, the linear models and encoding 2 (one-hot) of the deep learning model became intractable.

From these results it is not entirely clear which encoding is definitively the best for SNP

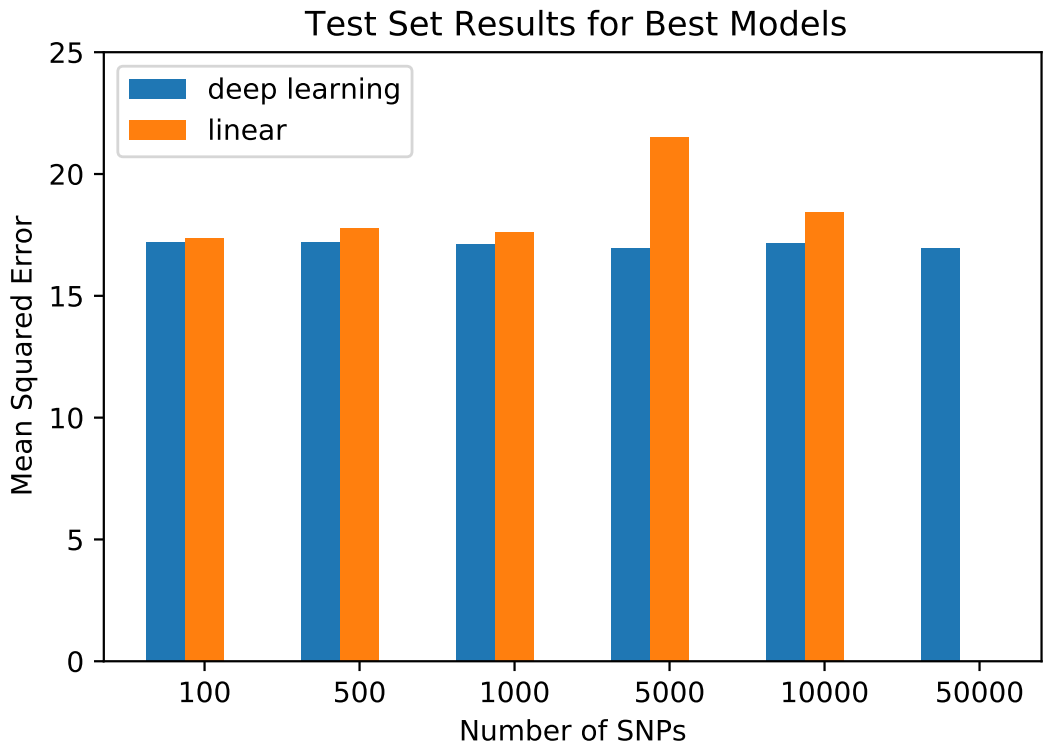


Figure 5.13: Graph comparing the test set MSE score for the best configuration/architecture of the overall best regressors (linear baseline and neural network), for each SNP set.

data, but it appears that alternatives to the SVE do potentially offer improved learning for ML models, and that less computationally intensive alternatives to one-hot encoding can offer at least equivalent performance. It appears that none of the encoding methods tested were definitively good representations of the underlying genetics, and that further research into the best ways to represent genetic data at the SNP level is warranted.

5.3.4 Performance analysis of different deep learning models

As mentioned, the CNNs were plagued by memory issues meaning that it was not possible to finish the gridsearch phase.

The Bayesian neural networks, aside from the 50000 SNP model, were able to complete training, though they were significantly slower. However, they showed significant signs of underfitting, which appeared to worsen as the SNP sets got larger. This can be seen in figure 5.7; it is

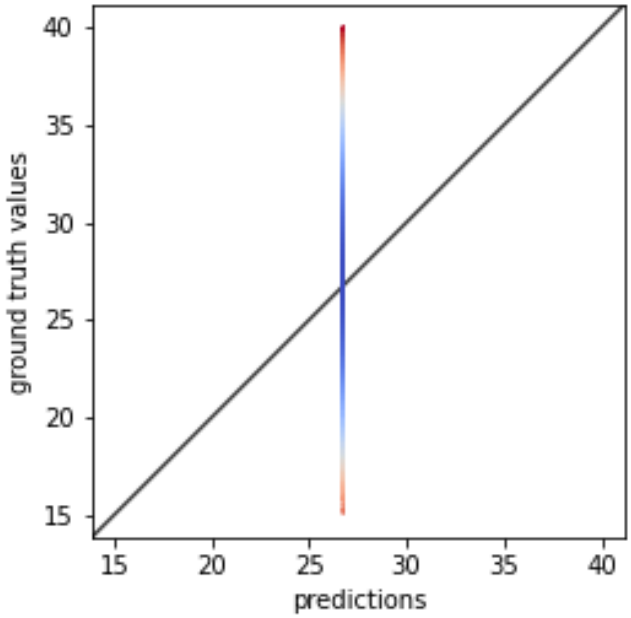


Figure 5.14: Graph of the predictions of the bayesian neural network against the ground truths for the 10000 SNP test set.

difficult to calculate MSE on a BNN so only Pearson R was used for comparison with the other models. This is unsurprising since most of the feedforward models were also erring on the side of underfitting; Bayesian neural networks are robust to overfitting but the variability in parameters introduces more uncertainty and can make it harder for the network to fit to the underlying distribution. The full scope of this underfitting is clearly illustrated by the prediction vs ground truth graph in figure 5.14, from which it is apparent that the model only predicts the mean value of the dataset.

Chapter 6

Interpretation for BMI Prediction

The aim of this chapter is to explore issues surrounding deep learning model interpretation for genetic data. The interpretation of deep learning models is still a relatively new field and lots of different types of approaches exist, which bring their own benefits and drawbacks; not all of these approaches will be suitable for all models, and so the choice of interpretation approach may necessitate certain constraints in model design. Even for model-agnostic interpretation methods, taking into consideration the type of input data and the problem space can greatly improve or limit the utility of interpretation results. In this chapter we will consider how best to design or adapt interpretation methods for SNP data.

The rest of this chapter is arranged as follows. Section 6.1 outlines a range of possible interpretation methods that have been used for models trained on genetic data in previous studies, providing an understanding of the theoretical basis of these methods as well as their strengths and weaknesses. Section 6.2 discusses some novel ways in which interpretation methods have been augmented for this project to make them better adapted for drawing useful conclusions from models trained on SNP data, as well as describing the sample preprocessing procedure for the interpretation step. Section 6.3 discusses the results of the different interpretation methods performed.

Study	Interpretation Method	Single Features or Interactions
Bhatt, Lucek & Ott 1999	Network weights	Single features only
Behravan et al. 2020	SHAP	Single features and interactions
Badré et al. 2021	DeepLift, LIME	Single features only
Tomita et al. 2004	Statistical overrepresentation analysis	Single features and interactions
Mieth et al. 2021	Layerwise Relevance Propagation & association testing	Single features only
Van Hilten et al. 2021	Biologically informed network design & network weights	Single features only
Yin et al. 2019	Biologically informed network design & selected inputs	Single features only
Montañez et al. 2018	Stacked Autoencoders, Association Rule Mining & Frequent Pattern Matching	Single features only
Cui et al. 2022	Biologically informed network design & network weights	Single features and interactions
Lee et al. 2022	GMDR & Pathway Enrichment Analysis	Interactions only
Motsinger et al. 2006	Biologically meaningful network design & network weights	Single features and interactions
Moore et al. 2006	MDR + model interpretation	Interactions only
Motsinger et al. 2008	Extraction of decision tree from trained network	Single features and interactions

Table 6.1: Literature review table showing the different interpretation approaches used on machine learning models in previous studies.

6.1 Interpretation Methods

Table 6.1 shows a brief summary of the interpretation methods used in the previous phenotype prediction studies that interpreted their trained models. As you can see in the table, quite a variety of different approaches have been used and there isn't much consensus on the best approach. Many interpretation methods rely on the neural network architecture itself to have interpretable biological meaning, a design approach we chose not to pursue for this study due to the lack of well-validated biological knowledge of the genetic basis of BMI and the evidence of potential interpretation pitfalls with this approach on poorly understood polygenic phenotypes [301]. Of the methods that don't rely on a biologically meaningful network architecture, there are a mixture of gradient-based (DeepLift, LRP), statistical (overrepresentation analysis, association testing) and perturbation-based (SHAP, LIME) methods, and many studies used a combination of several methods in conjunction. Badré *et al.* were the only team to compare multiple different interpretation methods on the same network, and found that there was low agreement between the features identified as important between the two methods [13], highlighting the importance of method choice. Of the studies who interpreted models trained on real instead of synthetic data, a few found that their derived important SNPs or genes/regions did not match existing domain knowledge at all, suggesting possible methodological issues [195, 217].

A commonality of all the reviewed research in table 6.1 is that the interpretation methods utilise the same features as the deep learning model does. For biologically designed models, this is logical as the network architecture, including input features, has been chosen with an interpre-

tation in mind. For other methods, such as Layerwise Relevance Propagation, it is a necessary requirement of the method. However, for perturbation-based methods such as LIME or SHAP, there is the opportunity to utilise features that are designed specifically for the interpretation step, and these features need not be the same as the input features of the model itself. This means, in the case of genotype-phenotype prediction models, that we can theoretically "have our cake and eat it too": we can design models that do not make limiting assumptions about involved regions, genes or SNPs, *and* we can make use of biological priors to make our explanations more useful and intuitive. We can do this by designing deep learning models with single SNPs as inputs, and then designing complementary perturbation-based interpretation methods which use "explainable" features that have more biological meaning, such as genes, regions or pathways. This saves us the usual difficulty of trying to translate the interpretation results into biological knowledge via a method such as gene enrichment analysis - we can instead design a single-step interpretation method that takes us directly to the biological insights. In the following subsections we will discuss several interpretation methods that lend themselves to the design of biological features, and contrast them with some that don't.

6.1.1 Feature Ablation

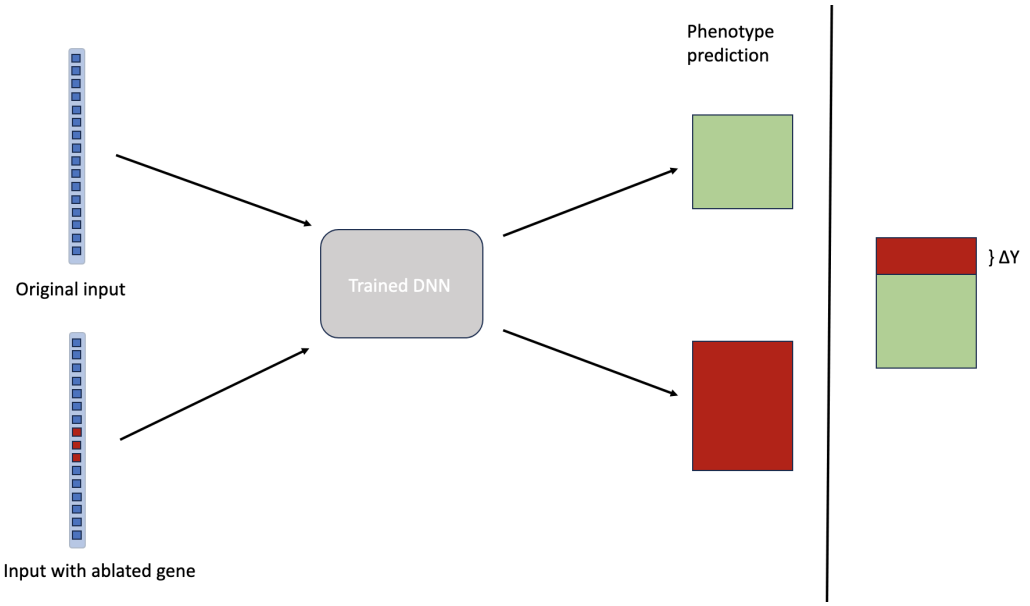


Figure 6.1: Illustration of the single gene feature ablation process for a single gene of a single sample.

Feature ablation is the process of perturbing a sample such that feature/s of interest are

replaced by a baseline value and the importance of the ablated feature is computed using the difference in output [188]. Being a perturbation-based method means that it does not require the ablated features to be the same as the original inputs; they may be either single input features or groups of features considered together to create features that are more comprehensible for humans as an explanation of the model's behaviour. In my case I have grouped the input SNPs into genes (using the closest gene for noncoding SNPs), as these are more biologically meaningful than individual point mutations, as usually the goal of interpretation of GWAS data is to find implicated genes and thus potential functional pathological pathways. Since I subdue the effects of entire genes at once, this feature ablation method can be considered to be conceptually similar to a genetic knockout study, in which the expression of genes of interest in an organism is artificially reduced in a lab in order to see how each gene affects a trait of interest. If the resulting difference in output of the model due to a perturbation of a gene is greater than expected, this indicates that the model uses this feature preferentially for prediction and sheds some light on the model's "understanding" of the relationship between features and predictions for the given sample. The process is illustrated in figure 6.1.

If we begin with a model $f(x)$ trained on data for sample X_k, and we are interested in learning about the importance of features in the interpretable feature set $\mathbb{F}$, we may obtain a prediction y as follows:

$$Y(X_k) = f(X_k; \mathbb{F}) \tag{6.1}$$

We then may observe the importance of explainable feature i, where $i \in \mathbb{F}$, by making a new prediction on the same sample that sets the value of i to its baseline so that it is not useful for prediction:

$$Y_i'(X_k) = f(X_k; \mathbb{F} \setminus \{i\}) \tag{6.2}$$

The quantity we are interested in is the resulting *difference* resulting from the removal of feature i, corresponding to the marginal contribution of i, defined as:

$$\Delta Y_i(X_k) = Y(X_k) - Y_i'(X_k) = f(X_k; \mathbb{F}) - f(X_k; \mathbb{F} \setminus \{i\}) \tag{6.3}$$

The quantity of $\Delta Y_i(X_k)$ can then be compared to or aggregated with the effects of ablation on

other genes in the same sample $\Delta Y_j(X_k)$, the same gene in other samples $\Delta Y_i(X_l)$, or the same gene and sample in the null hypothesis distribution, discussed below. By examining the effect of removing i from feature set $\mathbb{F}$, instead of computing $f(X_k;\{i\})$ directly, we are accounting for the possibility of interaction with other features, the effects of which would not be seen in $f(X_k;\{i\}$. These interaction effects are contained in the prediction on full feature set (F) but will be ablated along with the main effects by removing feature i and the resulting difference $\Delta Y_i(X_k)$ will thus contain interaction effects between i and any other features. An algorithmic representation of this process can be seen in figure 6.2.

```
FOR sample in sample set
    Function PREDICT(sample): pass sample through DL model
    FOR gene in gene feature set
        DO find related variants
        DO set related variants ← [0,0] baseline
        DO PREDICT(sample)
        DO calculate absolute prediction difference per gene
calculate mean absolute prediction difference per gene

```

Figure 6.2: Algorithmic representation of the single gene feature ablation process.

6.1.1.1 Generating a Null Hypothesis Distribution for Feature Ablation

While one of the major strengths of feature ablation is its non-parametric approach - since no simplifying assumptions are made about the relationship between $Y(X)$ and X, the method is able to detect the effects of any resulting relationship the original model has learned, regardless of complexity - this also creates one of its major drawbacks, namely the difficulty in establishing what constitutes an "important" gene signal. In cases where the resulting $\Delta Y(X)$ is similar for a large number of the genes tested, a not unlikely scenario for a highly polygenic phenotype, it is especially difficult to decide where to put the importance cutoff. Creating a null hypothesis distribution to use as a comparative baseline is a way to empirically establish the point at which a signal becomes "important".

The aim of a null hypothesis distribution is to provide a selection of results that represent the

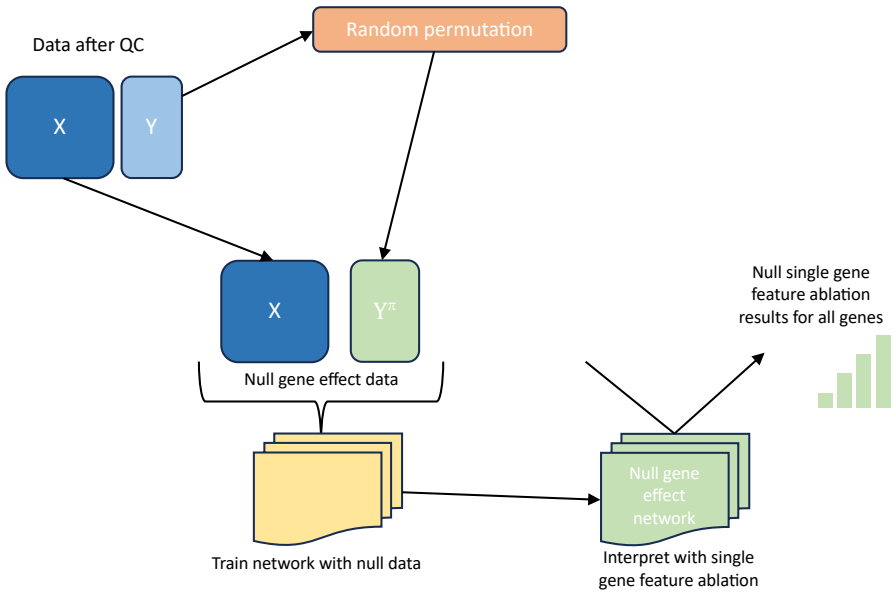


Figure 6.3: Illustration of the procedure used to generate a null result for feature ablation. A new copy of the network is trained on data where the Y values have been randomly permuted to destroy the real gene effects and ensure that the network only learns random feature association. The results of interpreting this new network copy can be used to indicate the upper limit of interpretation results that can be generated when no true effects are present. Several network copies trained on different random permutations of the Y data should be used to generate a null distribution.

condition of the null hypothesis being true, so that the results generated from the actual data can be compared against this. If the results from the real data are significantly different than most of the null hypothesis distribution, this provides a sound and compelling argument in favour of rejecting the null hypothesis.

In our case, the null hypothesis for single gene feature ablation is that a gene of interest i has no main effect on the phenotype. To generate a null hypothesis result for sample X_k, we must train a network in which we can guarantee no relation between gene i and $Y(X_k)$ will be learned, so that we can then compute $\Delta Y_i(X_k)$ for that network to use as a null baseline. This is done by randomly permuting the Y values for the entire dataset before training the network with the goal of any relationship between X and Y. Performing feature ablation for each gene on each sample of interest will then generate a distribution of possible results in the case of the null hypothesis being true. On the off-chance that a particular network has learned a spurious correlation between one of the features and the permuted phenotypes, several networks may be trained and ablated to provide a truly random distribution of feature ablation scores. A p-value

can then be calculated for feature ablation results from networks trained on the real unpermuted data, quantifying how likely these results would be in the case of the null hypothesis being true. This procedure is illustrated in figure 6.3.

6.1.2 Local Interpretable Model-Agnostic Explanation

Local Interpretable Model-Agnostic Explanations, or LIME, is an algorithm that aims to create an explanation of a prediction by any machine learning model by approximating it at a local level with a simpler model [252]. An explanation is defined as a representation of a relationship between input variable and response that conveys a qualitative understanding of the model's decision. LIME explanations do not aim for global fidelity to the predictor model, they only aim to explain the model's behaviour in the vicinity (in feature-space) of the sample being explained. The algorithm takes advantage of an interpretable representation of the input features to craft an explanation that's comprehensible to humans; in some cases, where the input feature space is already fairly simple, this may be the same as the representation used by the predictor model, but in other cases it may be a simplified representation chosen specifically for comprehension. A more simplistic and comprehensible model, such as a linear regression model or decision tree, is then used to explain the relationship between the interpretable representation of the input features and the output in the feature-space vicinity of the sample in question. Since the characteristics of the more complex predictor model are irrelevant to the mechanism of LIME, it is said to be model-agnostic.

The explanation generated by LIME can be understood as

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{6.4}$$

where f is the predictor model, such that $f(x)$ is a prediction made by the model given input x, and g is the interpretable model, which is a member of the class G of potentially interpretable models such as decision trees. Since not all individual instances of potentially interpretable models are actually comprehensible to humans (for example, a linear regression model with thousands of features is not easily interpretable), $\Omega(x)$ is a measure of model complexity, which

should be kept low in order to preserve interpretability. $\mathcal{L}(f, g, \pi_x)$ is then a loss function that is a measure of how unfaithful the interpretable model g is to the predictor model f in the local vicinity defined by π_x. So the goal is to maximise local fidelity while still keeping the complexity of the interpretable model low enough to be comprehensible to humans. Theoretically, different classes of explainable models G, fidelity functions $\mathcal{L}$ and complexity measures Ω can be used, but the creators of LIME used sparse linear models as their class of interpretable models, and I have done the same.

Additionally, if $x \in \mathbb{R}^d$ denotes the representation of a sample x in the original feature space used by the predictor model, then $x' \in \{0,1\}^{d'}$ is a binary vector representing the presence or absence of d' interpretable features. In order to minimise the loss function $\mathcal{L}(f, g, \pi_x)$ for a given sample input without making assumptions about the function f (in order to preserve model agnosticity), samples are drawn from the vicinity of x' using a perturbation method, where individual features within the interpretable space are set to zero or nonzero values at random, with the number of zero and nonzero elements in each perturbation also being uniformly randomly sampled. This generates a perturbed sample $z' \in \{0,1\}^{d'}$, which can be converted back to the feature space of the original predictor model to yield $z \in \mathbb{R}^d$. This perturbed sample in original feature space can then be fed into the predictor model f to produce a prediction, which will be used as a label for the interpretable model, to help it learn the local behaviour of the predictor model as inputs vary. After many perturbations, a dataset $\mathcal{Z}$ of perturbed samples labelled with their predictions by the original model is obtained, which can be used to train the interpretable model. A similarity kernel can be chosen to help prioritise perturbations in the feature-space vicinity of the original sample and down-weight those further away.

LIME has several advantages, such as its model-agnosticity, which means it can be used to compare the underlying "reasoning" of different types of models trained on the same data, as well as the flexibility in the design of interpretable features, meaning that it can be used even in situations where the original input feature space may be large, multimodal and difficult to explain. It also has a number of shortcomings. One of these is that it is designed to be applied per-sample, which not only means that it is difficult to draw larger conclusions about the model's performance

on a dataset or population, but also makes it more vulnerable to per-sample fluctuations in network performance which are generally smoothed out by aggregating results across a dataset; in other words, the derived explanation is only ever as good as the network's performance *on that sample*, even if the network performs much better on other samples in the data. Another issue is that the random perturbation method is liable to produce unrealistic samples with combinations of features that would never occur organically, meaning that the network is unlikely to have been exposed to similar samples during training. This means its predictions on these generated outlier samples may not be robust, which could potentially affect the weighting of explainable features in the explainable model. Finally, it is often difficult in practice to achieve the desired trade-off for the explainable model between fidelity to the original model and complexity; a complex original model tends to lead to a complex (and therefore less explainable) explainable model, regardless of how you design the interpretable features.

6.1.3 Layerwise Relevance Propagation

Another method which aims to explain a model's reasoning, but with quite a different approach, is Layerwise Relevance Propagation, or LRP [196]. The creators of LRP make the argument that in many cases it is difficult to perform feature selection for complex prediction tasks, as the exact way that different input features are used may be nuanced and non-fixed, so it is better to train a model with many input features to maximise the possibility of good prediction performance, but assume that not all of them will actually be useful. In this case it may be important to be able to tell which combination of input features is actually being used by the model in order to try to assure oneself that the basis of prediction is logical and will thus generalise to unseen data. Aside from this, examining which features are used by the model post-hoc can also allow researchers to go back and remove unused features and re-train the model on a "cleaned" feature set, reducing excess overhead without impacting performance.

A downside of LRP is that since it examines the importance of individual input features, it is not well-suited to prediction problems in which the input is highly multidimensional and granular, meaning that no individual feature is likely to be very important on its own. In this case it is the combination or interaction of multiple features working in conjunction that is of

more interest from an interpretation standpoint than individual feature importance. For example, in the case of my research, it is already known that no individual SNP is likely to be notably more important for phenotype prediction than the others, otherwise this SNP would have already been identified during GWAS. Hence the main thing that LRP is able to do for models trained with SNP data is validate the model by re-identifying statistically important GWAS SNPs, which does not contribute much to the corpus of biological knowledge.

6.1.4 Shapley Values

Shapley values are a concept borrowed from cooperative game theory which can be used in machine learning model interpretation to quantify the importance of different features to a prediction. In their game theory form, Shapley values fairly determine how much an individual player contributed to the outcome of a cooperative multiplayer game, i.e. the marginal contribution of each player [260]. By viewing machine learning models as a game in which features "collaborate" to produce an outcome (the prediction), this framework can be used to explain the reasoning behind a given prediction.

Like in LIME, Shapley values are calculated by perturbing the original input and observing how this affects the model's prediction. Normally, some kind of comparison set is selected to serve as a baseline against which to compare the marginal contributions of features; this can be the entire dataset, or it can be a subset containing a specific value for a feature of interest. A unique property of Shapley values is that they satisfy a set of rigorous axioms that most machine learning feature attribution methods do not. These include:

1. Completeness: the sum of feature contributions must add up to the difference in the prediction from the baseline, meaning that explanations generated using Shapley values are always complete.

2. Dummy: a player that does not contribute to the outcome of the game must be assigned a Shapley value of zero, meaning that features which are not used by the model will not be given any importance in the explanation.

3. Monotonicity: Shapley values of players that consistently contribute more to the outcome than other players must be higher, meaning that greater trends in the contributions of features will be reflected in the explanation.

To calculate the true Shapley value, the marginal contribution of each feature - i.e. the incremental contribution of the feature combined with the interaction with other features - must be computed across a variety of perturbed samples in which varying selections of features are set to a baseline value, as in LIME, but in this case, the baseline is usually set to be the average value of the feature in the comparison set. The Shapley value of the feature is then the weighted sum of the marginal contributions of the feature across different perturbations. Calculating the marginal contributions of many features across many perturbations is highly computationally expensive, usually prohibitively so for any model with a reasonable number of features, and hence a variety of algorithms exist to closely estimate Shapley values instead of calculate them directly. One of these is SHAP, discussed below.

Shapley values, again like LIME, are model-agnostic, as they only require black-box access to the model in order to be able to feed in perturbed inputs and access outputs. They also have the great advantage of being computable at the local level (on a single sample) while also being easily extrapolated to groups of samples, giving a glimpse into the more global reasoning of a model. As with other perturbation methods, unrealistic inputs (where perturbation creates a combination of features that for whatever reason could not actually occur in the data) is a potential pitfall. As discussed above, directly computing Shapley values based on the marginal contribution of each feature to the final prediction is prohibitively computationally expensive for deep learning models in most cases. One way of estimating Shapley values more efficiently is known as SHapley Additive exPlanations or SHAP [170]. Like LIME, SHAP employs a linear explanation model to learn to quantify the contribution of features using a perturbation-based approach. Also like LIME, the input features can be grouped into interpretable features that are more human-comprehensible than the feature space of the original input. This is an efficient way to estimate Shapley values in comparison to direct computation, and has the advantage that the resulting feature attributions are more meaningful and easier to compare across models and datasets than those produced by LIME.

6.2 Design of Interpretable Features

As discussed in the previous chapter, a major benefit of LIME is that the interpretable model is not constrained to use the same input features as the original model, allowing a set of features to be chosen such that the explanation generated by LIME is maximally useful or meaningful. This is also the case for other model-agnostic perturbation-based methods, such as feature ablation. In my case, the goal was to design a set of interpretable features that could provide insight into the genetic architecture of BMI. Using the same features as the original model - i.e. individual SNPs - allows for importance scores to be calculated per-SNP, which may be informative but suffers many of the same limitations as GWAS, namely that the effect sizes of individual SNPs in polygenic diseases tend to be very small and so this knowledge is not immediately useful for biological or clinical purposes. Instead using biologically-informed groupings of SNPs as interpretable features may allow for broader genetic insights, as well as allowing the discovery of effects that may be missed during a GWAS.

There are many potential groupings that could be used for this purpose, but an obvious choice is genes. Most SNPs in the UK Biobank Axiom array are non-coding, i.e. not actually situated within a gene, so to group SNPs into related genes I used the KEGG [124] database to link SNPs to the nearest gene; this means while associated SNPs may be in or close to the actual gene, they are more likely located in a regulatory region that influences the gene via for example enhancer pathways. SNPs not associated with any gene (3101 in the top 10000 set) were excluded from interaction analysis. SNPs associated with multiple genes (937 in the top 10000 set) were included in multiple gene features, such that if either or both of the gene features were perturbed, this SNP would be as well. In order to be able to perturb samples it is necessary to have a way to switch genes "off" before feeding inputs into the original model, thus I defined a baseline non-influential state of individual SNPs to be encoded as [0,0], corresponding to 0 magnitude along either REF or ALT axes.

I then was left with 1366 interpretable features in the form of genes, instead of 10000 SNP features. I hypothesised that different BMI subgroups in my sample may have different underlying genetic importance structures, so I stratified my sample by BMI category (underweight, healthy

weight, overweight, obese and extremely obese - see figure 4.2 for cutoffs) for the explanation phase. I trained a linear explainable model for samples in each BMI subgroup using LIME, where each model was trained using the output of the deep learning model as discussed in the previous chapter. The coefficients of these linear regression models can then be used as feature importance scores for individual genes for each BMI category, and genes with high importance scores throughout the BMI category subgroup may indicate genes involved in the development of obesity or extreme obesity, as well as genes involved with maintaining a healthy body weight.

6.2.1 Augmentation of LIME methods for genetic data

As discussed previously, sparse SNP data is unique in structure and any existing deep learning methods will need to be adapted to work with this kind of data. When adapting LIME to work with this kind of data it was first necessary to create a function to transform between the original feature space and the interpretable feature space, with interpretable features corresponding to groups of SNPs associated with the nearest gene. This method made use of gene masks, which could ablate SNPs corresponding to a given gene while leaving the rest of the sample unaltered, and could also be multiplied together element-wise to create joint masks corresponding to several genes being switched off, which allowed for SNPs to be associated with multiple genes and set to baseline if any of their associated genes were ablated.

It was then necessary to create a specialised perturbation function. Perturbation took place in the interpretable feature space and involved a random number of randomly selected genes being set to zero. The resulting binary gene vector could then be translated back into the original feature space using the joint gene mask. In order to quantify how similar a perturbed input was to the original input, it was necessary to define a similarity kernel. A common choice is an exponential kernel, but since the the original feature space is so large this can result in vanishingly small weights given to perturbations with even a few ablated genes. Instead I used a similarity function of $1 - \frac{|x-x^{\pi}|}{\gamma}$, where x is the original input, x^{π} is the perturbed input, and γ is a scaling factor provided at runtime. The best choice of γ generates a result that quantifies similarity as a proportion with 1 being an exact match; for example, since the length of x and x^{π} is 20,000 for a 10,000 top SNP feature space encoded in any projection encoding, in this case

20,000 would be a good choice for γ.

6.2.2 Aggregation of Interpretation Results

LIME and feature ablation are both only designed to be used at the individual sample level, with a new linear model trained for each sample in the case of LIME, so the question of how to generate results that are meaningful for a whole population is not immediately obvious. In our case, we wanted to discover aetiological trends that would be more broadly applicable rather than only look at individual causes, so we stratified individuals by BMI subcategory and kept only those whose BMI the model was able to predict with a high level of fidelity (prediction RMSE cutoff of 0.1) in an attempt to create a fairly homogeneous set, and then aggregated statistics across the set to produce an average attribution coefficient for each interpretable feature for that population. Since the direction of effect for individuals would be different depending on the original genotype (for example, someone who was already homozygous for the risk alleles in FTO would see a decrease in BMI as a result of blocking the FTO gene, while the opposite would be the case for someone who did not have any of the risk alleles), the absolute values of the resulting differences were utilised for aggregation so that the overall importance of the gene could be adequately assessed without being confounded by the sign difference for different samples. Attribution scores across individuals in the interpretation discovery set were thus aggregated by taking the mean absolute attribution per gene, which provided a measure of overall importance for the set.

6.2.3 Selection of Samples for Interpretation

Samples for interpretation were selected from the holdout test set, meaning they had not been used at any point to train the model. This was to minimise any chances of data leakage, whereby the model might somehow be recognising artefacts in the individual samples rather than more universal patterns in the genetic data. The total number of samples in the holdout test set was 62095. Samples were then filtered by prediction error, with only samples where the MSE of the network's prediction was less than 0.1 were retained for interpretation, corresponding to the top 25% of predictions. This was done to ensure that interpretation was only performed on

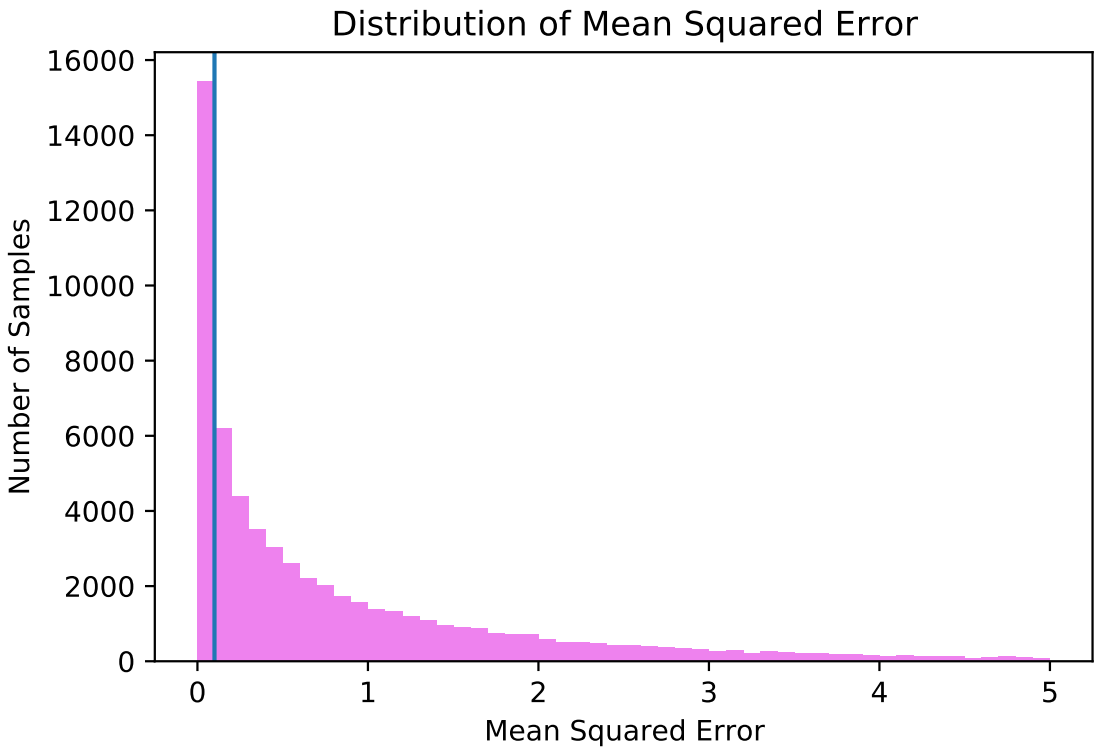


Figure 6.4: Plot showing the distribution of Mean Squared Error values across samples in the holdout test set. The blue line shows the cutoff for samples used in the interpretation phase.

samples where it seemed most likely that the network had learned something accurate about the underlying genetic causes, thus reducing noise in the interpretation results. Finally, since the underlying genetics of obese vs normal BMI could be different with different genes implicated, it is possible that the network prioritises different genes to make predictions for individuals in different BMI subgroups, so individuals were stratified by BMI subcategory and interpretation for obese samples and healthy/underweight samples were performed separately. After these filtering steps, I was left with 8979 samples with MSE < 0.1 in the obese discovery set. This was divided into two equal-sized subsets which were analysed and had results aggregated separately to test the robustness of the method; large differences in the important genes between the two sets would indicate that the methods used were highly susceptible to small differences in the data and would be very unlikely to generalise well.

6.3 Interpretation Results

6.3.1 Single Gene Feature Ablation Results

Gene Name	Mean Absolute Effect on BMI	Prior Relevant Associations
FTO	0.208	BMI, lean body mass, body weight
BDNF-AS	0.080	BMI, body weight
LINC01122	0.077	BMI
CPNE4	0.076	BMI
NRXN1	0.076	BMI
CSMD1	0.074	BMI
MSRA	0.074	BMI, body weight
MAP2K5	0.073	BMI
LINGO2	0.069	BMI
SGCZ	0.069	BMI
SOX5	0.069	BMI, body weight
ZFPM2	0.067	BMI
RBFOX1	0.064	BMI, body weight
CTNNA2	0.063	BMI, body weight
AGAP1	0.061	BMI, body weight
DLG2	0.061	BMI, body weight, fat body mass
ZBTB20	0.061	BMI
ADCY9	0.060	BMI, body weight, body surface area
RPTOR	0.059	BMI, body fat %
NLGN1	0.059	BMI, body fat %, body weight

Table 6.2: Test Set 1, Obese individuals. Top twenty genes as discovered by single gene feature ablation, their mean absolute difference on BMI, and prior relevant biological associations. Biological associations sourced from genecards.org.

As you can see from tables 6.2, 6.3, 6.4, and 6.5, there is a high level of agreement between the results of analysis performed on discovery set 1 and discovery set 2, suggesting the method is relatively robust to data variations and should be able to generalise to new data. The results also are similar regardless of whether the chosen samples were from obese individuals or healthy/underweight individuals, suggesting that the trained model appears to look for the same patterns in all individuals regardless of BMI category.

In all discovery sets the FTO gene is a standout, with an effect size that differs by a factor of ten, or greater than fifteen standard deviations, from the mean effect size across all genes. The average difference conferred to the network's prediction by the exclusion of the FTO gene would

Gene Name	Mean Absolute Effect on BMI	Prior Relevant Associations
FTO	0.208	BMI, lean body mass, body weight
BDNF-AS	0.084	BMI, body weight
NRXN1	0.081	BMI
LINC01122	0.080	BMI
MSRA	0.078	BMI, body weight
CPNE4	0.076	BMI
CSMD1	0.075	BMI
MAP2K5	0.073	BMI
ZFPM2	0.072	BMI
SGCZ	0.072	BMI
SOX5	0.071	BMI, body weight
LINGO2	0.071	BMI
CTNNA2	0.066	BMI, body weight
RBFOX1	0.063	BMI, body weight
AGAP1	0.062	BMI, body weight
ADCY9	0.062	BMI, body weight, body surface area
DLG2	0.062	BMI, body weight, fat body mass
ZBTB20	0.062	BMI
RPTOR	0.061	BMI, body fat %
PDE4B	0.060	BMI, body fat %, body weight

Table 6.3: Test Set 2, Obese individuals. Top twenty genes as discovered by single gene feature ablation, their mean absolute difference on BMI, and prior relevant biological associations. Biological associations sourced from genecards.org.

be equivalent, for a person of average height, to a 600g weight difference; this is an intuitive result, since the greatest weight difference caused by the FTO gene (the difference between homozygosity for the protective allele vs the risk allele) has been found to be about 3kg [82]. The second most important gene, BDNF-AS, is the antisense partner of BDNF, a brain-related gene shown to be heavily involved in the causation of severe monogenic obesity [22, 97]. It is somewhat surprising that the network found BDNF-AS to be more important than BDNF, and raises the question as to whether the network may have been deriving "leaked" information about BDNF via BDNF-AS, since they have overlapping associated SNPs but BDNF-AS has more associated SNPs (54) in my 10k dataset than BDNF (16), and all but two of the BDNF associated SNPs are also included in the set for BDNF-AS. This highlights a potential issue with using databases such as KEGG for grouping SNPs, and suggests the possibility that other genes on this list may be "standing in" for an overlapping gene. In any case, all genes in the top of the importance distribution have documented research associating them with BMI and other

Gene Name	Mean Absolute Effect on BMI	Prior Relevant Associations
FTO	0.201	BMI, lean body mass, body weight
BDNF-AS	0.084	BMI, body weight
NRXN1	0.079	BMI
CPNE4	0.077	BMI
LINC01122	0.077	BMI
MSRA	0.075	BMI, body weight
CSMD1	0.074	BMI
MAP2K5	0.073	BMI
ZFPM2	0.071	BMI
SGCZ	0.069	BMI
LINGO2	0.068	BMI
SOX5	0.068	BMI, body weight
RBFOX1	0.064	BMI, body weight
AGAP1	0.062	BMI, body weight
CTNNA2	0.061	BMI, body weight
ADCY9	0.061	BMI, body weight, body surface area
ZBTB20	0.061	BMI
TENM2	0.060	BMI
DLG2	0.060	BMI, body weight, fat body mass
PDE4B	0.060	BMI, body fat %, body weight

Table 6.4: Test Set 1, Healthy and underweight individuals. Top twenty genes as discovered by single gene feature ablation, their mean absolute difference on BMI, and prior relevant biological associations. Biological associations sourced from genecards.org.

related phenotypes. This result validates that the network has, without any deliberate input of such knowledge, automatically learned some important prior biological knowledge about BMI.

We can further validate this result using the null distribution, which was calculated following the procedure followed in section 5.1.1. As can be seen in figure 6.7, the vast majority of effects detected by the network trained on the null dataset were effectively zero, with the few higher importance scores tailing off at 0.025, whereas there is a tail of higher importance scores in the real data. This suggests that results over 0.025 would have been highly unlikely to impossible to generate from the null distribution, and are likely to be evidence of a genuine gene effect. 615 genes in the real dataset have mean BMI effects larger than this value, corresponding to 20% of the total genes tested.

None of the effect sizes for individual genes (including FTO) are large, which is to be expected

Gene Name	Mean Absolute Effect on BMI	Prior Relevant Associations
FTO	0.201	BMI, lean body mass, body weight
BDNF-AS	0.084	BMI, body weight
NRXN1	0.081	BMI
LINC01122	0.080	BMI
MSRA	0.078	BMI, body weight
CPNE4	0.076	BMI
CSMD1	0.075	BMI
MAP2K5	0.073	BMI
ZFPM2	0.072	BMI
SGCZ	0.072	BMI
SOX5	0.071	BMI, body weight
LINGO2	0.071	BMI
CTNNA2	0.066	BMI, body weight
RBFOX1	0.063	BMI, body weight
AGAP1	0.062	BMI, body weight
ADCY9	0.062	BMI, body weight, body surface area
DLG2	0.062	BMI, body weight, fat body mass
ZBTB20	0.062	BMI
RPTOR	0.061	BMI, body fat %
PDE4B	0.060	BMI, body fat %, body weight

Table 6.5: Test Set 2, Healthy and underweight individuals. Top twenty genes as discovered by single gene feature ablation, their mean absolute difference on BMI, and prior relevant biological associations. Biological associations sourced from genecards.org.

since any single gene with large effect on the phenotype would likely have already been discovered by existing methods. As this brute-force ablation method considers genes in isolation from one another, it is largely limited to reproducing existing biological knowledge about important genes and does not suggest specific candidates for investigation of interactions. We now move on to the more sophisticated method of LIME, and revisit feature ablation for specific pairwise analysis later.

6.3.2 LIME Results

LIME was performed on the same partitioned dataset as the single gene ablation analysis, with data stratified by BMI category. Because LIME generates a linear model and the gene importance scores are the coefficients that correspond with each gene, the actual values are less immediately interpretable than the occlusion results. As the trained linear model had 3013 gene features, the coefficients of each are also quite small, so it can be hard to determine what scores count

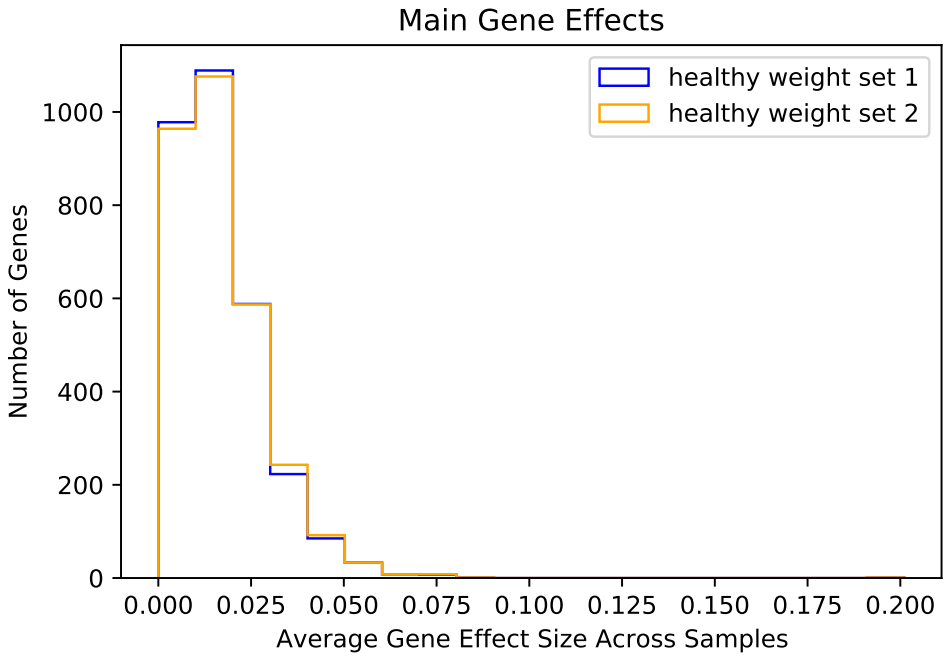


Figure 6.5: Histogram showing the high level of agreement between the results from healthy weight discovery set 1 and healthy weight discovery set 2.

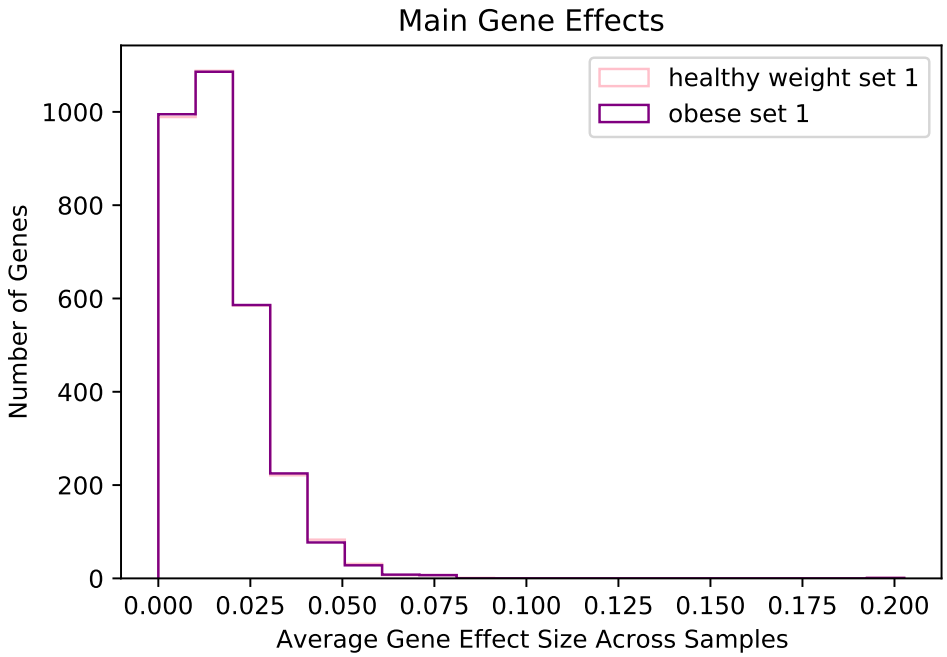


Figure 6.6: Histogram showing the high level of agreement between the results from healthy weight discovery set 1 and obese discovery set 1.

as important. For this reason LIME was also performed on the neural network trained on the permuted data, to produce a null distribution, which can be seen in figure 6.11. From this

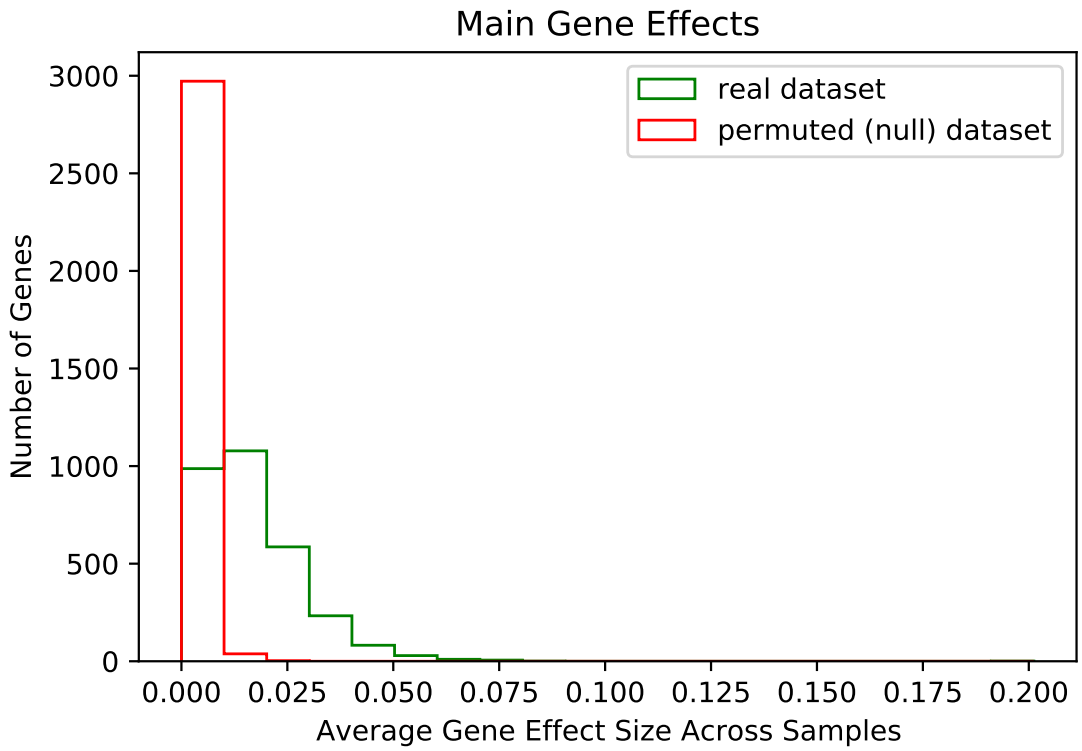


Figure 6.7: Graph showing the gene effects detected by probing the deep learning model trained on the null dataset, in comparison to those generated from the interpretation of the same model architecture trained on the real data.

distribution you can see that no genes in the null network achieved an importance score of greater than 0.00013, meaning that all scores above this could be considered to be evidence of genuine importance. However, since the network trained on the real phenotype data produced mostly importance scores above this threshold, it is not very useful for prioritising important genes. A cutoff of 0.000675 selects only the top tail (approximately 1%) of the importance score distribution for further analysis (see figures 6.8 to 6.10 for importance score distribution). However, I acknowledge that a different cutoff could be used, leading to different results, and hence the lack of a definitive feature importance threshold for LIME is problematic.

The first thing to note about the results of LIME as opposed to occlusion is that there is a much more noticeable difference between the results on obese discovery set 1 and obese discovery set 2, as well as on healthy weight discovery set 1 and healthy weight discovery set 2, suggesting that the method is less robust to data variation and may not generalise as easily.

Gene Name	Feature Importance Score	Prior Relevant Associations
FTO	0.00104	BMI, lean body mass, body weight
LINC01122	0.00072	BMI
MSRA	0.00072	BMI, body weight
SOX5	0.00071	BMI, body weight
CSMD1	0.00071	BMI
CADM2	0.00070	BMI, body weight, body fat %
ADCY9	0.00070	BMI, body weight, body surface area
NRXN1	0.00069	BMI
AGBL4	0.00069	BMI, body weight, body fat %
CPNE4	0.00069	BMI
PDE4B	0.00069	BMI, body weight, body fat %
SPTB	0.00069	BMI
DLG2	0.00069	BMI, body weight, fat body mass
BDNF-AS	0.00069	BMI, body weight
ADGRB3	0.00069	BMI, body weight
ARNTL	0.00069	none
RBFOX1	0.00069	BMI, body weight
GIPR	0.00069	BMI, body weight, body fat %, body surface area
CRTC1	0.00069	BMI, body fat %, body weight

Table 6.6: Test Set 1, Obese individuals. Top twenty genes as discovered by LIME, their assigned importance score (corresponding to the weight of their corresponding feature in the trained linear model), and relevant prior associations. Biological associations sourced from genecards.org.

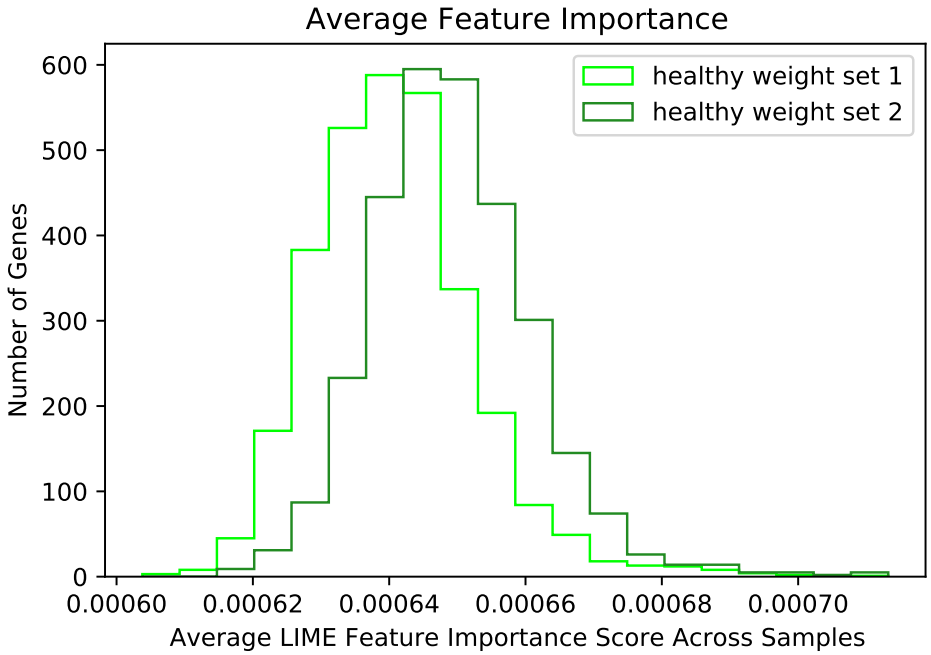


Figure 6.8: Histogram showing the distribution of feature importance scores for gene features in the two healthy weight dicovery sets, as derived by LIME.

Gene Name	Feature Importance Score	Prior Relevant Associations
FTO	0.00105	BMI, lean body mass, body weight
LINC01122	0.00073	BMI
NRXN1	0.00073	BMI
CPNE4	0.00071	BMI
CSMD1	0.00070	BMI
DLG2	0.00070	BMI, body weight, fat body mass
MSRA	0.00070	BMI, body weight
DPF3	0.00070	BMI
RBFOX1	0.00070	BMI, body weight
TRPM3	0.00070	BMI
PDE4B	0.00070	BMI, body weight, body fat %
AGAP1	0.00070	BMI, body weight
MAP2K5	0.00070	BMI
NLGN1	0.00070	BMI, body fat %, body weight
PTPRD	0.00070	BMI, lean body mass
LINGO2	0.00069	BMI
SGCZ	0.00069	BMI
CTNNA2	0.00069	BMI, body weight
ZFPM2	0.00069	BMI
C19orf48	0.00069	none

Table 6.7: Test Set 2, Obese individuals. Top twenty genes as discovered by LIME, their assigned importance score (corresponding to the weight of their corresponding feature in the trained linear model), and relevant prior associations. Biological associations sourced from genecards.org.

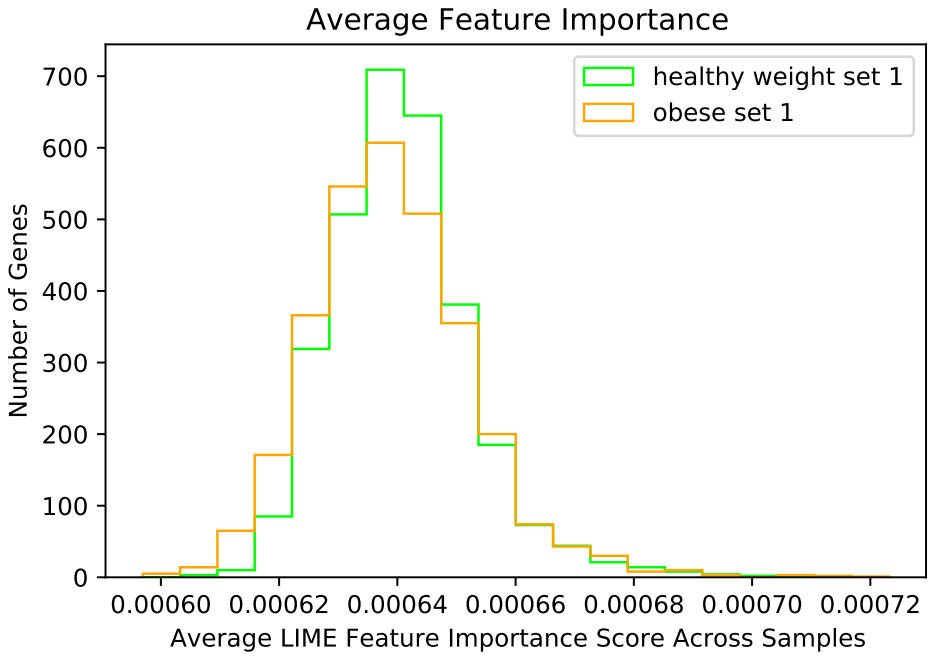


Figure 6.9: Histogram showing the contrasting distribution of feature importance scores for gene features in the obese and healthy weight dicovery sets, as derived by LIME.

Gene Name	Feature Importance Score	Prior Relevant Associations
FTO	0.00102	BMI, lean body mass, body weight
NRXN1	0.00071	BMI
TENM2	0.00070	BMI
CSMD1	0.00070	BMI
ZFPM2	0.00070	BMI
MSRA	0.00069	BMI, body weight
RBFOX1	0.00069	BMI, body weight
SGCZ	0.00069	BMI
CTNNA2	0.00069	BMI, body weight
LINGO2	0.00069	BMI
FLT3	0.00069	BMI
LINC01122	0.00069	BMI
SEMA6D	0.00069	BMI, body weight
CADM2	0.00069	BMI, body fat %, body weight
L3MBTL4	0.00069	none
MAP2K5	0.00069	BMI
ADCY9	0.00068	BMI, body weight, body surface area
ROBO2	0.00068	BMI
GIPR	0.00068	BMI, body weight, body fat %, body surface area
SOX5	0.00068	BMI, body weight

Table 6.8: Test Set 1, healthy weight and underweight individuals. Top twenty genes as discovered by LIME, their assigned importance score (corresponding to the weight of their corresponding feature in the trained linear model), and relevant prior associations. Biological associations sourced from genecards.org.

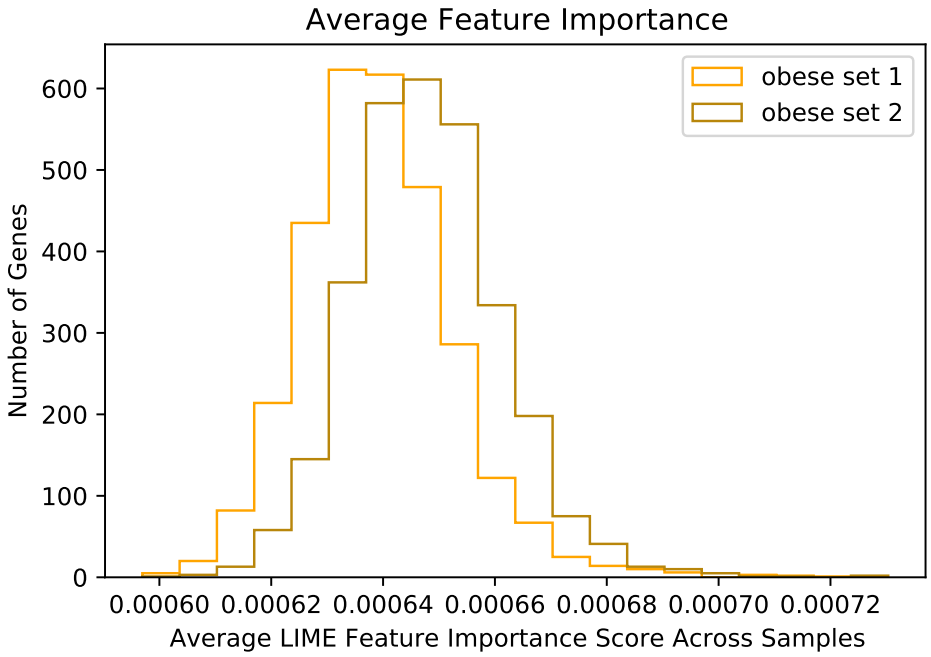


Figure 6.10: Histogram showing the distribution of feature importance scores for gene features in the two obese dicovery sets, as derived by LIME.

Gene Name	Feature Importance Score	Prior Relevant Associations
FTO	0.00103	BMI, lean body mass, body weight
AGAP1	0.00071	BMI, body weight
LINGO2	0.00071	BMI
NRXN1	0.00071	BMI
NLGN1	0.00071	BMI, body weight, body fat %
MSRA	0.00071	BMI, body weight
BNC2	0.00071	BMI, body fat %, body SA, body fat distribution
CPNE4	0.00070	BMI
SGCZ	0.00070	BMI
LINC01122	0.00070	BMI
NTM	0.00070	BMI, body fat %
ADGRB3	0.00070	BMI, body weight
PPP1R3B-DT	0.00070	BMI, body fat %
TCF7L2	0.00070	BMI, body fat %, body weight
MAP2K5	0.00070	BMI
CSMD1	0.00069	BMI
HRH1	0.00069	BMI, body weight, body SA, fat & lean body mass
KSR2	0.00069	BMI, body weight
ADCY9	0.00069	BMI, body weight, body surface area
DLG2	0.00069	BMI, body weight, fat body mass

Table 6.9: Test Set 2, healthy weight and underweight individuals. Top twenty genes as discovered by LIME, their assigned importance score (corresponding to the weight of their corresponding feature in the trained linear model), and relevant prior associations. Biological associations sourced from genecards.org.

There is also more difference between the results of the overweight and healthy weight sets, but it is not clear that this is due to a real effect and not just a lack of stability across different samples. The much higher agreement between subsets for feature ablation suggests that this is a methodological issue rather than a case of genuine significant differences in gene importance across different samples. This is potentially unsurprising as LIME is more prone to the accidental generation of highly unrealistic genotypes; since I used [0,0] as a baseline it is possible that LIME's perturbation function would occasionally generate a genotype that was mostly zeros, leading to unpredictable behaviour in the deep learning model and less deterministic outcomes overall.

In obese discovery set 2, the linear model scored almost twice as many genes above my determined "highly important threshold". Similar results were seen in the obese and healthy weight discovery sets, and as with the feature ablation results, there did not seem to be a clear difference

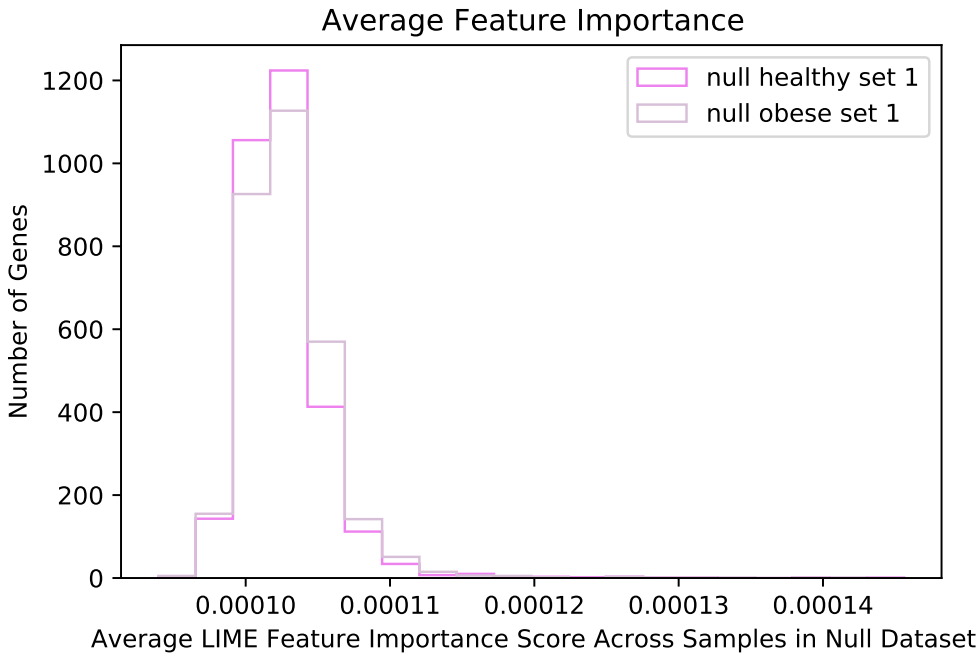


Figure 6.11: Histogram showing the distribution of feature importance scores for gene features in the obese and healthy weight discovery sets, as derived by LIME performed on a network trained on permuted phenotype data to produce a null distribution.

in which genes were prioritised for obese vs healthy samples. This suggests that there may be very significant differences in obesity aetiology at the individual level, especially where interaction is involved, such that the method becomes highly sensitive to the individuals it trains on. A large enough dataset would, however, help to smooth out such fluctuations.

It is also interesting to note that the distribution of gene importance differs greatly for the two methods. From single gene feature ablation we can see that the vast majority of genes are of low importance when omitted from prediction and have mean effects near zero, and the distribution has only a small tail of important genes, whereas with LIME the distribution is more normal, with most genes being considered of "moderate" importance and tails on either end of highly important and highly unimportant genes. There is also only a very small difference between the most and least important genes as prioritised by LIME. This could possibly be due to the fact that LIME considers combinations of genes where single gene feature ablation only considers one at a time; most genes are not important when considered alone but the all genes are at least somewhat influential in conjunction with other genes, something we might expect to

see in a polygenic phenotype like BMI. However, this property somewhat limits the utility of the results of gene-feature LIME on polygenic phenotypes.

Like the single gene feature ablation method, LIME singles out FTO as a gene of standout importance in every discovery set, validating that it is detecting similar information from the network as feature ablation. As you can see from tables 6.6 and 6.7, there is a high overlap of genes selected by LIME as important and those selected by ablation, and many of the genes selected by LIME are already well backed up by literature as being associated with BMI. However, unlike single gene feature ablation (and all other state of the art methods for identifying important genes from GWAS data), LIME considers genes not in isolation but in randomly generated sets, meaning that it is able to identify genes which do not have a significant effect on the phenotype when investigated in isolation but may have strong effects in conjunction with other genes. Of particular interest, then, are genes which are *not* prioritised by the occlusion method but nonetheless are flagged as important by LIME as these may be promising candidates for interaction analysis. Of particular interest within this subgroup are LIME-only important genes which have no prior association with BMI in literature, as these may have slipped through the net of traditional methods. One such candidate gene is ARNTL, which is above the importance threshold in both obese discovery sets for LIME, but achieves a comparatively lower 0.048 mean BMI effect from single gene feature ablation (ranking it 61st) when tested with ablation alone, and to our knowledge has not been found to be associated with BMI in the past. ARNTL has been identified as integral to the mechanism of circadian rhythm in mammals [223, 299], and appears to influence other disorders with a neurobehavioural component [181, 228, 222, 118, 18]. There is also evidence to suggest that the circadian system plays a role in the development of obesity [145, 89, 272], so this could be an interesting result. However, due to the variation in LIME's results across dataset partitions, it is possible that results like this may just be random.

Chapter 7

Gene-Gene Interaction in BMI

The purpose of this chapter is to examine different ways in which the model interpretation methods discussed in chapter 6 can be extended to search for evidence of gene-gene interactions. The ability to extract evidence of gene-gene interactions from trained phenotype prediction models could greatly streamline the process of interaction discovery and improve the utility of deep learning models as genetic discovery tools.

The chapter is divided into two main sections. The first section gives an overview of different techniques that could be used to detect and verify gene-gene interactions in this dataset, including detailing the novel method proposed in this PhD project. The second section explores the results of the gene-gene interaction method performed and provides further analysis to help glean as much biologically useful information as possible from the trained deep neural network.

7.1 Detection of Interactions

Of greater interest than individually important genes is how these genes may interact, something that is almost impossible to detect via state of the art methods such as GWAS or PRS. Interactions between genes can indicate involved biological or developmental processes, offering directions for future research or insights for new treatment or diagnostic options. The grouping of SNPs into biologically interesting groups for interpretation can be extended beyond single genes; to probe the model for interpretations, we will now demonstrate that it is possible to design gene

pair features to test for evidence of statistical epistasis. For this stage, as with the single gene analysis step, it is sensible to select a few representative samples for each subgroup, for which the predictive performance of the original network is very good, suggesting the network has learned a particularly useful representation of these people's genomic data; this reduces computational complexity as well as increasing the likelihood that the detected interactions will be meaningful.

7.1.1 Statistical Interaction Analysis

Before interaction analysis can occur it is necessary to define what we mean by a genetic interaction in a statistical sense. From a statistical point of view, the simplest definition of pairwise interaction would be any deviation from a linear additive model describing how two predictors x_1 and x_2 predict a phenotype y [83]. The simplest such linear additive model would be linear regression, in which a quantitative outcome y is modelled as a function of predictor x using the formula $y(x) = \beta x + c$ where c is a constant bias term. This can be extended to include n predictors in the format of

$$y(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{7.1}$$

It is assumed that each predictor variable has a linear relationship with y such that a unit increase in x_i would result in a β_i increase in y. Equation 6.1 can be extended to account for interactions between an arbitrary number of predictors using the Taylor series expansion:

$$y(x_1, ..., x_n) = \beta_0 + \sum_i \beta_i x_i + \sum_{i<j} \beta_{i,j} x_i x_j + \sum_{i<j<k} \beta_{i,j,k} x_i x_j x_k + ... \tag{7.2}$$

It is clear that modelling interactions parametrically using this equation would be computationally intractable, so often the simplification is made of discarding interactions of order 3 and higher and limiting consideration to pairwise effects, giving:

$$y(x_1, ..., x_n) \approx \beta_0 + \sum_i \beta_i x_i + \sum_{i<j} \beta_{i,j} x_i x_j \tag{7.3}$$

It is worth noting that even within the limitation of only considering pairwise interactions, equation 7.3 is parametric in that it assumes that the interaction effect is multiplicative. This

would cause issues in the case where there are interaction effects in the absence of main effects, as the interaction term would go to zero even if $\beta_{i,j}$ was nonzero. A general equation to quantify a specific pairwise interaction between features x_1 and x_2, which does not make any parametric assumptions about the nature of interaction, would thus be:

$$y(x_1,x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \sigma(x_1,x_2) \tag{7.4}$$

where σ is an unknown function which may include any number of higher-order terms, and the null hypothesis $H_0 : \sigma(x_1,x_2) = 0$, indicating that the relationship between the two variables and the output is perfectly linear and additive. The goal of a simple pairwise interaction method is thus to determine whether the value of $\sigma(x_1,x_2)$ is nonzero.

Note that while equation 7.4 is derived for genes interacting from the Taylor Series expansion, it is very similar to the definition of epistasis between loci for quantitative traits seen in [57], which defines the linear model $y_{i,j} = \mu + \alpha_i + \beta_j$ as the effect of two loci in absence of any epistatic effect.

7.1.2 Linear Regression with Interaction Terms

Quite a large number of different statistical methods exist for detecting interactions, either between individual markers or between groups of markers, each with varying degrees of effectiveness, and different advantages and drawbacks [111]. However, most of these statistical methods are based on some kind of comparison of an interaction-related statistic between two populations, which means they cannot easily be applied to continous data [58]. To our knowledge, the only method for continuous data that is in common use involves performing linear regression on the markers of interest with respect to the phenotype of interest, using a model that contains higher order multiplicative terms or even specifically designed interaction terms [182]. Searching over all possible pairs for interactions with this method is sometimes prohibitively computationally expensive [201], so there are several strategies which may be employed to reduce computational burden, for example applying a thresholding step to individual loci before searching for pairs (though this increases the likelihood of missing loci involved in interactions

in the absence of main effects).

A major drawback of this method is that it is highly parametric, with the model limited to detecting interactions that fit the terms designed for them. This means that real interactions may be missed altogether if simple multiplicative terms are used, and while using more sophisticated terms to detect known types of interactions has the advantage of the results showing exactly what's going on, they require considerable labour and expertise to design and may *still* miss signals of interactions that don't fit these parameters. Also, while exhaustive pairwise search can be feasible using this method with sufficient resources [182], the computational complexity climbs exponentially if more loci are considered, limiting the possibility of testing groups of interacting loci.

7.1.3 Multi-gene Feature Ablation

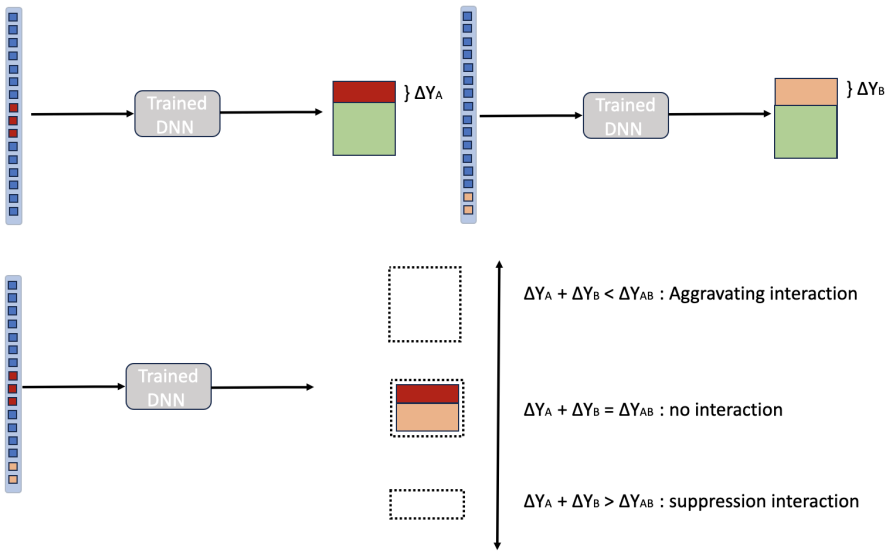


Figure 7.1: Illustration of the multi gene feature ablation process for a gene pair in a single sample.

As discussed in the previous section, the ablation of features can be done individually or with any combination of features in groups. This leads to an opportunity to efficiently analyse pairs or larger groups of genes for interaction effects. It also has the benefit that it is detecting pairs using "reasoning" that the model has learned based on the patterns it has detected in the training data, as opposed to just blindly comparing a test statistic. This means that it should be somewhat more

robust to dataset variation.

As with single gene feature ablation, we will be analysing interested in the resulting difference in prediction from perturbing feature i, again defined as:

$$\Delta Y_i(X_k) = Y(X_k) - Y_i'(X_k) = f(X_k;\mathbb{F}) - f(X_k;\mathbb{F}\setminus\{i\}) \tag{7.5}$$

However, this time we are not interested in main gene effects but in the interaction effect of the two genes. If there is no interaction between genes i and j, we would expect that the contribution of the two genes together for a given sample X_k would equal the sum of the individual contributions, such that:

$$f(X_k,\mathbb{F}) - f(X_k,\mathbb{F}\setminus\{i,j\}) = f(X_k,\mathbb{F}) - f(X_k,\mathbb{F}\setminus\{i\}) + f(X_k,\mathbb{F}) - f(X_k,\mathbb{F}\setminus\{j\}) \tag{7.6}$$

or,

$$\Delta Y_{i,j}(X_k) = \Delta Y_i(X_k) + \Delta Y_j(X_k) \tag{7.7}$$

We can then define the interaction term $\sigma^f_{i,j}$ as:

$$\sigma^f_{i,j}(X_k) = \Delta Y_{i,j}(X_k) - \Delta Y_i(X_k) - \Delta Y_j(X_k) = Y(X_k) - Y_i'(X_k) - Y_j'(X_k) + Y_{i,j}'(X_k) \tag{7.8}$$

which can be conceived of as the effect on the value of Y for sample X_k resulting from the difference in the individual and joint effects of i and j, or the difference in Y for sample X_k due to interaction between i and j. No parametric assumptions are made about the nature of this interaction, but a nonzero difference in the individual and joint effects suggests an interaction of some description. A $\sigma^f_{i,j}(X_k)$ value of zero corresponds with the truth of the null hypothesis, namely no interaction between features i and j for sample X_k, but as with single gene feature ablation, we need a null distribution for comparison in order to reject the null hypothesis.

As discussed, it can be difficult to choose candidate pairs to analyse. Genes which are identified as important by LIME but not by single-gene feature ablation are potentially good candidates for interaction effects; however, genes with a strong main effect may also have interaction effects, so

if the computational burden is not too high, it is best to test exhaustively. Fortunately feature ablation is efficient, so it was possible for us to exhaustive ablate every gene with every other gene in the set; if the mean result of perturbing both genes together is notably different than the result of perturbing each separately, this flags a potential interaction being detected by the network. Pair interactions can be built upon by further feature ablation with the pair and all other genes to look for three-way interactions, and so on. This offers a quick and efficient way to mine large amounts of data for potential pairs to analyse using more computational expensive traditional methods.

```
Function DIFFERENCE(sample):
                calculate prediction difference as a result of perturbation
FOR gene 1 in gene set:
    FOR gene 2 in (gene set - gene 1):
        FOR sample in sample set:
            DO set gene 1 related variants to [0,0] baseline
            DO Function DIFFERENCE(sample) ← difference 1
            DO set gene 2 related variants to [0,0] baseline
            DO Function DIFFERENCE(sample) ← difference 2
            DO set genes 1 & 2 related variants to [0,0] baseline
            DO Function DIFFERENCE(sample) ← difference 3
            DO calculate |difference 3 - (difference 1 + difference 2)|
        DO calculate mean absolute pair difference

```

Figure 7.2: Algorithmic representation of the gene pair feature ablation process.

7.1.4 Generating a Null Interaction Distribution

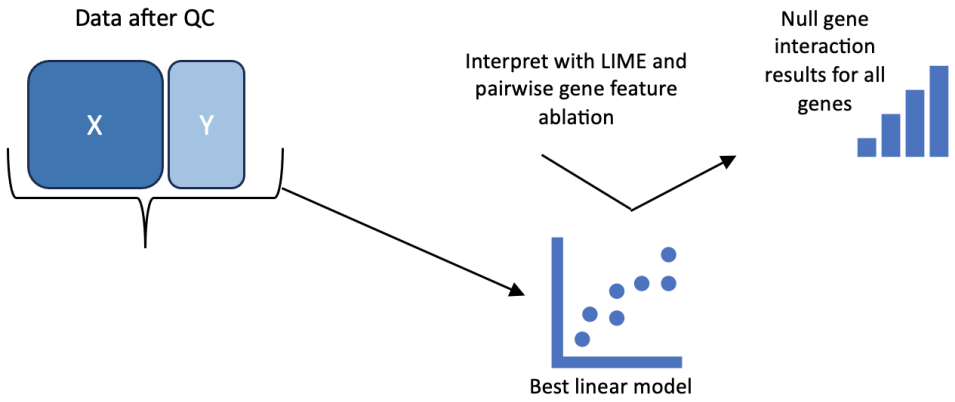


Figure 7.3: Illustration of the procedure to generate a null distribution for multi-gene feature ablation to use as a baseline for gene-gene interaction.

As with main effects, quantifying significance of interactions is difficult if we don't have some examples of spurious deviations from the linear additive model that may arise even in the case of the null hypothesis being true. The procedure for generating a null distribution for interactions is similar to the permutation procedure described in section 5.1.1, except that now the null hypothesis is that there is no *interaction* between genes of interest i and j, which may or may not be individually important; more concretely, the null hypothesis is that for genes i and j the term $\sigma^f_{i,j}(X_k)$ as defined in equation 6.6 is zero. Since neural networks are not entirely deterministic processes and the model is not perfect, there may be some deviation in the prediction results when different features are ablated that is the result of artefacts in the model or the data rather than true interaction in the relationship between genotype and phenotype. The goal of the null hypothesis distribution is thus to determine what nonzero value is the minimum sufficient to reject the null hypothesis.

Since we are testing interactions, in order to get a null distribution we need to ensure we are probing a model of the data in which no interactions are present. However, to ensure a good comparison it is important that this data still contains the majority of the *main* effects, to simulate a situation in which genes may still be important individually but there is no interaction to detect, i.e. $y(x_1,...,x_n) \approx \beta_0 + \sum_i \beta_i x_i$. Fortunately we already have a linear regression model trained on the data with no multiplicative terms; this performs almost as well as the DNN, suggesting it has learned most of the main effects, but it is unable to model the interactions between features by design. Thus the results of permuting and testing the linear model should contain the main effects and any artefacts resulting from said main effects, but no true interaction signals, providing a baseline against which to compare a potential interaction signal. This procedure is not perfect, since the neural network itself may create spurious interaction artefacts during training that would not arise in a linear model. Nonetheless, the linear model derived interaction distribution is a good basic test for possible interactions, without the need to generate synthetic data with the interactions removed.

7.2 Interaction Results

7.2.1 Null Distribution

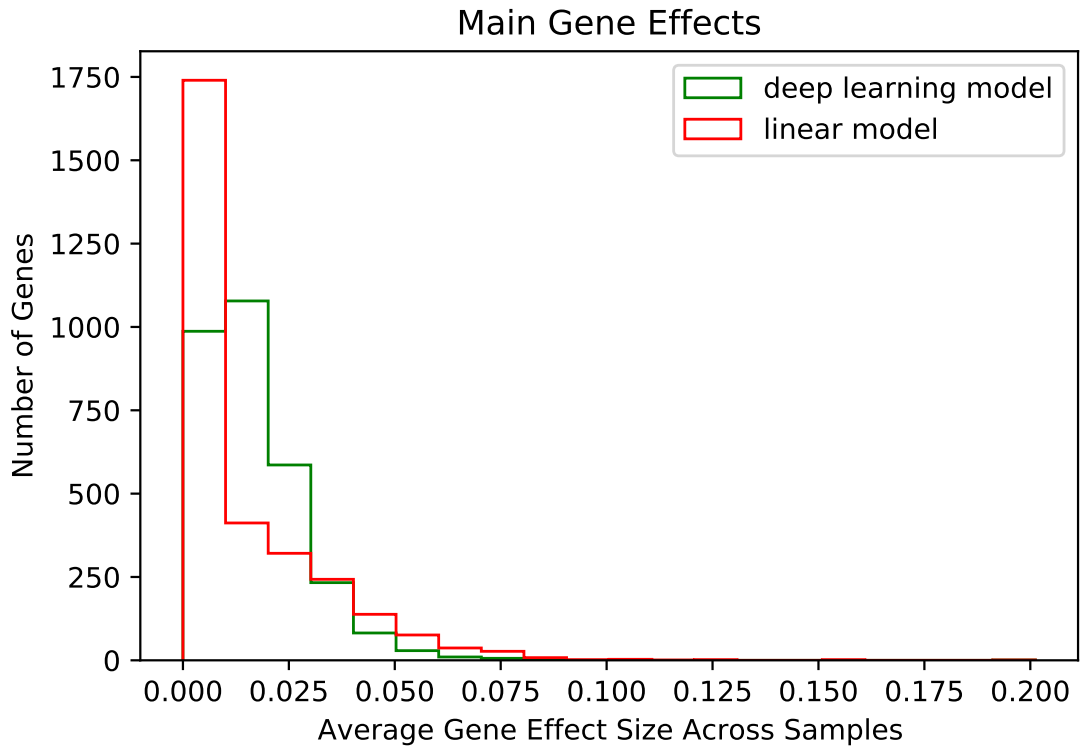


Figure 7.4: The distribution of main gene effect results for the deep learning model (green) vs the linear regression model (red).

When probed for main effects using single gene ablation, the linear regression model showed a fairly similar distribution to the deep learning model, as can be seen in figure 7.4. We can see that while the linear model groups a larger proportion of genes into the near-zero effect category and misses some moderate effect genes, it captures a large proportion of the larger single-gene effects and produces a main effect distribution that is not dissimilar to that of the deep learning model, demonstrating that it adequately learned to model most of the main effects. Like the deep learning model, it found a standout main effect for the FTO gene and yielded a very similar mean absolute BMI difference as a result of ablating this gene, at 0.198 (contrasting with 0.201 for the deep learning model). This suggests that it is a reasonable model of the main gene effects in absence of interaction.

Across pairs studied using both the linear and deep learning models, the interaction effects for most genes were small. There were a small number of strong effects ($\sim 10^{-2}$) in the pairwise ablation results from the linear model which corresponded to genes with a number of shared SNPs in the dataset, such as sense-antisense pairs or genes with separately labelled promoters, as well as genes with a "second degree overlap", meaning they both shared SNPs with a mutual third gene. Overlapping variants in the dataset will cause a deviation from the linear additive model without being evidence of real interaction, so it makes sense that ablation of the linear model would be able to indicate them. Otherwise, the linear model did not detect any non-zero interactions (see figure 7.5), suggesting that the false positive "overlap interactions" were the only signals present in the null distribution. This means we can be fairly confident that all nonzer interactions in the deep learning model that do not share either a first or second degree SNP overlap originate from some kind of authentic signal in the data that has been detected by the deep learning model, as opposed to being an artefact of the ablation method. The greatest non-interlapping interaction value in the linear model results was $1 * 10^{-4}$. We considered all interaction results greater than this to be potentially real results, and considered anything greater than $1 * 10^{-3}$ to be convincing evidence of detected interaction.

7.2.2 Deep Learning Model Pair Ablation Results

We performed an exhaustive search of all combinations of genes, which meant testing 3.64 million pairs, around 40% of which showed no interaction. 140 pairs showed especially convincing evidence of interaction with results of greater than $1 * 10^{-3}$, which is an order of magnitude greater than the largest interaction detected by the linear model. This is roughly 0.003% of all interactions tested, suggesting that pairwise gene interactions of noticeable effect size are something of a rarity for this phenotype in this cohort. A comparison between the pair results of the linear and deep learning models can be seen in figure 7.5. If we zoom in on the detected effects from the deep learning model (figures 7.6 and 7.7) we can see that interaction results become increasingly rare as they get higher. A 0.001 pair result corresponds to a modest main effect, with 95% of single genes having a measured main effect at this value or greater. This suggests that on the whole single gene effects may be more of a driver of BMI in this cohort than interaction effects.

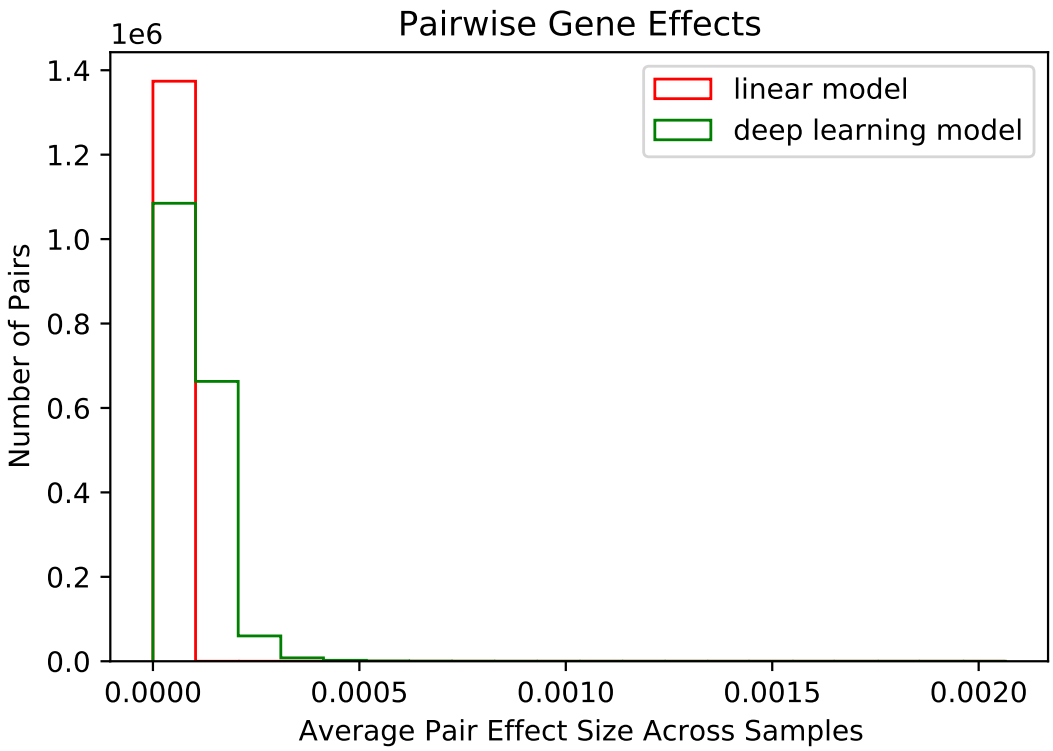


Figure 7.5: Pair ablation results of the linear model contrasted with the deep learning model.

Since the number of pairs with interaction results greater than seen in the null distribution was large (more than half of pairs investigated), we focussed further analysis on the top 140 "convincing interaction" pairs. The top 20 pair results are shown in the table 7.1. "Convincing interaction" pairs were filtered for both first and second degree overlap (to ensure they neither directly shared SNPs nor both shared SNPs with an unrelated third gene) by comparing the sets of overlapping genes for the two genes in the pair and excluding pairs where there was any set intersect.

Running this analysis was a lot more computationally intensive than the single gene feature ablation, so it was not feasible to run the analysis for every pair twice on different sample sets to test for reproducibility, like we did for the single gene features. However, in order to make sure that results were not unduly skewed by the choice of samples, we made an additional benchmarking set of the same size out of other samples in the test set that had not passed the MSE filter, and exhaustively ran the analysis for FTO and every other gene in the gene set

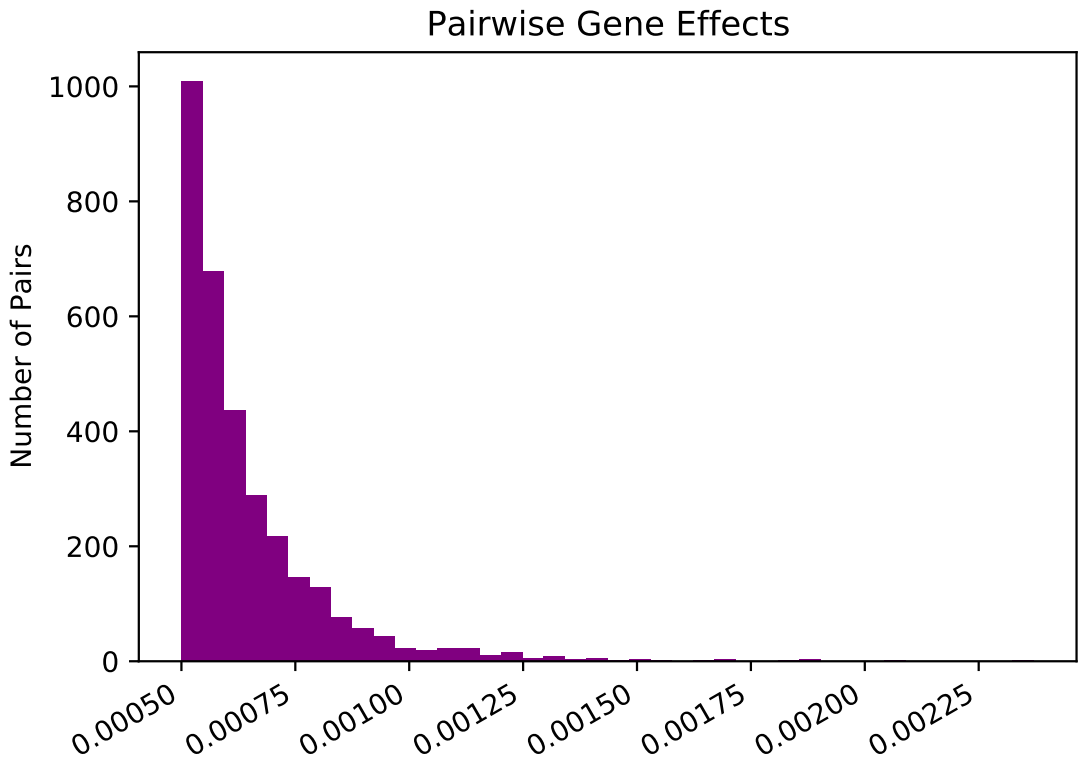


Figure 7.6: The distribution of mean absolute BMI differences resulting from the ablation of 3.64 million pairs of genes in the dataset. This histogram shows the distribution of results of 0.0005 or greater, corresponding to the top 3238 interactions.

on this group of samples. The results of this are shown in comparison with the results of the MSE-filtered sample set in figure 7.8. The figure shows an extremely high level of agreement between the different sample groups, showing both that the method is robust across different data and also that the MSE filtering does not alter results, which is interesting as one would expect that the model would only be able to make accurate interpretations on samples it had accurately learned to represent. This suggests that possibly the model has in fact learned a relatively good representation of the common variant basis of BMI for all the samples, and that the poor prediction performance on some samples could potentially be due to confounding environmental or rare gene effects that do not show up in the data but do influence the phenotype.

Most of the genes involved in the most strongly interacting pairs are familiar from the top main effects results, and perhaps unsurprisingly FTO is clearly a main player. As can be seen in table 7.1, the pair effects tend to be only a small fraction of the main effect strength, and there are

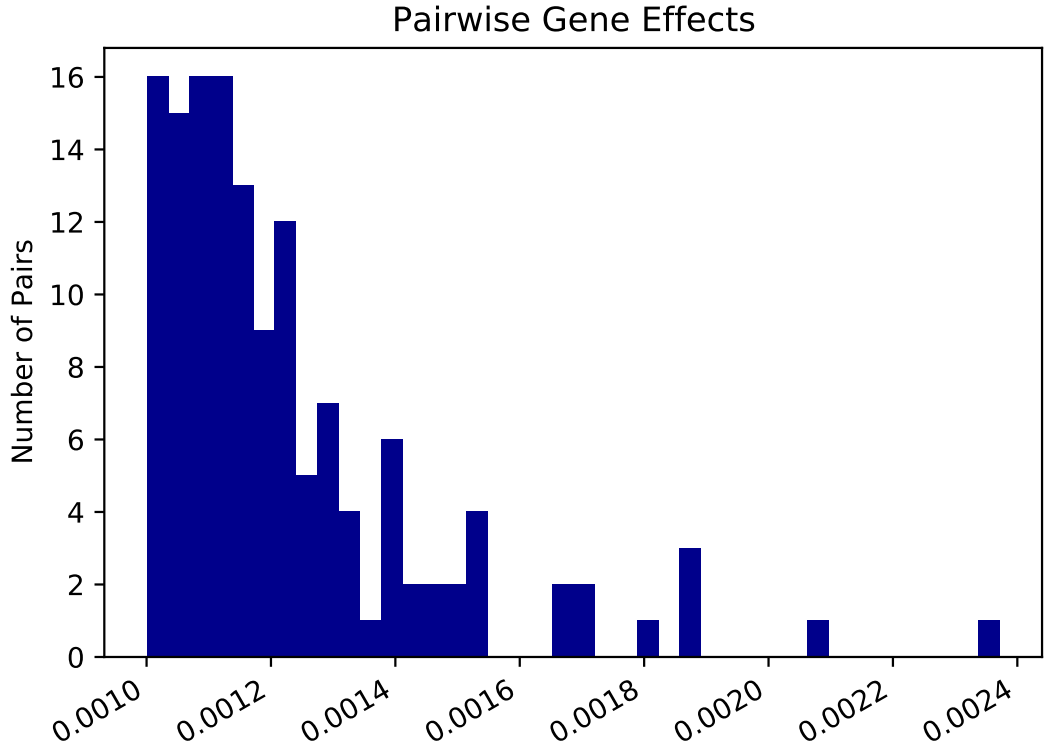


Figure 7.7: This histogram shows the distribution of interaction results greater than 0.001, which we defined as convincing evidence for detection of real interaction.

Gene Pair	Pair Effect	Individual Effects	Interaction Difference	Percentage of Main Effects
FTO, BDNF-AS	0.2361	0.2182, 0.0884	0.00243	1%, 3%
FTO, AGBL4	0.2250	0.2182, 0.0599	0.00212	1% , 3.5%
FTO, INPP4B	0.2202	0.2182, 0.0510	0.00193	0.8%, 4%
FTO, LINC01122	0.2354	0.2182, 0.0858	0.0020	0.9%, 2%
FTO, CPNE4	0.2340	0.2182, 0.0823	0.00195	0.8%, 2.3%
FTO, MSRA	0.2262	0.2128, 0.07967	0.00188	0.8%, 2%
FTO, NRXN1	0.2305	0.2182, 0.0823	0.00180	0.8%, 2%
FTO, SGCZ	0.2267	0.2182, 0.0741	0.00182	0.8%, 2%
NRXN1, CADM2	0.100	0.0823, 0.0578	0.00171	2%, 3%
FTO, PTPRD	0.2242	0.2182, 0.0581	0.00176	0.8%, 3%
FTO, ZBTB20	0.2308	0.2182, 0.064	0.00161	0.7%,2.5%
FTO, TRPM3	0.2249	0.2182, 0.0535	0.00159	0.7%, 3%
BDNF-AS, AGBL4	0.1003	0.0884, 0.0599	0.00156	2%, 3%
FTO, CADM2	0.2233	0.2182, 0.0578	0.00157	0.7%, 3%
FTO, LINGO2	0.2285	0.2182, 0.0735	0.00154	0.7%, 2%
FTO, CSMD1	0.2344	0.2182, 0.0783	0.00156	0.7%, 2%
FTO, PPP1R3B-DT	0.2304	0.2182, 0.0624	0.00154	0.7%, 2.5%
FTO, CCDC171	0.2249	0.2182, 0.0535	0.00151	0.6%, 3%
FTO, AGAP1	0.2230	0.2182, 0.0654	0.00154	0.7%, 2%

Table 7.1: Table showing the top 20 pair interactions found by pairwise gene ablation. The interaction difference corresponds to the average difference between the BMI difference that resulted from ablating both genes in the pair simultaneously and the result of adding the BMI differences of the individual genes.

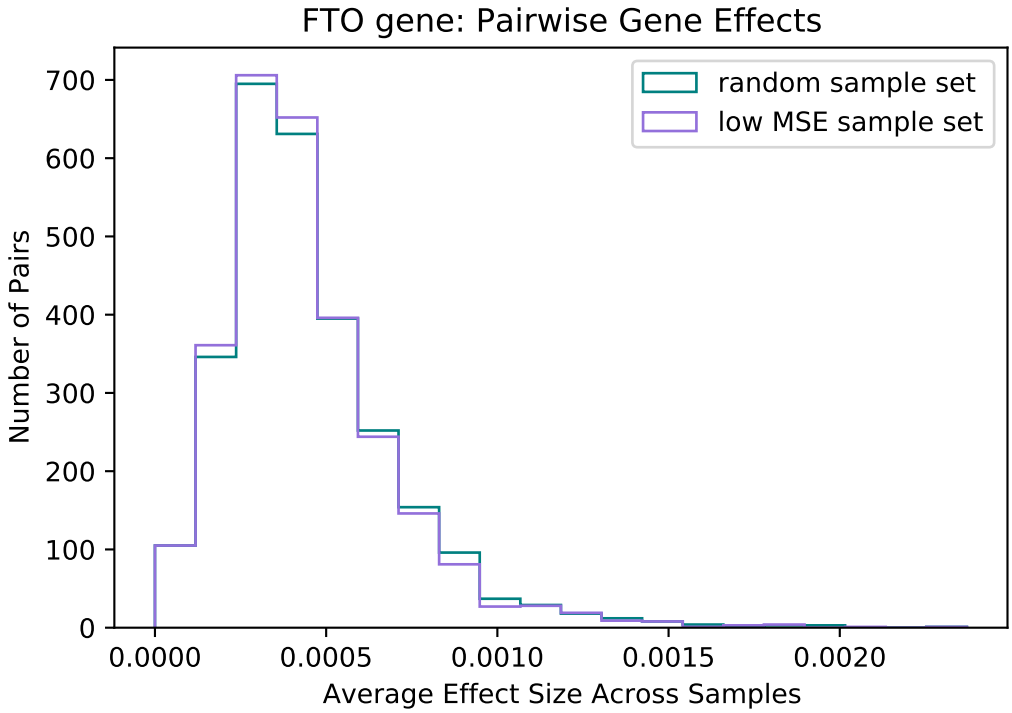


Figure 7.8: Histogram showing the results of perturbing FTO pairs on the MSE filtered sample set and on a set of randomly selected samples.

a lot of repeating genes in the top pairs, suggesting a possible network of interacting genes. Investigation of individual genes with numerous strong interactions can provide insights into the mechanism of genetic interactions that influence BMI. Some of these genes are examined in more detail below; information about prior associations and other biological knowledge is from GWAScatalog [278], GeneCards [280] and the NIH NCBI [265]. It is worth noting that there *is* a correlation between gene size and individual importance as discovered by our single-gene ablation method and this effects continues to be visible in the results of pairwise ablation; many large genes such as FTO have effects that are still very large even when corrected for their size, but for other large genes such as AGBL4 it is possible that their detected importance is an artefact rather than a true effect. Future studies may want to find a way to adjust for gene size in feature importance results.

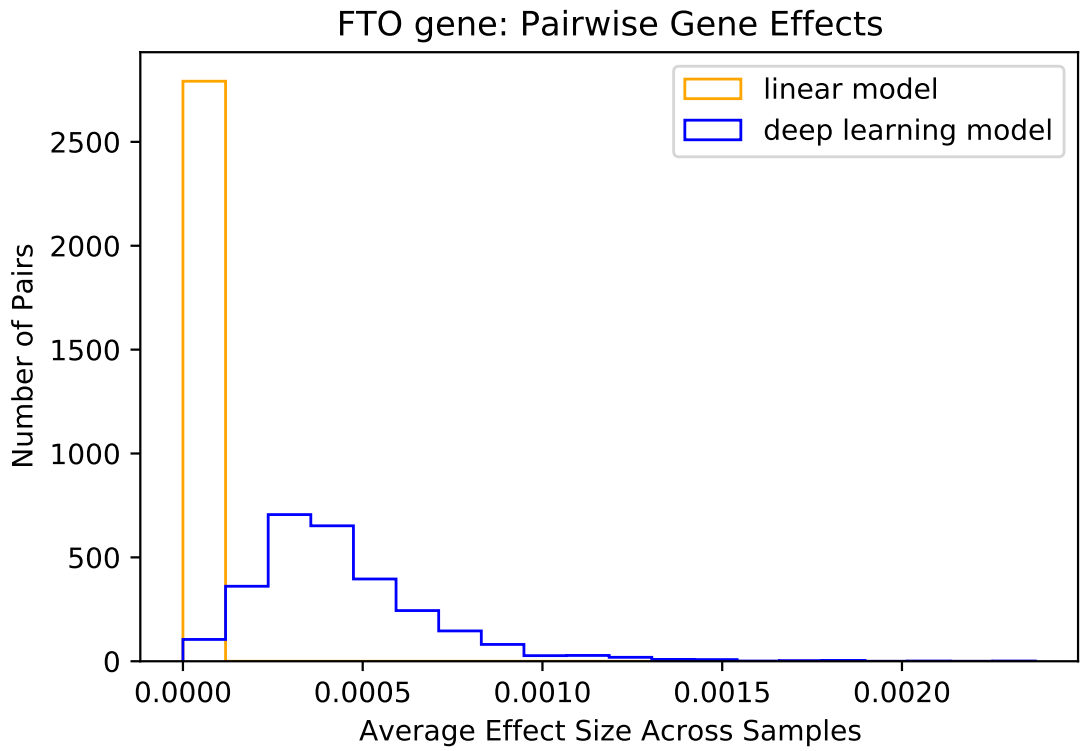


Figure 7.9: Pair ablation results of pairs involving the FTO gene, which displays convincing interactions with the majority of other genes in the dataset.

7.2.2.1 FTO

FTO appears prolifically throughout the top pairs, which is perhaps unsurprising given its importance individually. Plotting its pair scores shows an unexpected distribution in which it appears to interact with the vast majority of genes in the dataset. 96% of other genes in the dataset have non-zero interactions with FTO and FTO shows interactions of greater than 0.0005 with 30% of all genes. In fact, if we sum all of the pair effects associated with the FTO gene, it appears to indicate that it is possible that all of the FTO gene's effect derives from its interactions with other genes, which is an interesting possibility since as yet the exact mechanism of FTO on BMI is unknown [279]. This result raises the question as to whether FTO could act as a regulatory gene at the center of a BMI gene network.

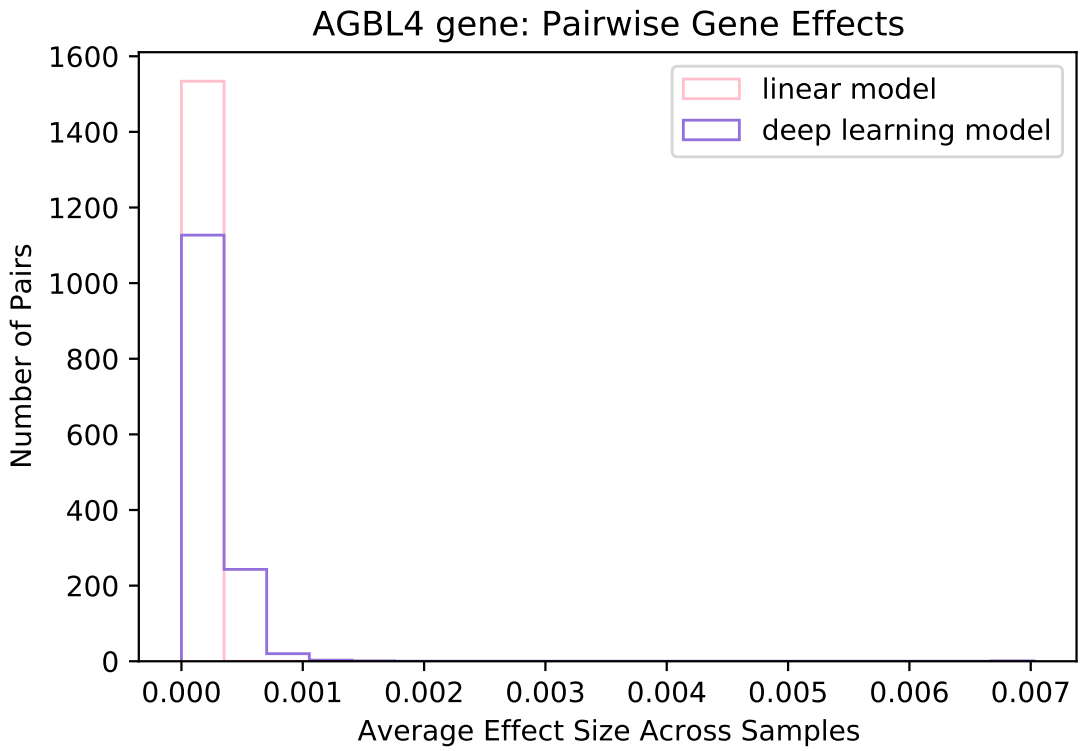


Figure 7.10: Pair ablation results of pairs involving the AGBL4 gene, which ranks 26th individually.

7.2.2.2 AGBL4

To compare with FTO, let's investigate another individually significant gene that appears to also be highly interactive, AGBL4. AGBL4 ranks 26th individually and is previously associated with BMI by a number of different studies. It does not appear to interact with the majority of other genes, suggesting this might be a unique property of FTO. AGBL4 shows interactions of greater than 0.0005 with only 6% of other genes.

7.2.2.3 XKR6

As discussed in previous sections, one of the biggest challenges to conventional interaction detection methods is the detection of interaction effects in the absence of main effects. Strong interactions are therefore especially interesting when one or both genes involved do not show a noticeable individual effect. Our pairwise feature ablation method does not include main effects in any part of the interaction calculation and is therefore capable of revealing several candidate interacting genes that do not show significant main effects on BMI either from single gene feature

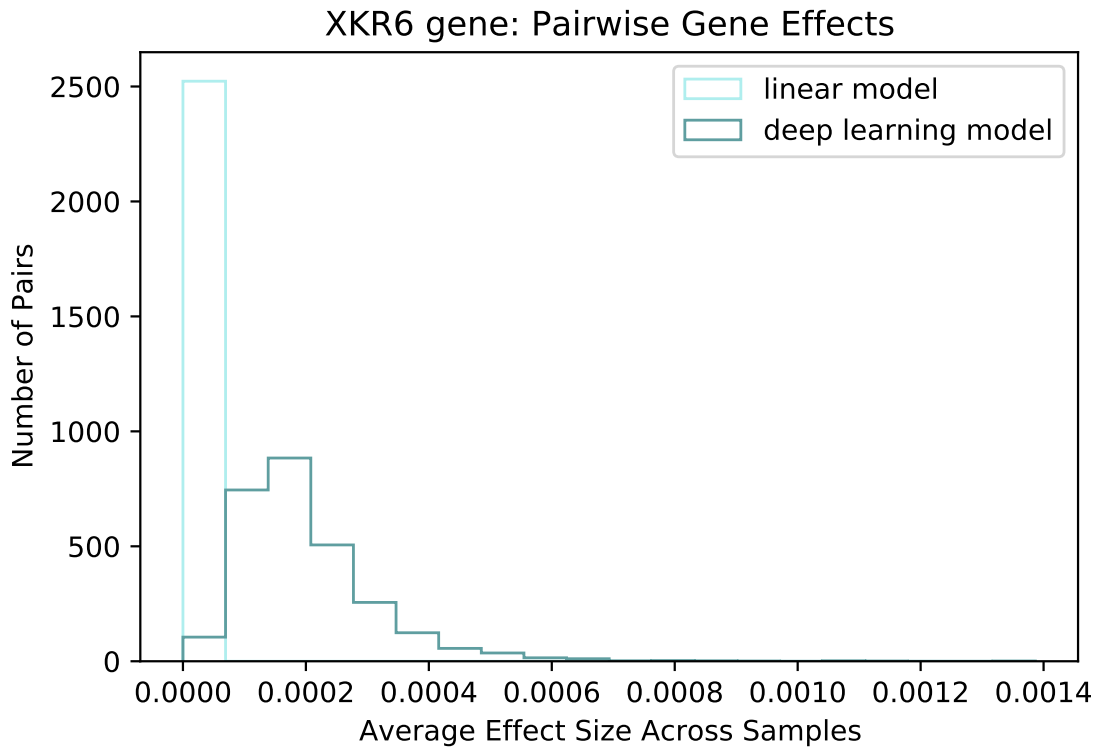


Figure 7.11: Pair ablation results of pairs involving the XKR6 gene, which ranks 618th individually.

ablation, LIME, or in previous literature.

One such gene is XKR6. Individually it does not stand out in single gene ablation, ranking 618th out of 3000 genes, with a mean BMI effect of 0.026, just barely over the highest null distribution result of 0.025. However, it is involved in 4 interactions of 0.001 or greater, and appears from figure 7.11 to be more interactive than one might have guessed. Its strongest interaction is with BDNF-AS, followed by FTO, CADM2 and CTBP2, which are all individually in the top 100 main effect genes. While this is not exactly a case of interaction in the absence of main effect, it is nonetheless an unexpected result for a highly interactive gene, and its interactions with BDNF-AS, FTO, CADM2 and CTBP2 account for 5%, 4%, 4% and 4% of its main effect respectively.

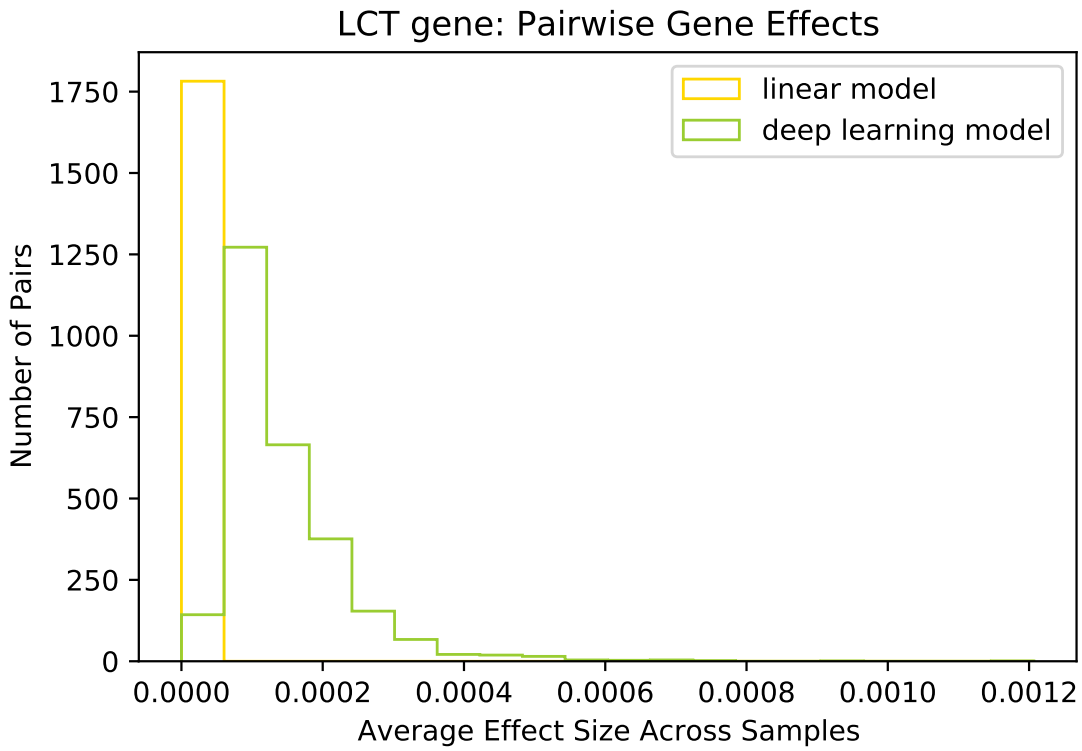


Figure 7.12: Pair ablation results of pairs involving the LCT gene, which ranks 619th individually.

7.2.2.4 LCT

Another unexpected interaction involves the LCT gene, which codes for the lactase enzyme, which in turn governs the persistence of lactose tolerance into adulthood [70]. To our knowledge this gene has not previously been associated with BMI, and our network finds it to be individually unimportant, ranking 619th with a mean BMI effect of 0.025, a result which the null distribution network demonstrates could be produced spuriously. However it has one strong >0.001 interaction with the gene CADM2, and like XKR6 appears slightly unexpectedly interactive. The implication of this interaction is potentially interesting; CADM2 is the second most strongly interactive gene after FTO, which again begs the question as to whether it is part of a BMI-influencing regulatory network.

7.2.2.5 C2

C2 is a protein coding gene in the complement system [207]. One SNP in it has a previous association with lean body mass via a single study, and our network did not find it to be

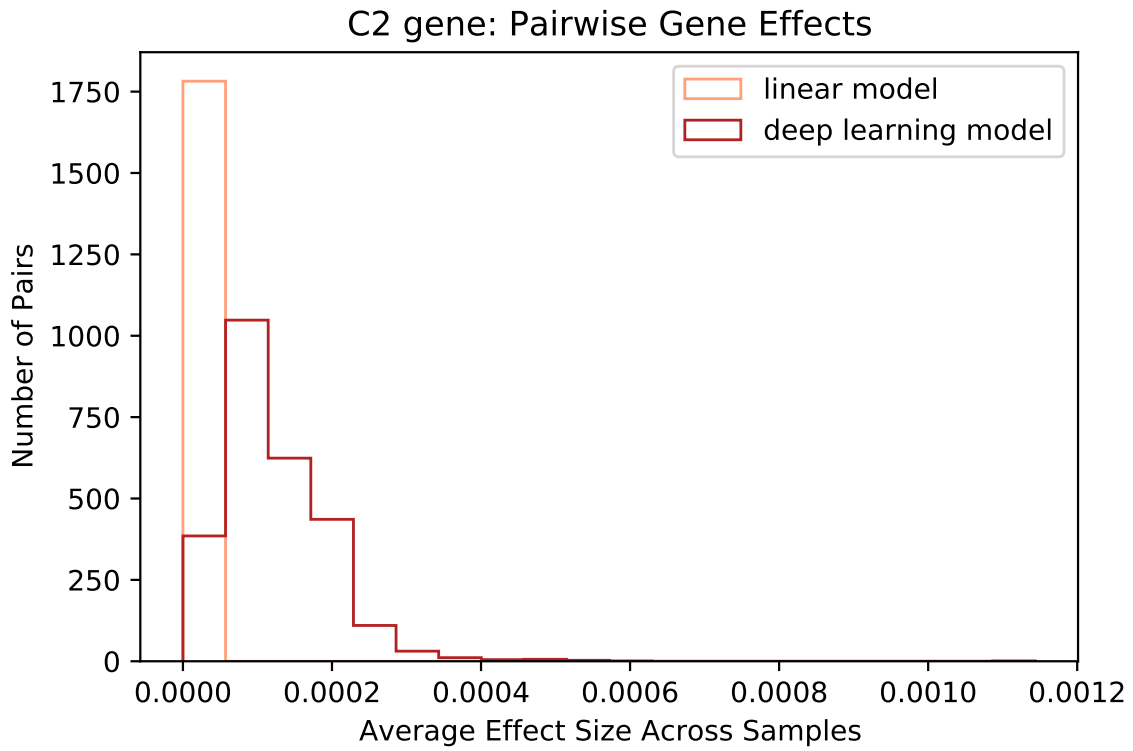


Figure 7.13: Pair ablation results of pairs involving the C2 gene, which ranks 1960th individually.

individually important, but its interaction with AGBL4 accounts for over 11% of its total phenotypic effect.

7.2.3 Gene Interaction Network

There are many genes that appear multiple times in pairs with interactions of greater than 0.001. This suggests the possible existence of a gene interaction network, which could be of great interest in terms of shedding light on the underlying disease aetiology. We can try to illustrate this interaction network by creating a graph of pairs, where nodes represent individual genes and edges represent the interactions between them. We can then rank nodes by degree: the number of connections they share with other nodes. If we constrain the problem by only considering pairs with >0.001 interaction results, we start to see certain genes appear as hubs at the center of strong interaction networks. The network of the 140 >0.001 interaction genes can be seen in figure 7.14.

	Number of Interactions >0.001	Individual Importance Rank	Chromosome
FTO	26	1	16
BDNF-AS	13	2	11
CADM2	9	35	3
AGBL4	9	26	1
PPP1R3B-DT	7	23	8
MSRA	7	6	8
TRPS1	5	75	8
XKR6	4	619	8
ZFHX3	3	70	16
CTBP2	3	67	10
CCDC171	3	48	9
STK24	3	50	13
NRXN1	3	4	2
SLIT2	3	49	4
ZFPM2	3	10	8
AGAP1	2	16	2
SSH2	2	41	17
SGMS1	2	123	10
MFHAS1	2	86	8
KCNQ5	2	92	6
SMG6	2	685	17
CPNE4	2	5	3
RPTOR	2	18	17
SGCZ	2	9	8
ZBTB20	2	17	3
PATJ	2	58	1
BNC2	2	31	9
DLG2	2	20	11

Table 7.2: Table showing the top 20 genes with the most strong (>0.001 difference from linear additive model values) interactions as established by pairwise feature ablation, and their importance rank as established by individual feature ablation.

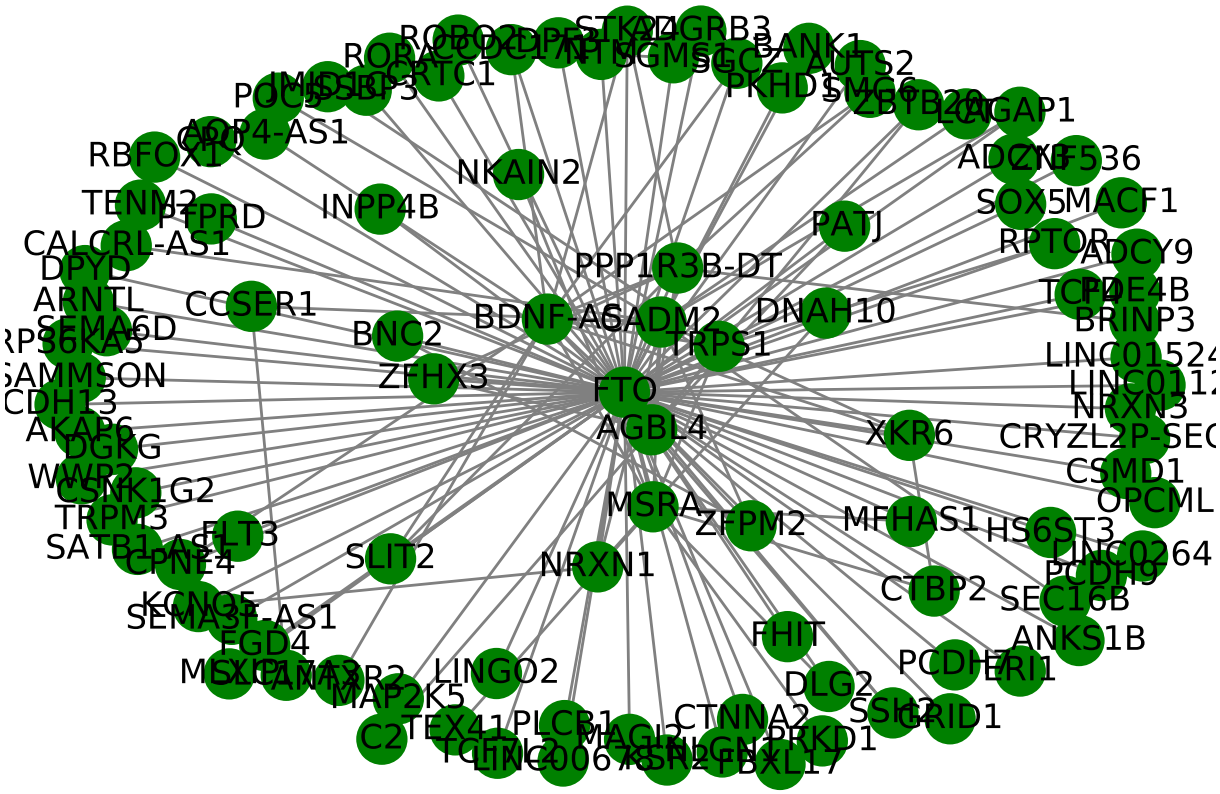


Figure 7.14: Representation of the network of genes with interactions greater than 0.001.

If we look at figure 7.14, we can see that there appears to be a central more complicated interaction network surrounded by a ring of genes with only a single connection to one gene in the network. This might indicate that the outer ring genes are not part of the regulatory network, but only acted on by it. The complex inner network is of particular interest, and we can zoom in on this by constraining the set to only include nodes with two or more connections, and re-drawing the graph (see figure 7.15). Some of the genes in this interactions are further examined in table 7.2. It is apparent that the individual importance of the genes (as determined by single gene ablation) is somewhat variable, and they are found all over the genome.

To our knowledge, most of these interactions have not been detected or reported on before. A notable exception to this is that an interaction network has been previously identified that involves FTO and ZBTB20 [247] via the NRTK2 signalling pathway. ZBTB20 also appears to be connected to AGBL4 via the enhancer GH01J050023 [281].

Since 0.001 is a fairly arbitrary choice of an importance cutoff, it may result in some missed

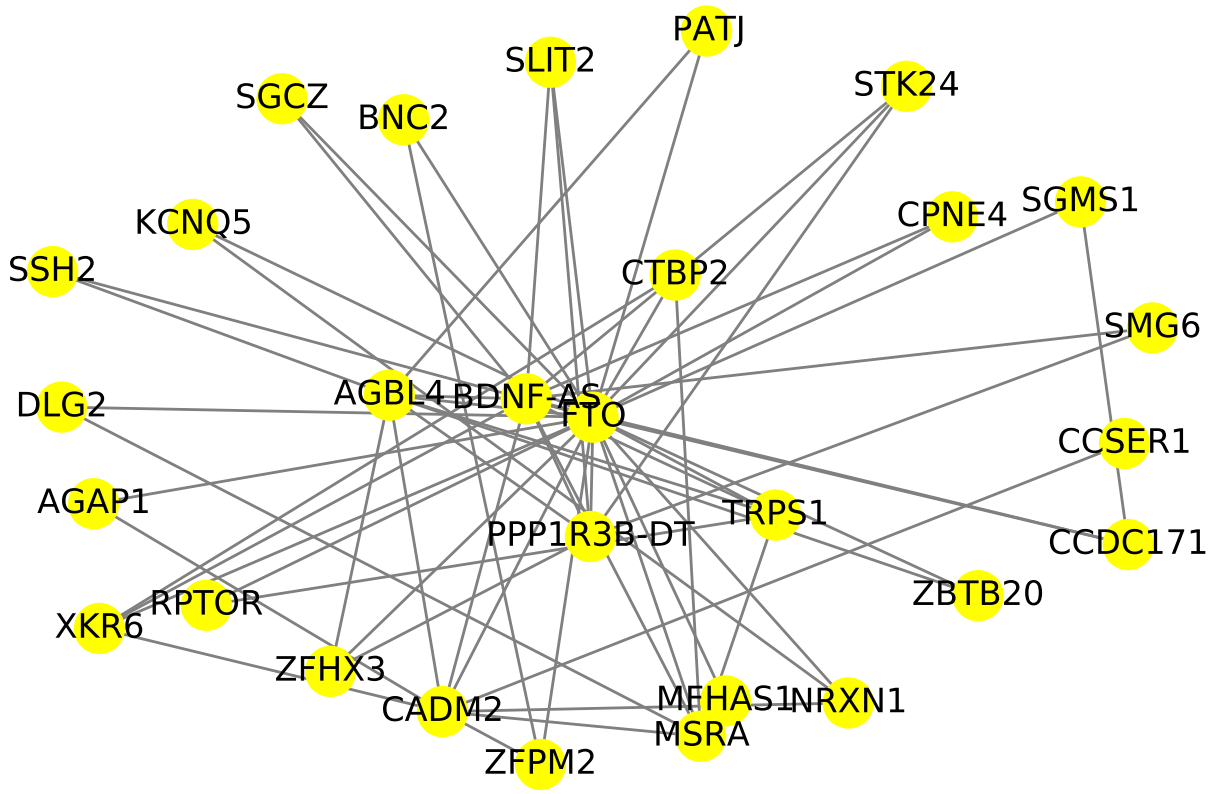


Figure 7.15: Representation of the network of genes with more than one interaction greater than 0.001, comprising a network of more interactive genes.

interactions. However, when we tried using 0.0005 as the interaction cutoff, and then prioritising nodes in the resulting highly connective 3238-node graph with 20 connections or more in order to look for the most interactive genes, the resulting set of highly interactive genes was identical to the top genes with the 0.001 cutoff, suggesting that the genes with the strongest interactions are also the ones with the most interactions in general.

A previous study [67] investigated gene-gene interactions in BMI using multifactor dimensionality reduction; unfortunately, due to data removed in our data filtering and SNP-gene association steps, there were no pairs of genes identified as interacting in their study for which both genes were represented in our dataset, meaning we were unable to test if our method had detected the same interactions. However, our set did contain the gene MAP2K5, which was found in one of their detected interactions, and we found a strong interaction (>0.001) between this gene and FTO. FTO was excluded from the MDR study, so this interaction would not have been detected. MAP2K5 also appeared as one of the highly interactive genes in our gene interaction

graph.

Overall, these results display an intriguing combination of known genetic effects and previously undocumented pathways for BMI. This outcome suggests that our proposed technique could offer a promising new tool for efficiently probing predictive models for genetic architectural information, potentially leading to insights on a variety of phenotypes. We leave further interpretation of our BMI results to geneticists.

Chapter 8

Conclusions and Future Work

The era of Genetic Big Data has provided unprecedented new opportunities for learning about the aetiology of complex phenotypes, which could potentially lead to better understanding of and the development of new treatments for common diseases. However, being able to take advantage of this data for scientific and medical research requires the development of appropriate tools. Deep learning models show promise in this area, but it is not yet clear how best to design or optimise them for genetic data and phenotype prediction tasks, and there is even less understanding of how to extract biologically useful information from trained models. In this thesis I have proposed several techniques for adapting deep learning models for SNP data and interpreting these models using biologically-informed features. This chapter summarises the main contributions of my research, and then discusses how this work fits in to the larger landscape of deep learning for genetics and the challenges still facing the field, and then provides some recommendations for the direction of future work.

8.1 Contributions

In this section we will discuss the novel contributions of this thesis, and evaluate their efficacy and discuss how they add to the existing corpus of knowledge in this domain.

1. **Novel Encodings for SNP data (Chapter 5)**

   We proposed a series of novel biologically-informed methods to encode data at the SNP-

level, as discussed in chapter 5. We then trialled our encoding methods in a variety of linear and deep learning models and contrasted their performance against the more traditional standard variable encoding as well as the non-genetics-specific one-hot variable encoding. Our encoding methods relied on a two dimensional vector representation of SNP space, parametrised either by REF/ALT alleles or by allele effect direction as discovered by GWAS. We also trialled two vector magnitudes for the heterozygous case, one which constrained the effect size to be the same for all three cases and one which considered the heterozygous case as having 50% magnitude along each of the axes, leading to a lower overall effect magnitude for heterozygosity.

We found our novel encoding methods provided a modest improvement in model performance across different types and configurations of models. In particular, all proposed novel encodings outperformed the standard variable encoding across the board. Performance of the one-hot variable encoding was comparable with our novel encodings for many models; however, since the novel encodings require only a two dimensional vector space instead of three, the resulting encoded datasets were much smaller, which meant they were far more efficient than the one-hot representation, leading to faster training times and allowing for use with larger datasets without encountering memory problems. We found that encodings where the effect size of the heterozygous genotype was slightly smaller than those of the homozygous genotypes appeared to lead to marginally better performance, and that including information about allele effect direction did not improve model performance but *did* increase model training time.

Overall our results indicate that there is potential utility in the design and use of specialised encoding methods for SNP data for deep learning models. Attempts to encode data in ways that better reflected the biological reality of the genotypes appeared to improve performance while allowing for a reduction in computational complexity in comparison to a categoric variable encoding. However, our observed improvements over the standard approach were limited and we were not able to pinpoint a definitive best encoding. Exactly which biological information should be included in a genetic data encoding and how this

would best be represented remains an open question.

2. **Evaluation of Deep Learning Models for Body Mass Index Prediction (Chapter 5)**

   In an effort to examine whether deep neural networks could automatically learn to accurately model BMI from genetic data, we evaluated a number of deep learning models trained on data from the UK Biobank, and contrasted these with linear baseline models trained on the same data. Since there is no consensus on the best network architecture or parameters for genotype-phenotype networks, we trialled a wide variety of model configurations using a gridsearch approach. Our analysis did not appear to uncover an optimal architecture for neural networks predicting phenotype from genotype data, and the performance differences between different activation functions, node architectures, optimisers, etc. were marginal. However, we broadly found that deep learning models outperformed linear models in phenotype prediction even when the deep learning model architecture was suboptimal, and that this performance difference increased with the number of inputs. We found that including a larger number of more weakly significant SNPs improved model performance for both deep learning and linear models, but moreso for the deep learning models. We encountered significant performance issues with deep learning models aside from feedforward fully connected networks (CNNs, Bayesian neural networks). Our result that including more SNPs improved prediction performance across different model types validates similar findings in numerous other studies [17, 128].

   Overall, we were not able to uncover an obviously optimal architecture for BMI prediction networks for genetic data. However, our results yielded several pieces of useful information about how to design models for this task, the main ones being that neural networks are a better choice than linear models, that a basic fully-connected feedforward neural network is a reasonable architectural starting point, and that including as many input SNPs as is computationally feasible is likely to lead to best performance.

3. **Biologically-Informed Feature Design for Deep Learning Model Interpretation (Chapter 6)**

While the field of deep learning is improving at a rapid pace, and many new methods to apply deep learning to genomic problems have been developed, the question of how to usefully interpret "black-box" models trained on genetic data remains an open one. We present an approach to augment existing deep learning model interpretation methods to be better suited to models trained on genomic data, allowing these methods to yield more biologically informative results. Our approach uses biological prior knowledge to design interpretable features that mimic real genetic phenomena such as genes or pathways, so that perturbing them yields intuitively comprehensible information about the aetiology of the phenotype, therefore making the model itself more useful as a research or clinical tool. Our method for the design of biologically-informed features can be used with any perturbation-based model interpretation method, but we found that feature ablation yielded results with the greatest intuitive value, that were more reproducible than those of other methods. We used this approach to probe our best performing network for its most influential genes, and found that these genes closely corresponded with existing findings about the aetiology of BMI.

This result demonstrated that our genetic feature ablation approach can be used to validate that a model has learned biologically accurate information even when prediction accuracy metrics are not sufficiently high to definitively suggest that an accurate representation of the relationship between independent and dependent variables has been modelled. We showed that feature ablation with biologically-informed interpretable gene features is efficient, produces results that are robust across dataset partitions and phenotypic subgroups, and produces interpretable values in the same scale as the original phenotype, making results easy to understand in the context of the wider prediction problem. Our approach is highly flexible in that any deep learning model can be used for the prediction stages, meaning this framework offers multiple avenues for augmentation, extension or integration with existing methods. It is also entirely nonparametric from start to finish, meaning that it can easily be applied in conjunction with predictive models with many input features operating in complex problem domains with limited prior knowledge. We demonstrated that even

a deep learning model with a relatively poor fit on the data was able to capture enough accurate information to make it a useful tool for probing phenotype aetiology.

4. **Pairwise Gene Interaction Detection Method (Chapter 7)**

   As discussed in chapter 7, detecting interactions between genes from genetic data can be extremely methodologically and computationally challenging. We present an effective and comprehensible framework for detecting gene-gene interactions in the aetiology of any phenotype for which a moderately successful predictive deep learning model has been developed. Our proposed framework extends our single-gene feature ablation method to look for deviations from a linear additive model when pairs of genes are ablated.

   We showed that once ablation with single gene features has validated that a model has learned real biological information, ablation can then be used with combinations of features to search for feature interactions in reasonable computational time. We used this method to search for gene interactions influencing BMI in data from the UK Biobank, and then used our most prominent detected gene pairs to construct a gene interaction map, validating some existing interaction results and suggesting many possible new interactions. In the case of our dataset and models, the fact that the performance improvement of the deep learning model over the linear model was limited suggests that gene-gene interactions probably didn't play a very significant role in the aetiology of this phenotype in this particular population; this was confirmed by the small size of even the most prominent interactions discovered by the pair interpretation step. However, our method is flexible and model-agnostic and thus could easily be applied to other datasets and phenotypes in which gene-gene interactions may prove significant. Even in the case of our small detected interactions, there were interesting implications for BMI aetiology, showing that the method has promise for illuminating potential disease pathways.

8.2 Discussion and Recommendations for Future Work

In this section we will outline some of the major challenges faced in this project and discuss the methodological limitations of our approach. We will also make recommendations for ways in which our findings and approaches could inform the direction of or be incorporated into future research.

8.2.1 Limitations of Genomic Methods

There are a few methodological shortcomings associated with the earlier genomics stages of my pipeline, which could potentially be overcome in future work by the use of new technologies. We still used GWAS as the first filtering step for the input SNPs, meaning that there is the possibility that our exclusion of "unimportant" SNPs may have removed SNPs that were actually involved in interactions, and their removal hampered our efforts to detect a BMI interaction network. It is also possible that the inclusion of these "unimportant" would have improved prediction performance, particularly for the deep learning models. Teams with access to more computing resources could potentially look at wider networks with a greater number of inputs, but it would also be worth considering alternatives to GWAS for dimensionality reduction, such as newer ML-driven SNP prioritisation tools [16], or even unsupervised methods like autoencoders [237].

8.2.2 Methodological Issues with Genetic Ablation Methods

There are a number of decisions to be made in the design of a feature ablation procedure, and choices relating to these have an influence on the success of the method. A poorly designed feature ablation procedure can lead to unrealistic inputs and a lack of robustness, or results that do not actually conform to the true model importance of the features being tested. In this study we were not able to exhaustively test different feature ablation procedures, and this section discusses how our approach could be modified in future to potentially improve outcomes.

One design decision that can have significant effects on the efficacy of the protocol is in the choice of baseline value. Currently it is not easy to infer effect direction from our method due to the fact that when SNPs are knocked out as part of gene ablation, the resulting direction of

BMI difference will depend on what SNP was there in the first place. This also means that in pairwise gene feature ablation it is not possible to indicate what kind of interaction is being detected between genes, only that there has been a deviation from the linear model. A different choice of baseline values could potentially overcome this; commonly the baseline of choice for ablation methods is the dataset mean value of the feature, but since genotype values are discreet and not continuous the mean would provide a highly unrealistic input. An alternative could be to use the modal genotype instead, which would mean that the ablation results would indicate the effect direction of the minor allele, but could be overall less informative (since the modal genotype may actually be the effective one when considered in conjunction with other variants, treating it as the "no effect" default is risky). Another option could be to keep the zero baseline but to consider alleles, instead of SNPs, as the knock-out-able features, which could show a more granular picture of effect direction and in particular make it easier to see the effects of dominance interactions. However, this would become quite complex when one is trying to ablate entire genes or gene systems, and is complicated by the lack of stranding information in most SNP datasets.

Another issue we encountered, discussed slightly in chapter 6, was to do with SNP labelling and overlapping genes. Since we used KEGG to group SNPs into gene features based on genes they were associated with in the database, we had a large minority of SNPs that were simultaneously grouped into multiple explainable features. This meant that some duplicate features had to be removed, leading to some gene features becoming stand-ins for multiple genes, and also we had some gene features with very high SNP overlap, such as the case of BDNF-AS containing all but two of BDNF's SNPs. This meant that in single gene interpretation it could be difficult to determine which gene was the actual effect gene due to data leakage between gene features, and in the pair ablation stage it meant that we were unable to measure interactions between certain pairs, as the overlap meant they automatically deviated from the linear model. For example, BDNF and BDNF-AS showed a very high interaction score with a mean absolute difference of 0.0358, which could have corresponded to a real effect, but it is impossible to separate a real signal from the noise caused by the shared SNPs. To avoid this issue, a different, more definitive chromosomal location-based SNP clumping method could potentially have been used

which would have prevented SNPs from being in multiple features, or else SNPs could have been labelled in the same way but constrained to appear in only one gene feature (though it might be difficult to work out which one when several were available). Additionally, our gene feature construction method discarded the minority of SNPs which did not have any known associations with genes in KEGG, despite the fact that these SNPs might have been of interest. Again, grouping SNPs by region instead of gene could address this issue. As lack of clarity around best practice for SNP-gene labelling is a major confounder for GSA methods as well, this may be a broad obstacle for genetic interpretation methods in general. Also like GSA, we are limited in our interpretation by the current limits of KEGG, which will be true to some degree as long as features are designed with biological prior information. Our method was also intrinsically biased in favour of very large genes, and it could be difficult to separate true signal from artefacts caused by gene size. Future studies could adjust feature importance results to account for gene size, allowing for example the comparison of "per-SNP gene importance" instead of absolute gene importance.

A major pitfall of perturbation-based methods in general is that they are prone to problems when features are interdependent because impossible combinations can be generated by accident. These kinds of issues are potentially especially problematic with one-size perturbation methods like LIME; with the baseline I used it was possible to end up with majority-zero genotypes which will create extreme outlier samples that the network won't be able to perform well on, which may have contributed to the lack of reproducibility of LIME results across data subsets. While this was not an issue in our novel ablation method, it still may have generated unrealistic inputs by perturbing certain genes in isolation or in conjunction with other genes in a way that would never be seen in the real data due to factors such as linkage. Although different choice of baseline could go some way towards mitigating this, there would still be a risk of generating impossible allelic combinations that the model would by default not have been exposed to during training, leading to potentially uninformative and unexpected results from perturbations. In future this could possibly be circumvented by designing explainable features in a way that takes linkage disequilibrium into account, for example perturbing haplotype blocks instead of genes, or analysing explainable features for correlation in some other way and implementing requisite

co-ablation of correlated features.

8.2.3 Future Possibilities

In general, the flexibility of my genetic feature design protocol offers many possibilities for future work involving this approach. Aside from the greater number of possible groupings increasing the combinatorial search space, there is no associated computational cost with analysing higher-order interactions, and with a smaller input set or a filtering step it would be quite possible to search for three-, four- or ten-SNP interaction clusters. It also could be interesting to design interpretable features around for example known pathways or GO accession terms, allowing the possibility of searching for pathway involvement or even inter-pathway interactions.

A potential limitation of this project was the relatively poor performance of the neural network being interpreted. Since the entire pipeline is nonparametric and this interpretation method can be applied to any model capable of detecting nonlinear feature interactions, it would be trivial to integrate this with other contemporary methods aiming to produce better neural networks, such as those evolved using genetic algorithms [213, 298] or ones that use MDR for feature selection [203]. If a network with a clearer performance improvement over the linear baseline was developed and a strong interaction pair or cluster *was* discovered during the interpretation phase, it would be interesting to try redesigning the network with only the identified interacting features as inputs and seeing how this changed the performance of the phenotype prediction. This could give an insight into whether a phenotype is more polygenic with additional SNPs serving as noise, or whether the genetic background is still important in the phenotypic aetiology even in the presence of strong interactions.

Aside from this, the nonparametric nature of our proposed pipeline means there is no reason our interaction detection approach couldn't be applied on a model trained with both genetic features and multimodal features, offering the possibility of detecting gene-environment interactions as well as gene-gene interactions (subject to the suitable selection of a baseline environment value for perturbation). Since there is a growing body of evidence to suggest that gene-environment

interactions are a crucial part of the development of obesity [248], the ability to build interpretable models that include both genetic and environmental factors could significantly improve prediction performance, making them potentially useful in a clinical context.

Finally, recent research into the genetics of obesity has indicated that more extreme cases can be used to detect rare variants, which can themselves be used to shed light on genetic regions or biological processes involved in the more common, milder phenotype [310, 88, 263]. This could mean that models that consider rare variants alongside common ones could have increased prediction power, and their interpretation could yield more biological insights. Incorporating rare variants into traditional statistical methods such as GWAS or even PRS is challenging as the low prevalnce of such variants in any sample can make it difficult to validate results; these problems persist to a degree with neural networks, as the network is unlikely to learn much from a pattern it is only exposed to a handful of times. On the other hand, our interpretation method demonstrated that deep neural networks are able to automatically learn some relational chromosomal location information even from sparse SNP data. This suggests that if rare variants were inserted into the data in such a way that they were surrounded by related common variants (perhaps associated with the same gene), the network might be able to use the rare variant's surroundings to enhance its "understanding" of the effect of the rare variant despite the allelic imbalance. Simultaneously, the presence of the rare variant and the likely stronger phenotypic effect it confers could also enhance the model's "understanding" of the gene region. A trained common and rare variant model could then use our proposed interpretation methods to look for interactions between common variants and rare variants or between rare variants and genes.

8.3 Conclusion

In this thesis, I have explored the opportunities offered by interpretable Deep Learning models to shed light on the genetic basis of Body Mass Index. I have developed and tested a number of novel encoding methods for SNP data that will be used to train deep learning models, and tested a variety of model configurations for predicting BMI. I have also proposed and tested several novel approaches for model interpretation that are designed and tailored specifically

to DL models trained on genetic data. Experiments have shown that the proposed techniques, which are flexible and can be used with a variety of different DL models, can increase the utility of DL prediction models as biological discovery tools and allow for nuanced and efficient model probing in reasonable computational time, while avoiding many of the pitfalls of traditional statistical methods. We have also demonstrated that our approach can be used to shed light on the genetic architecture of BMI, and potentially other traits, in an exciting and novel way.

Understanding the genetic basis of complex heterogeneous phenotypes remains a major challenge, but one that, if solved, could provide revolutionary new scientific insights for the field as a whole. Many questions remain unanswered, and the rapid pace of technological advancement on both the genetic sequencing technology side and the Artificial Intelligence side means that new opportunities are being created all the time. In this project we have opened a number of avenues for future work involving interpretable DL models and genetic data, and hope that researchers will be able to use our insights to design ever more effective tools for understanding the genome.

Bibliography

[1] UK Biobank research ethics approval, Dec 2021.

[2] UK Biobank - about us, Aug 2022.

[3] R. Abdollahi-Arpanahi, D. Gianola, and F. Peñagaricano. Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genetics Selection Evolution*, 52(1), Feb 2020.

[4] J. U. Adams. Genetics: big hopes for big data. *Nature*, 527(7578):S108–S109, 2015.

[5] D. Albuquerque, C. Nóbrega, L. Manco, and C. Padez. The contribution of genetics and environment to obesity. *British medical bulletin*, 123(1):159–173, 2017.

[6] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838, Jul 2015.

[7] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1), Mar 2021.

[8] S.-i. Amari. Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5(4-5):185–196, 1993.

[9] C. A. Anderson, F. H. Pettersson, G. M. Clarke, L. R. Cardon, A. P. Morris, and K. T. Zondervan. Data quality control in genetic case-control association studies. *Nature Protocols*, 5(9):1564–1573, Jul 2010.

[10] H. Araki, C. Knapp, P. Tsai, and C. Print. Genesetdb: A comprehensive meta-database, statistical and visualisation framework for gene set analysis. *FEBS Open Bio*, 2(1):76–82, Jan 2012.

[11] R. A. Armstrong. Should pearson's correlation coefficient be avoided? *Ophthalmic and Physiological Optics*, 39(5):316–327, Aug 2019.

[12] E. A. Ashley. Towards precision medicine. *Nature Reviews Genetics*, 17(9):507–522, 2016.

[13] A. Badré, L. Zhang, W. Muchero, J. C. Reynolds, and C. Pan. Deep neural network improves the estimation of polygenic risk scores for breast cancer. *Journal of Human Genetics*, 66(4):359–369, Oct 2020.

[14] N. Barton, A. Etheridge, and A. Véber. The infinitesimal model: Definition, derivation, and implications. *Theoretical Population Biology*, 118:50–73, Dec 2017.

[15] H. Behravan, J. M. Hartikainen, M. Tengström, V. Kosma, and A. Mannermaa. Predicting breast cancer risk using interacting genetic and demographic factors and machine learning. *Scientific Reports*, 10(1), Jul 2020.

[16] H. Behravan, J. M. Hartikainen, M. Tengström, K. Pylkäs, R. Winqvist, V. Kosma, and A. Mannermaa. Machine learning identifies interacting genetic variants contributing to breast cancer risk: A case study in finnish cases and controls. *Scientific Reports*, 8(1), Sep 2018.

[17] P. Bellot, G. de los Campos, and M. Pérez-Enciso. Can deep learning improve genomic prediction of complex human traits? *Genetics*, 210(3):809–819, Aug 2018.

[18] T. Bering, M. B. Carstensen, G. Wörtwein, P. Weikop, and M. F. Rath. The circadian oscillator of the cerebral cortex: molecular, biochemical and behavioral effects of deleting the ARNTL clock gene in cortical neurons. *Cerebral Cortex*, 28(2):644–657, 2018.

[19] A. Bhat, P. R. Lucek, and J. Ott. Analysis of complex traits using neural networks. *Genetic Epidemiology*, 17(S1), Jan 1999.

[20] S. Billiard, V. Castric, and V. Llaurens. The integrative biology of genetic dominance. *Biological Reviews*, 96(6):2925–2942, 2021.

[21] E. G. Bochukova, N. Huang, J. Keogh, E. Henning, C. Purmann, K. Blaszczyk, S. Saeed, J. Hamilton-Shield, J. Clayton-Smith, S. O'Rahilly, et al. Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature*, 463(7281):666–670, 2010.

[22] E. G. Bochukova, K. Lawler, S. Croizier, J. M. Keogh, N. Patel, G. Strohbehn, K. K. Lo, J. Humphrey, A. Hokken-Koelega, L. Damen, et al. A transcriptomic signature of the hypothalamic response to fasting and BDNF deficiency in prader-willi syndrome. *Cell reports*, 22(13):3401–3408, 2018.

[23] V. Botta, G. Louppe, P. Geurts, and L. Wehenkel. Exploiting SNP correlations within random forest for genome-wide association studies. *PLoS ONE*, 9(4), Apr 2014.

[24] E. A. Boyle, Y. I. Li, and J. K. Pritchard. An expanded view of complex traits: From polygenic to omnigenic. *Cell*, 169(7):1177–1186, Jun 2017.

[25] A. P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[26] N. Brandes, O. Weissbrod, and M. Linial. Open problems in human trait genetics. *Genome Biology*, 23(1), Jun 2022.

[27] A. J. Brookes. The essence of SNPs. *Gene*, 234(2):177–186, 1999.

[28] J. D. Brown. The coefficient of determination, 2009.

[29] W. S. Bush, S. M. Dudek, and M. D. Ritchie. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. In *Biocomputing 2009*, pages 368–379. World Scientific, 2009.

[30] W. S. Bush and J. H. Moore. Chapter 11: Genome-wide association studies. *PLoS Computational Biology*, 8(12), Dec 2012.

[31] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, and et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, Oct 2018.

[32] M. P. Calus and J. Vandenplas. SNPrune: An efficient algorithm to prune large SNP array and sequence datasets based on high linkage disequilibrium. *Genetics Selection Evolution*, 50(1), Jun 2018.

[33] E. Cano-Gamez and G. Trynka. From GWAS to function: Using functional genomics to identify the mechanisms underlying complex diseases. *Frontiers in Genetics*, 11, May 2020.

[34] Carlborg and C. S. Haley. Epistasis: Too often neglected in complex trait studies? *Nature Reviews Genetics*, 5(8):618–625, Aug 2004.

[35] A. Cecile, J. W. Janssens, and M. J. Joyner. Polygenic risk scores that predict common diseases using millions of single nucleotide polymorphisms: Is more, better? *Clinical Chemistry*, 65(5):609–611, May 2019.

[36] M. E. Celebi and K. Aydin. *Unsupervised learning algorithms*, volume 9. Springer, 2016.

[37] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao, T. D. Kelley, D. Braines, M. Sensoy, C. J. Willis, and P. Gurram. Interpretability of deep learning models: A survey of results. In *2017 IEEE Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing, UIC-ATC*, pages 1–6, 2017.

[38] N. Chami, M. Preuss, R. W. Walker, A. Moscati, and R. J. Loos. The role of polygenic susceptibility to obesity among carriers of pathogenic mutations in MC4R in the UK Biobank population. *PLOS Medicine*, 17(7), Jul 2020.

[39] P. B. Chandrashekar, J. Wang, G. E. Hoffman, C. He, T. Jin, S. Alatkar, S. Khullar, J. Bendl, J. F. Fullard, P. Roussos, and D. Wang. Deepgami: Deep biologically guided auxiliary learning for multimodal integration and imputation to improve phenotype prediction. *bioRxiv*, 2022.

[40] N. Chatterjee, B. Wheeler, J. Sampson, P. Hartge, S. J. Chanock, and J.-H. Park. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nature Genetics*, 45(4):400–405, Mar 2013.

[41] A. Chattopadhyay and T.-P. Lu. Gene-gene interaction: the curse of dimensionality. *Annals of Translational Medicine*, 7(24), 2019.

[42] A. Cheerla and O. Gevaert. Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics*, 35(14):i446–i454, 07 2019.

[43] L. Chen, C. Cai, V. Chen, and X. Lu. Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model. *BMC Bioinformatics*, 17(S1), Jan 2016.

[44] S.-H. Chen, J. Sun, L. Dimitrov, A. R. Turner, T. S. Adams, D. A. Meyers, B.-L. Chang, S. L. Zheng, H. Grönberg, J. Xu, and et al. A support vector machine approach for detecting gene-gene interaction. *Genetic Epidemiology*, 32(2):152–167, Feb 2008.

[45] X. Chen, L. Wang, B. Hu, M. Guo, J. Barnard, and X. Zhu. Pathway-based analysis for genome-wide association studies using supervised principal components. *Genetic epidemiology*, 34(7):716–724, 2010.

[46] Y. Chen, Y. Li, R. Narayan, A. Subramanian, and X. Xie. Gene expression inference with deep learning. *Bioinformatics*, 32(12):1832–1839, 02 2016.

[47] J. M. Cheverud and E. J. Routman. Epistasis and its contribution to genetic variance components. *Genetics*, 139(3):1455–1461, Mar 1995.

[48] Y. Cho, M. Ritchie, J. Moore, J. Park, K.-U. Lee, H. Shin, H. Lee, and K. Park. Multifactor-dimensionality reduction shows a two-locus interaction associated with type 2 diabetes mellitus. *Diabetologia*, 47:549–554, 2004.

[49] S. W. Choi, T. S.-H. Mak, and P. F. O'Reilly. Tutorial: A guide to performing polygenic risk score analyses. *Nature Protocols*, 15(9):2759–2772, Jul 2020.

[50] H. Choquet and D. Meyre. Genetics of obesity: What have we learned? *Current Genomics*, 12(3):169–179, May 2011.

[51] D. M. Church, V. A. Schneider, T. Graves, K. Auger, F. Cunningham, N. Bouk, H.-C. Chen, R. Agarwala, W. M. McLaren, G. R. Ritchie, et al. Modernizing reference genome assemblies. *PLoS biology*, 9(7):e1001091, 2011.

[52] K. W. CHURCH. Word2vec. *Natural Language Engineering*, 23(1):155–162, 2017.

[53] M. Claussnitzer, J. H. Cho, R. Collins, N. J. Cox, E. T. Dermitzakis, M. E. Hurles, S. Kathiresan, E. E. Kenny, C. M. Lindgren, D. G. MacArthur, et al. A brief history of human disease genetics. *Nature*, 577(7789):179–189, 2020.

[54] C. S. Coffey, P. R. Hebert, M. D. Ritchie, H. M. Krumholz, J. M. Gaziano, P. M. Ridker, N. J. Brown, D. E. Vaughan, and J. H. Moore. An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene-gene interactions on risk of myocardial infarction: the importance of model validation. *BMC bioinformatics*, 5:1–10, 2004.

[55] A. Collins and N. Morton. Mapping a disease locus by allelic association. *Proceedings of the National Academy of Sciences*, 95(4):1741–1745, 1998.

[56] R. Collins. What makes UK Biobank special? *The Lancet*, 379(9822):1173–1174, 2012.

[57] H. J. Cordell. Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20):2463–2468, Oct 2002.

[58] H. J. Cordell. Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6):392–404, Jun 2009.

[59] M. Costanzo, A. Baryshnikova, J. Bellay, Y. Kim, E. D. Spear, C. S. Sevier, H. Ding, J. L. Koh, K. Toufighi, S. Mostafavi, and et al. The genetic landscape of a cell. *Science*, 327(5964):425–431, Jan 2010.

[60] D. J. Crouch and W. F. Bodmer. Polygenic inheritance, GWAS, polygenic risk scores, and the search for functional variants. *Proceedings of the National Academy of Sciences*, 117(32):18924–18933, Aug 2020.

[61] H. Cui, C. Zhou, X. Dai, Y. Liang, R. Paffenroth, and D. Korkin. Boosting gene expression clustering with system-wide biological information: A robust autoencoder approach. *BioRXiV*, Nov 2017.

[62] T. Cui, K. El Mekkaoui, J. Reinvall, A. S. Havulinna, P. Marttinen, and S. Kaski. Gene–gene interaction detection with deep learning. *Communications Biology*, 5(1), Nov 2022.

[63] R. Culverhouse, T. Klein, and W. Shannon. Detecting epistatic interactions contributing to quantitative traits. *Genetic Epidemiology*, 27(2):141–152, Sep 2004.

[64] R. Culverhouse, B. K. Suarez, J. Lin, and T. Reich. A perspective on epistasis: Limits of models displaying no main effect. *The American Journal of Human Genetics*, 70(2):461–471, Jan 2002.

[65] P. Cunningham, M. Cord, and S. J. Delany. *Supervised Learning*, pages 21–49. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[66] D. CURTIS, B. V. NORTH, and P. C. SHAM. Use of an artificial neural network to detect association between a disease and multiple marker genotypes. *Annals of Human Genetics*, 65(1):95–107, Jan 2001.

[67] R. De, S. S. Verma, F. Drenos, E. R. Holzinger, M. V. Holmes, M. A. Hall, D. R. Crosslin, D. S. Carrell, H. Hakonarson, G. Jarvik, and et al. Identifying gene-gene interactions that are highly associated with body mass index using quantitative multifactor dimensionality reduction (QMDR). *BioData Mining*, 8(1), Dec 2015.

[68] C. A. de Leeuw, B. M. Neale, T. Heskes, and D. Posthuma. The statistical properties of gene-set analysis. *Nature Reviews Genetics*, 17(6):353–364, Apr 2016.

[69] B. Devlin, N. Risch, and K. Roeder. Disequilibrium mapping: Composite likelihood for pairwise disequilibrium. *Genomics*, 36(1):1–16, 1996.

[70] L. Diekmann, K. Pfeiffer, and H. Y. Naim. Congenital lactose intolerance is triggered by severe mutations on both alleles of the lactase gene. *BMC Gastroenterology*, 15(1), Mar 2015.

[71] H. U. Dike, Y. Zhou, K. K. Deveerasetty, and Q. Wu. Unsupervised learning based on artificial neural network: A review. In *2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)*, pages 322–327, 2018.

[72] M. E. Doche, E. G. Bochukova, H.-W. Su, L. R. Pearce, J. M. Keogh, E. Henning, J. M. Cline, A. Dale, T. Cheetham, I. Barroso, et al. Human SH2B1 mutations are associated with maladaptive behaviors and obesity. *The Journal of clinical investigation*, 122(12):4732–4736, 2012.

[73] Y. Dodge. *Mean Squared Error*, pages 337–339. Springer New York, New York, NY, 2008.

[74] S. Draghici, P. Khatri, A. L. Tarca, K. Amin, A. Done, C. Voichita, C. Georgescu, and R. Romero. A systems biology approach for pathway level analysis. *Genome Research*, 17(10):1537–1545, Sep 2007.

[75] F. Dudbridge. Power and predictive accuracy of polygenic risk scores. *PLoS Genetics*, 9(3), Mar 2013.

[76] L. Duncan, H. Shen, B. Gelaye, J. Meijsen, K. Ressler, M. Feldman, R. Peterson, and B. Domingue. Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications*, 10(1), Jul 2019.

[77] S. D'Silva, S. Chakraborty, and B. Kahali. Concurrent outcomes from multiple approaches of epistasis analysis for human body mass index associated loci provide insights into obesity biology. *Scientific Reports*, 12(1), May 2022.

[78] C. E. Elks, M. den Hoed, J. H. Zhao, S. J. Sharp, N. J. Wareham, R. J. Loos, and K. K. Ong. Variability in the heritability of body mass index: A systematic review and meta-regression. *Frontiers in Endocrinology*, 3, Feb 2012.

[79] P. Esposito. Blitz - bayesian layers in torch zoo (a bayesian deep learing library for torch). `https://github.com/piEsposito/blitz-bayesian-deep-learning/`, 2020.

[80] J. Euesden, C. M. Lewis, and P. F. O'Reilly. PRSICE: Polygenic risk score software. *Bioinformatics*, 31(9):1466–1468, Dec 2014.

[81] I. Farooqi and S. O'rahilly. Genetic factors in human obesity. *Obesity reviews*, 8:37, 2007.

[82] K. A. Fawcett and I. Barroso. The genetics of obesity: FTO leads the way. *Trends in Genetics*, 26(6):266–274, Jun 2010.

[83] R. A. Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1919.

[84] E. B. Ford. The theory of dominance. *The American Naturalist*, 64(695):560–566, 1930.

[85] T. M. Frayling, N. J. Timpson, M. N. Weedon, E. Zeggini, R. M. Freathy, C. M. Lindgren, J. R. Perry, K. S. Elliott, H. Lango, N. W. Rayner, and et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, 316(5826):889–894, May 2007.

[86] B. L. Fridley and J. M. Biernacka. Gene set analysis of SNP data: Benefits, challenges, and future directions. *European Journal of Human Genetics*, 19(8):837–843, Apr 2011.

[87] J. M. Friedman. Obesity in the new millennium. *Nature*, 404(6778):632–634, Apr 2000.

[88] P. Froguel and A. I. Blakemore. The power of the extreme in elucidating obesity. *New England Journal of Medicine*, 359(9):891–893, 2008.

[89] O. Froy. Metabolism and circadian rhythms—implications for obesity. *Endocrine reviews*, 31(1):1–24, 2010.

[90] S. B. Gabriel, S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly, and D. Altshuler. The structure of haplotype blocks in the human genome. *Science*, 296(5576):2225–2229, 2002.

[91] D. M. Gatti, W. T. Barry, A. B. Nobel, I. Rusyn, and F. A. Wright. Heading down the wrong pathway: On the influence of correlation within gene sets. *BMC Genomics*, 11(1), Oct 2010.

[92] J. Gayon. From mendel to epigenetics: History of genetics. *Comptes rendus biologies*, 339(7-8):225–230, 2016.

[93] V. H. Gazestani and N. E. Lewis. From genotype to phenotype: augmenting deep learning with networks and systems biology. *Current Opinion in Systems Biology*, 15:68–73, 2019. Gene regulation.

[94] E. Goan and C. Fookes. Bayesian neural networks: An introduction and survey. *Case Studies in Applied Bayesian Data Science: CIRM Jean-Morlet Chair, Fall 2018*, pages 45–87, 2020.

[95] J. J. Goeman and P. Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, 02 2007.

[96] D. Golan, E. S. Lander, and S. Rosset. Measuring missing heritability: Inferring the contribution of common variants. *Proceedings of the National Academy of Sciences*, 111(49), Nov 2014.

[97] J. Gray, G. S. Yeo, J. J. Cox, J. Morton, A.-L. R. Adlam, J. M. Keogh, J. A. Yanovski, A. El Gharbawy, J. C. Han, Y. L. Tung, and et al. Hyperphagia, severe obesity, impaired cognitive function, and hyperactivity associated with functional loss of one copy of the brain-derived neurotrophic factor (BDNF) gene. *Diabetes*, 55(12):3366–3371, Dec 2006.

[98] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, et al. Recent advances in convolutional neural networks. *Pattern recognition*, 77:354–377, 2018.

[99] D. F. Gudbjartsson, G. B. Walters, G. Thorleifsson, H. Stefansson, B. V. Halldorsson, P. Zusmanovich, P. Sulem, S. Thorlacius, A. Gylfason, S. Steinberg, and et al. Many sequence variants affecting diversity of adult human height. *Nature Genetics*, 40(5):609–615, Apr 2008.

[100] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[101] F. Günther, N. Wawro, and K. Bammann. Neural networks for modeling gene-gene interactions in association studies. *BMC genetics*, 10:1–14, 2009.

[102] E. Génin and F. Clerget-Darpoux. Revisiting the polygenic additive liability model through the example of diabetes mellitus. *Human Heredity*, 80(4):171–177, Sep 2015.

[103] J. T. Hancock and T. M. Khoshgoftaar. Survey on categorical data for neural networks. *Journal of Big Data*, 7(1), Apr 2020.

[104] D. He and L. Parida. Does encoding matter? a novel view on the quantitative genetic trait prediction problem. *BMC Bioinformatics*, 17(S9), Jul 2016.

[105] P. Hebbar and S. K. Sowmya. Genomic variant annotation: A comprehensive review of tools and techniques. In *International Conference on Intelligent Systems Design and Applications*, pages 1057–1067. Springer, 2021.

[106] A. G. Heidema, J. M. Boer, N. Nagelkerke, E. C. Mariman, D. L. van der A, and E. J. Feskens. The challenge for genetic epidemiologists: How to analyze large numbers of SNPs in relation to complex diseases. *BMC Genetics*, 7(1), Apr 2006.

[107] W. G. Hill, M. E. Goddard, and P. M. Visscher. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics*, 4(2), Feb 2008.

[108] A. Hinney and J. Hebebrand. Polygenic obesity in humans. *Obesity facts*, 1(1):35–42, 2008.

[109] J. Hoh and J. Ott. Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Reviews Genetics*, 4(9):701–709, Sep 2003.

[110] T.-C. Hsia, H.-C. Chiang, D. Chiang, L.-W. Hang, F.-J. Tsai, and W.-C. Chen. Prediction of survival in surgical unresectable lung cancer by artificial neural networks including genetic polymorphisms and clinical parameters. *Journal of Clinical Laboratory Analysis*, 17(6):229–234, Nov 2003.

[111] J. K. Hu, X. Wang, and P. Wang. Testing gene-gene interactions in genome wide association studies. *Genetic Epidemiology*, 38(2):123–134, Jan 2014.

[112] T. Hu, N. Chitnis, D. Monos, and A. Dinh. Next-generation sequencing technologies: An overview. *Human Immunology*, 82(11):801–811, 2021. Next Generation Sequencing and its Application to Medical Laboratory Immunology.

[113] D. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, Nov 2008.

[114] P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, Mar 1964.

[115] J.-H. Hung, T.-H. Yang, Z. Hu, Z. Weng, and C. DeLisi. Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in bioinformatics*, 13(3):281–291, 2012.

[116] S. K. Iyengar and R. C. Elston. *The Genetic Basis of Complex Traits*, pages 71–84. Humana Press, Totowa, NJ, 2007.

[117] A. Jakulin and I. Bratko. Quantifying and visualizing attribute interactions, 2004.

[118] K. S. Jankowski and M. Dmitrzak-Weglarz. ARNTL, CLOCK and PER3 polymorphisms–links with chronotype and affective dimensions. *Chronobiology International*, 34(8):1105–1113, 2017.

[119] A. C. Janssens. Validity of polygenic risk scores: Are we measuring what we think we are? *Human Molecular Genetics*, 28(R2), Aug 2019.

[120] M. Jaritz, R. D. Charette, E. Wirbel, X. Perrotton, and F. Nashashibi. Sparse and dense data with CNNs: Depth completion and semantic segmentation. In *2018 International Conference on 3D Vision (3DV)*, pages 52–60, 2018.

[121] P. Jia, L. Wang, H. Y. Meltzer, and Z. Zhao. Common variants conferring risk of schizophrenia: a pathway analysis of GWAS data. *Schizophrenia research*, 122(1-3):38–42, 2010.

[122] R. Jiang, W. Tang, X. Wu, and W. Fu. A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinformatics*, 10(S1), Jan 2009.

[123] L. V. Jospin, H. Laga, F. Boussaid, W. Buntine, and M. Bennamoun. Hands-on bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2):29–48, 2022.

[124] M. Kanehisa, M. Furumichi, Y. Sato, M. Kawashima, and M. Ishiguro-Watanabe. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research*, 51(D1):D587–D592, 10 2022.

[125] K. KC, R. Li, F. Cui, Q. Yu, and A. R. Haake. Gne: A deep learning framework for gene network inference by aggregating biological information. *BMC Systems Biology*, 13(S2), Apr 2019.

[126] D. R. Kelley, J. Snoek, and J. L. Rinn. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, 26(7):990–999, May 2016.

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

[128] A. V. Khera, M. Chaffin, K. H. Wade, S. Zahid, J. Brancale, R. Xia, M. Distefano, O. Senol-Cosar, M. E. Haas, A. Bick, and et al. Polygenic prediction of weight and obesity trajectories from birth to adulthood. *Cell*, 177(3), Apr 2019.

[129] T. O. Kilpeläinen, L. Qi, S. Brage, S. J. Sharp, E. Sonestedt, E. Demerath, T. Ahmad, S. Mora, M. Kaakinen, C. H. Sandholt, and et al. Physical activity attenuates the influence of FTO variants on obesity risk: A meta-analysis of 218,166 adults and 19,268 children. *PLoS Medicine*, 8(11), Nov 2011.

[130] H. Kim, A. Grueneberg, A. I. Vazquez, S. Hsu, and G. de los Campos. Will big data close the missing heritability gap? *Genetics*, 207(3):1135–1145, Sep 2017.

[131] M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310–315, Feb 2014.

[132] D. Kleftogiannis, P. Kalnis, and V. B. Bajic. DEEP: a general computational framework for predicting enhancers. *Nucleic Acids Research*, 43(1):e6–e6, 11 2014.

[133] A. Kong, G. Thorleifsson, M. L. Frigge, B. J. Vilhjalmsson, A. I. Young, T. E. Thorgeirsson, S. Benonisdottir, A. Oddsson, B. V. Halldorsson, G. Masson, and et al. The nature of nurture: Effects of parental genotypes. *Science*, 359(6374):424–428, Jan 2018.

[134] T. Konuma and Y. Okada. Statistical genetics and polygenic risk score for precision medicine. *Inflammation and Regeneration*, 41(1), Jun 2021.

[135] C. L. Koo, M. J. Liew, M. S. Mohamad, and A. H. Mohamed Salleh. A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology. *BioMed Research International*, 2013:1–13, Oct 2013.

[136] C. Kooperberg and I. Ruczinski. Identifying interacting SNPs using monte carlo logic regression. *Genetic Epidemiology*, 28(2):157–170, Feb 2005.

[137] A. Korte and A. Farlow. The advantages and limitations of trait analysis with GWAS: a review. *Plant methods*, 9(1):1–9, 2013.

[138] I. Korvigo, A. Afanasyev, N. Romashchenko, and M. Skoblov. Generalising better: Applying deep learning to integrate deleteriousness prediction scores for whole-exome snv studies. *PLOS ONE*, 13(3), Mar 2018.

[139] J. A. Kosmicki, C. L. Churchhouse, M. A. Rivas, and B. M. Neale. Discovery of rare variants for complex phenotypes. *Human genetics*, 135:625–634, 2016.

[140] J. Koza III and B. FHB. Hf & andre, d.(1999) genetic programming iii: Automatic programming and automatic circuit synthesis.

[141] D. Krentzel, S. L. Shorte, and C. Zimmer. Deep learning in image-based phenotypic drug discovery. *Trends in Cell Biology*, 33(7):538–554, Jan 2023.

[142] L. Krishnamurthy, J. Nadeau, G. Ozsoyoglu, M. Ozsoyoglu, G. Schaeffer, M. Tasan, and W. Xu. Pathways database system: an integrated system for biological pathways. *Bioinformatics*, 19(8):930–937, 2003.

[143] R. Krishnan, G. Sivakumar, and P. Bhattacharya. Extracting decision trees from trained neural networks. *Pattern recognition*, 32(12), 1999.

[144] J. Kumuthini, B. Zick, A. Balasopoulou, C. Chalikiopoulou, C. Dandara, G. El-Kamah, L. Findley, T. Katsila, R. Li, E. B. Maceda, and et al. The clinical utility of polygenic risk scores in genomic medicine practices: A systematic review. *Human Genetics*, 141(11):1697–1704, Apr 2022.

[145] J. Laermans and I. Depoortere. Chronobesity: role of the circadian system in the obesity epidemic. *Obesity reviews*, 17(2):108–125, 2016.

[146] S. A. Lambert, G. Abraham, and M. Inouye. Towards clinical utility of polygenic risk scores. *Human Molecular Genetics*, 28(R2), Jul 2019.

[147] J. LANCHANTIN, R. SINGH, B. WANG, and Y. QI. Deep motif dashboard: Visualizing and understanding genomic sequences using deep neural networks. *Biocomputing 2017*, Nov 2016.

[148] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[149] Y. LeCun, L. Bottou, G. B. Orr, and K. R. Müller. Efficient backprop. *Lecture Notes in Computer Science*, page 9–50, Mar 2002.

[150] Y.-C. Lee, J. J. Christensen, L. D. Parnell, C. E. Smith, J. Shao, N. M. McKeown, J. M. Ordovás, and C.-Q. Lai. Using machine learning to predict obesity based on genome-wide and epigenome-wide gene–gene and gene–diet interactions. *Frontiers in Genetics*, 12, Jan 2022.

[151] A. C. Lewis, R. C. Green, and J. L. Vassy. Polygenic risk scores in the clinic: Translating risk into action. *Human Genetics and Genomics Advances*, 2(4):100047, Oct 2021.

[152] C. M. Lewis and E. Vassos. Polygenic risk scores: From research tools to clinical instruments. *Genome Medicine*, 12(1), May 2020.

[153] J. Li, R. Tang, J. M. Biernacka, and M. de Andrade. Identification of gene-gene interaction using principal components. *BMC Proceedings*, 3(S7), Dec 2009.

[154] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. 2016.

[155] L. Li, K. Jamieson, A. Rostamizadeh, E. Gonina, J. Ben-tzur, M. Hardt, B. Recht, and A. Talwalkar. A system for massively parallel hyperparameter tuning. In *Third Conference on Systems and Machine Learning*, 2020.

[156] X. Li, H. Xiong, X. Li, X. Wu, X. Zhang, J. Liu, J. Bian, and D. Dou. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems*, 64(12):3197–3234, Sep 2022.

[157] Y. Li, W. Shi, and W. W. Wasserman. Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. *BMC Bioinformatics*, 19(1), May 2018.

[158] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12):6999–7019, 2022.

[159] P. Liashchynskyi and P. Liashchynskyi. Grid search, random search, genetic algorithm: a big comparison for NAS. *arXiv preprint arXiv:1912.06059*, 2019.

[160] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, and I. Stoica. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*, 2018.

[161] A. W.-C. Liew, H. Yan, and M. Yang. Pattern recognition techniques for the emerging field of bioinformatics: A review. *Pattern Recognition*, 38(11):2055–2073, 2005.

[162] F. Liu, H. Li, C. Ren, X. Bo, and W. Shu. Pedla: Predicting enhancers with a deep learning-based algorithmic framework. *Scientific Reports*, 6(1), Jun 2016.

[163] Y. Liu, N. Li, X. Zhu, and Y. Qi. How wide is the application of genetic big data in biomedicine. *Biomedicine Pharmacotherapy*, 133:111074, 2021.

[164] Y. Liu, D. Wang, F. He, J. Wang, T. Joshi, and D. Xu. Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. *Frontiers in Genetics*, 10, Nov 2019.

[165] A. E. Locke, B. Kahali, S. I. Berndt, A. E. Justice, T. H. Pers, F. R. Day, C. Powell, S. Vedantam, M. L. Buchkovich, J. Yang, and et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206, Feb 2015.

[166] K. E. Lohmueller, C. L. Pearce, M. Pike, E. S. Lander, and J. N. Hirschhorn. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature genetics*, 33(2):177–182, 2003.

[167] R. J. Loos and A. C. Janssens. Predicting polygenic obesity using genetic information. *Cell Metabolism*, 25(3):535–543, Mar 2017.

[168] R. J. Loos and G. S. Yeo. The bigger picture of FTO—the first GWAS-identified obesity gene. *Nature Reviews Endocrinology*, 10(1):51–61, 2014.

[169] R. J. Loos and G. S. Yeo. The genetics of obesity: From discovery to biology. *Nature Reviews Genetics*, 23(2):120–133, Sep 2021.

[170] S. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions, 2017.

[171] D. Ma, P. Whitehead, M. Menold, E. Martin, A. Ashley-Koch, H. Mei, M. Ritchie, G. Delong, R. Abramson, H. Wright, et al. Identification of significant association and gene-gene interaction of gaba receptor subunit genes in autism. *The American Journal of Human Genetics*, 77(3):377–388, 2005.

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

[173] D. J. MacKay. Bayesian neural networks and density networks. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 354(1):73–80, 1995. Proceedings of the Third Workshop on Neutron Scattering Data Analysis.

[174] N. S. Madhukar, O. Elemento, and G. Pandey. Prediction of genetic interactions using machine learning and network properties. *Frontiers in Bioengineering and Biotechnology*, 3, Oct 2015.

[175] H. H. Maes, M. C. Neale, and L. J. Eaves. Genetic and environmental factors in relative body weight and human adiposity. *Behavior Genetics*, 27(4):325–351, Jul 1997.

[176] B. S. Maher. Polygenic scores in epidemiology: Risk prediction, etiology, and clinical utility. *Current Epidemiology Reports*, 2(4):239–244, Sep 2015.

[177] F. Maleki. and A. J. Kusalik. Gene set overlap: An impediment to achieving high specificity in over-representation analysis. In *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2019) - BIOINFORMATICS*, pages 182–193. INSTICC, SciTePress, 2019.

[178] F. Maleki, K. Ovens, D. J. Hogan, and A. J. Kusalik. Gene set analysis: Challenges, opportunities, and future research. *Frontiers in Genetics*, 11, Jun 2020.

[179] F. Manna, G. Martin, and T. Lenormand. Fitness landscapes: an alternative theory for the dominance of mutation. *Genetics*, 189(3):923–937, 2011.

[180] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, and et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, Mar 2009.

[181] H. A. Mansour, T. H. Monk, and V. L. Nimgaonkar. Circadian genes and bipolar disorder. *Annals of medicine*, 37(3):196–205, 2005.

[182] J. Marchini, P. Donnelly, and L. R. Cardon. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics*, 37(4):413–417, Mar 2005.

[183] W. E. Marcílio and D. M. Eler. From explanations to feature selection: assessing SHAP values as feature selection mechanism. In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 340–347, 2020.

[184] A. T. Marees, H. de Kluiver, S. Stringer, F. Vorspan, E. Curis, C. Marie-Claire, and E. M. Derks. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, 27(2), Feb 2018.

[185] E. Marouli, M. Graff, C. Medina-Gomez, K. S. Lo, A. R. Wood, T. R. Kjaer, R. S. Fine, Y. Lu, C. Schurmann, H. M. Highland, and et al. Rare and low-frequency coding variants alter human adult height. *Nature*, 542(7640):186–190, Aug 2017.

[186] E. Martin, M. Ritchie, L. Hahn, S. Kang, and J. Moore. A novel method to identify gene–gene effects in nuclear families: The MDR-PDT. *Genetic Epidemiology*, 30(2):111–123, Feb 2006.

[187] B. A. McKinney, D. M. Reif, M. D. Ritchie, and J. H. Moore. Machine learning for detecting gene-gene interactions. *Applied Bioinformatics*, 5(2):77–88, Aug 2012.

[188] L. Merrick. Randomized ablation feature importance, 2019.

[189] B. Mieth, M. Kloft, J. A. Rodríguez, S. Sonnenburg, R. Vobruba, C. Morcillo-Suárez, X. Farré, U. M. Marigorta, E. Fehr, T. Dickhaus, et al. Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies. *Scientific reports*, 6(1):36671, 2016.

[190] B. Mieth, A. Rozier, J. A. Rodriguez, M. M. Höhne, N. Görnitz, and K.-R. Müller. Deepcombi: Explainable artificial intelligence for the analysis and discovery in genome-wide association studies. *NAR Genomics and Bioinformatics*, 3(3), Jul 2021.

[191] S. W. Mifflin and A. Strack. Obesity and the central nervous system. *The Journal of Physiology*, 583(2):423–423, Aug 2007.

[192] J. Millstein, D. V. Conti, F. D. Gilliland, and W. J. Gauderman. A testing framework for identifying susceptibility genes in the presence of epistasis. *The American Journal of Human Genetics*, 78(1):15–27, Jan 2006.

[193] P. Mishra, P. Törönen, Y. Leino, and L. Holm. Gene set analysis: limitations in popular existing methods and proposed improvements. *Bioinformatics*, 30(19):2747–2756, 2014.

[194] N. S. Mitchell, V. A. Catenacci, H. R. Wyatt, and J. O. Hill. Obesity: Overview of an epidemic. *Psychiatric Clinics of North America*, 34(4):717–732, Dec 2011.

[195] C. A. Montanez, P. Fergus, C. Chalmers, N. H. Malim, B. Abdulaimma, D. Reilly, and F. Falciani. Saerma: Stacked autoencoder rule mining algorithm for the interpretation of epistatic interactions in GWAS for extreme obesity. *IEEE Access*, 8:112379–112392, Jun 2020.

[196] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller. Layer-wise relevance propagation: An overview. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, page 193–209, Sep 2019.

[197] C. A. C. Montañez, P. Fergus, A. Hussain, D. Al-Jumeily, B. Abdulaimma, J. Hind, and N. Radi. Machine learning approaches for the prediction of obesity using publicly available genetic profiles. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2743–2750, 2017.

[198] C. A. C. Montañez, P. Fergus, A. C. Montañez, and C. Chalmers. Deep learning classification of polygenic obesity using genome wide association study SNPs, 2018.

[199] R. Moonesinghe, M. J. Khoury, T. Liu, and A. C. Janssens. Discriminative accuracy of genomic profiling comparing multiplicative and additive risk models. *European Journal of Human Genetics*, 19(2):180–185, Nov 2010.

[200] J. Moore, J. Lamb, N. Brown, and D. Vaughan. A comparison of combinatorial partitioning and linear regression for the detection of epistatic effects of the ACE I/D and PAI-1 4G/5G polymorphisms on plasma PAI-1 levels. *Clinical Genetics*, 62(1):74–79, 2002.

[201] J. H. Moore. The challenges of whole-genome approaches to common diseases. *JAMA: The Journal of the American Medical Association*, 291(13):1642–1643, Apr 2004.

[202] J. H. Moore. A global view of epistasis. *Nature Genetics*, 37(1):13–14, Jan 2005.

[203] J. H. Moore, J. C. Gilbert, C.-T. Tsai, F.-T. Chiang, T. Holden, N. Barney, and B. C. White. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *Journal of Theoretical Biology*, 241(2):252–261, Jul 2006.

[204] J. H. Moore and S. M. Williams. New strategies for identifying gene-gene interactions in hypertension. *Annals of Medicine*, 34(2):88–95, Jan 2002.

[205] J. H. Moore and S. M. Williams. Traversing the conceptual divide between biological and statistical epistasis: Systems biology and a more modern synthesis. *BioEssays*, 27(6):637–646, Jun 2005.

[206] J. H. Moore and S. M. Williams. Epistasis and its implications for personal genetics. *The American Journal of Human Genetics*, 85(3):309–320, 2009.

[207] B. P. Morgan. The complement system: an overview. *Complement methods and protocols*, pages 1–13, 2000.

[208] H. Mostafavi, A. Harpak, D. Conley, J. K. Pritchard, and M. Przeworski. Variable prediction accuracy of polygenic scores within an ancestry group. *bioRxiv*, 2019.

[209] A. A. Motsinger, B. S. Donahue, N. J. Brown, D. M. Roden, and M. D. Ritchie. Risk factor interactions and genetic effects associated with post-operative atrial fibrillation. In *Biocomputing 2006*, pages 584–595. World Scientific, 2006.

[210] A. A. Motsinger, S. M. Dudek, L. W. Hahn, and M. D. Ritchie. Comparison of neural network optimization approaches for studies of human genetics. In F. Rothlauf, J. Branke, S. Cagnoni, E. Costa, C. Cotta, R. Drechsler, E. Lutton, P. Machado, J. H. Moore, J. Romero, G. D. Smith, G. Squillero, and H. Takagi, editors, *Applications of Evolutionary Computing*, pages 103–114, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[211] A. A. Motsinger, S. L. Lee, G. Mellick, and M. D. Ritchie. GPNN: Power studies and applications of a neural network method for detecting gene-gene interactions in studies of human disease. *BMC Bioinformatics*, 7(1), Jan 2006.

[212] A. A. Motsinger and M. D. Ritchie. Multifactor dimensionality reduction: An analysis strategy for modelling and detecting gene - gene interactions in human genetics and pharmacogenomics studies. *Human Genomics*, 2(5), Mar 2006.

[213] A. A. Motsinger-Reif, S. M. Dudek, L. W. Hahn, and M. D. Ritchie. Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene

interactions in genetic epidemiology. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 32(4):325–340, 2008.

[214] A. A. Motsinger-Reif, T. J. Fanelli, A. C. Davis, and M. D. Ritchie. Power of grammatical evolution neural networks to detect gene-gene interactions in the presence of error. *BMC Research Notes*, 1(1), Aug 2008.

[215] M. Muneeb, S. Feng, and A. Henschel. Transfer learning for genotype–phenotype prediction using deep learning models. *BMC Bioinformatics*, 23(1), Nov 2022.

[216] S. K. Musani, D. Shriner, N. Liu, R. Feng, C. S. Coffey, N. Yi, H. K. Tiwari, and D. B. Allison. Detection of gene × gene interactions in genome-wide association studies of human population data. *Human Heredity*, 63(2):67–84, Feb 2007.

[217] K. M. Naimul Hassan, M. Shamiul Alam Hridoy, N. Tasnim, A. F. Chowdhury, T. Alam Roni, S. Tabrez, A. Subhana, and C. Shahnaz. Alsnet: A dilated 1-D CNN for identifying ALS from raw EMG signal. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1181–1185, 2022.

[218] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic. Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1), Feb 2015.

[219] M. Nelson, S. Kardia, R. Ferrell, and C. Sing. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Research*, 11(3):458–470, Mar 2001.

[220] R. M. Nelson, M. E. Pettersson, and Carlborg. A century after Fisher: Time for a new paradigm in quantitative genetics. *Trends in Genetics*, 29(12):669–676, Oct 2013.

[221] H. L. Nicholls, C. R. John, D. S. Watson, P. B. Munroe, M. R. Barnes, and C. P. Cabrera. Reaching the end-game for GWAS: machine learning approaches for the prioritization of complex disease loci. *Frontiers in genetics*, 11:350, 2020.

[222] C. M. Nievergelt, D. F. Kripke, T. B. Barrett, E. Burg, R. A. Remick, A. D. Sadovnick, S. L. McElroy, P. E. Keck Jr, N. J. Schork, and J. R. Kelsoe. Suggestive evidence for association of the circadian genes PERIOD3 and ARNTL with bipolar disorder. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 141(3):234–241, 2006.

[223] M. N. Nitabach and J. Blau. Cellular clockwork. *nature genetics*, 32(4):559–560, 2002.

[224] C. O'Dushlaine, E. Kenny, E. A. Heron, R. Segurado, M. Gill, D. W. Morris, and A. Corvin. The SNP ratio test: pathway analysis of genome-wide association datasets. *Bioinformatics*, 25(20):2762–2763, 07 2009.

[225] D. W. Otter, J. R. Medina, and J. K. Kalita. A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):604–624, 2021.

[226] G. P. Page, V. George, R. C. Go, P. Z. Page, and D. B. Allison. "are we there yet?": Deciding when one has demonstrated specific genetic causation in complex diseases and quantitative traits. *The American Journal of Human Genetics*, 73(4):711–719, Oct 2003.

[227] Y. Park and M. Kellis. Deep learning for regulatory genomics. *Nature Biotechnology*, 33(8):825–826, Aug 2015.

[228] T. Partonen, J. Treutlein, A. Alpman, J. Frank, C. Johansson, M. Depner, L. Aron, M. Rietschel, S. Wellek, P. Soronen, et al. Three circadian clock genes PER2, ARNTL, and NPAS2 contribute to winter depression. *Annals of medicine*, 39(3):229–238, 2007.

[229] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

[230] L. R. Pearce, N. Atanassova, M. C. Banton, B. Bottomley, A. A. van der Klaauw, J.-P. Revelli, A. Hendricks, J. M. Keogh, E. Henning, D. Doree, et al. KSR2 mutations are associated with obesity, insulin resistance, and impaired cellular fuel oxidation. *Cell*, 155(4):765–777, 2013.

[231] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau,

M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[232] S. Peng, Y. Zhu, F. Xu, X. Ren, X. Li, and M. Lai. FTO gene polymorphisms and obesity risk: A meta-analysis. *BMC Medicine*, 9(1), Jun 2011.

[233] J. R. Perry, M. I. McCarthy, A. T. Hattersley, E. Zeggini, W. T. C. C. Consortium, M. N. Weedon, and T. M. Frayling. Interrogating type 2 diabetes genome-wide association data using a biological pathway-based approach. *Diabetes*, 58(6):1463–1467, 2009.

[234] P. C. Phillips. The language of gene interaction. *Genetics*, 149(3):1167–1171, 1998.

[235] P. C. Phillips. Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9(11):855–867, Nov 2008.

[236] F. X. PI-SUNYER. Comorbidities of overweight and obesity: Current evidence and research issues. *Medicine amp; Science in Sports amp; Exercise*, 31(Supplement 1), Nov 1999.

[237] S. Pirmoradi, M. Teshnehlab, N. Zarghami, and A. Sharifi. A self-organizing deep auto-encoder approach for classification of complex diseases using SNP genomics data. *Applied Soft Computing*, 97:106718, 2020.

[238] F. Privé, B. J. Vilhjálmsson, H. Aschard, and M. G. Blum. Making the most of clumping and thresholding for polygenic scores. *The American Journal of Human Genetics*, 105(6):1213–1221, 2019.

[239] F. Privé, J. Arbel, and B. J. Vilhjálmsson. LDPRED2: Better, faster, stronger. *Bioinformatics*, 36(22–23):5424–5431, Dec 2020.

[240] H. Qi, C. Chen, H. Zhang, J. J. Long, W. K. Chung, Y. Guan, and Y. Shen. MVP: predicting pathogenicity of missense variants by deep neural networks. *bioRxiv*, 2018.

[241] Q. Qi, A. Y. Chu, J. H. Kang, J. Huang, L. M. Rose, M. K. Jensen, L. Liang, G. C. Curhan, L. R. Pasquale, J. L. Wiggs, and et al. Fried food consumption, genetic risk, and body

mass index: Gene-diet interaction analysis in three us cohort studies. *BMJ*, 348(mar19 1), Mar 2014.

[242] D. Qin. Next-generation sequencing and its clinical application. *Cancer Biology amp;amp; Medicine*, 16(1):4–10, Feb 2019.

[243] S. Qin, X. Zhao, Y. Pan, J. Liu, G. Feng, J. Fu, J. Bao, Z. Zhang, and L. He. An association study of the N-methyl-D-aspartate receptor NR1 subunit gene (GRIN1) and NR2B subunit gene (GRIN2B) in schizophrenia with universal DNA microarray. *European Journal of Human Genetics*, 13(7):807–814, 2005.

[244] D. Quang, Y. Chen, and X. Xie. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*, 31(5):761–763, 10 2014.

[245] D. Quang and X. Xie. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Research*, 44(11):e107–e107, 04 2016.

[246] A. Rammos, L. A. Gonzalez, D. R. Weinberger, K. J. Mitchell, and K. K. Nicodemus. The role of polygenic risk score gene-set analysis in the context of the omnigenic model of schizophrenia. *Neuropsychopharmacology*, 44(9):1562–1569, May 2019.

[247] M. Rask-Andersen, M. S. Almén, H. R. Olausen, P. K. Olszewski, J. Eriksson, R. A. Chavan, A. S. Levine, R. Fredriksson, and H. B. Schiöth. Functional coupling analysis suggests link between the obesity gene FTO and the BDNF-NTRK2 signaling pathway. *BMC Neuroscience*, 12(1), Nov 2011.

[248] H. Reddon, J.-L. Guéant, and D. Meyre. The importance of gene–environment interactions in human obesity. *Clinical science*, 130(18):1571–1597, 2016.

[249] E. Reed, S. Nunez, D. Kulp, J. Qian, M. P. Reilly, and A. S. Foulkes. A guide to genome-wide association analysis and post-analytic interrogation. *Statistics in Medicine*, 34(28):3769–3792, Sep 2015.

[250] D. E. Reich and E. S. Lander. On the allelic spectrum of human disease. *Trends in Genetics*, 17(9):502–510, Sep 2001.

[251] S. K. Remold and R. E. Lenski. Pervasive joint influence of epistasis and plasticity on mutational effects in escherichia coli. *Nature genetics*, 36(4):423–426, 2004.

[252] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016.

[253] M. Riedmiller and H. Braun. A direct adaptive method for faster backpropagation learning: the RPROP algorithm. In *IEEE International Conference on Neural Networks*, pages 586–591 vol.1, 1993.

[254] M. D. Ritchie, L. W. Hahn, and J. H. Moore. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genetic Epidemiology*, 24(2):150–157, Feb 2003.

[255] M. D. Ritchie, L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl, and J. H. Moore. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics*, 69(1):138–147, Jul 2001.

[256] M. D. Ritchie and K. Van Steen. The search for gene-gene interactions in genome-wide association studies: challenges in abundance of methods, practical considerations, and biological interpretation. *Annals of translational medicine*, 6(8), 2018.

[257] M. D. Ritchie, B. C. White, J. S. Parker, L. W. Hahn, and J. H. Moore. *BMC Bioinformatics*, 4(1):28, Jul 2003.

[258] M. Robnik-Šikonja and I. Kononenko. *Machine Learning*, 53(1/2):23–69, Oct 2003.

[259] F. Rosenblatt et al. *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*, volume 55. Spartan books Washington, DC, 1962.

[260] A. E. Roth. *Introduction to the Shapley value*, page 1–28. Cambridge University Press, 1988.

[261] S. Ruder. An overview of gradient descent optimization algorithms. *CoRR*, abs/1609.04747, 2016.

[262] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, Oct 1986.

[263] A. Saint Pierre and E. Génin. How important are rare variants in common disease? *Briefings in functional genomics*, 13(5):353–361, 2014.

[264] H. Satam, K. Joshi, U. Mangrolia, S. Waghoo, G. Zaidi, S. Rawool, R. P. Thakare, S. Banday, A. K. Mishra, G. Das, et al. Next-generation sequencing technology: Current trends and advancements. *Biology*, 12(7):997, 2023.

[265] E. W. Sayers, E. E. Bolton, J. R. Brister, K. Canese, J. Chan, D. Comeau, R. Connor, K. Funk, C. Kelly, S. Kim, and et al. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 50(D1), Dec 2021.

[266] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, Jan 2015.

[267] J. Schmitz, F. Abbondanza, and S. Paracchini. Genome-wide association study and polygenic risk score analysis for hearing measures in children. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 186(5):318–328, Jul 2021.

[268] P. Schober, C. Boer, and L. A. Schwarte. Correlation coefficients: Appropriate use and interpretation. *Anesthesia amp;amp; Analgesia*, 126(5):1763–1768, May 2018.

[269] A. Scuteri, S. Sanna, W.-M. Chen, M. Uda, G. Albai, J. Strait, S. Najjar, R. Nagaraja, M. Orrú, G. Usala, and et al. Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genetics*, 3(7), Jul 2007.

[270] H. Shi, G. Kichaev, and B. Pasaniuc. Contrasting the genetic architecture of 30 complex traits from summary association data. *The American Journal of Human Genetics*, 99(1):139–153, Jun 2016.

[271] J. Shi and M. Walker. Gene set enrichment analysis (GSEA) for interpreting gene expression profiles. *Current Bioinformatics*, 2(2):133–137, May 2007.

[272] S.-q. Shi, T. S. Ansari, O. P. McGuinness, D. H. Wasserman, and C. H. Johnson. Circadian disruption leads to insulin resistance and obesity. *Current Biology*, 23(5):372–381, 2013.

[273] A. Shrestha and A. Mahmood. Review of deep learning algorithms and architectures. *IEEE Access*, 7:53040–53065, 2019.

[274] D. M. Skapura. *Building neural networks*. Addison-Wesley Professional, 1996.

[275] M. Slatkin. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–485, Jun 2008.

[276] J. L. Slunecka, M. D. van der Zee, J. J. Beck, B. N. Johnson, C. T. Finnicum, R. Pool, J.-J. Hottenga, E. J. de Geus, and E. A. Ehli. Implementation and implications for polygenic risk scores in healthcare. *Human Genomics*, 15(1), Jul 2021.

[277] Y. B. Sohn. The genetics of obesity: A narrative review. *Precision and Future Medicine*, 6(4):226–232, Dec 2022.

[278] E. Sollis, A. Mosaku, A. Abid, A. Buniello, M. Cerezo, L. Gil, T. Groza, O. Güneş, P. Hall, J. Hayhurst, and et al. The NHGRI-EBI GWAS catalog: Knowledgebase and deposition resource. *Nucleic Acids Research*, 51(D1), Nov 2022.

[279] J. R. Speakman. The 'fat mass and obesity related'(FTO) gene: mechanisms of impact on obesity and energy balance. *Current obesity reports*, 4(1):73–91, 2015.

[280] G. Stelzer, N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, T. Iny Stein, R. Nudel, I. Lieder, Y. Mazor, and et al. The genecards suite: From gene data mining to disease genome sequence analyses. *Current Protocols in Bioinformatics*, 54(1), Jun 2016.

[281] G. Stelzer, N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, T. I. Stein, R. Nudel, I. Lieder, Y. Mazor, S. Kaplan, D. Dahary, D. Warshawsky, Y. Guan-Golan, A. Kohn, N. Rappaport, M. Safran, and D. Lancet. The genecards suite: From gene

data mining to disease genome sequence analyses. *Current Protocols in Bioinformatics*, 54(1):1.30.1–1.30.33, 2016.

[282] S. Stringer, N. R. Wray, R. S. Kahn, and E. M. Derks. Underestimated effect sizes in GWAS: fundamental limitations of single SNP analysis for dichotomous phenotypes. *PloS one*, 6(11):e27964, 2011.

[283] A. Sud, R. H. Horton, A. D. Hingorani, I. Tzoulaki, C. Turnbull, R. S. Houlston, and A. Lucassen. Realistic expectations are key to realising the benefits of polygenic scores. *BMJ*, Mar 2023.

[284] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, and et al. 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), Mar 2015.

[285] S. Taghavi Namin, M. Esmaeilzadeh, M. Najafi, T. B. Brown, and J. O. Borevitz. Deep phenotyping: Deep learning for temporal phenotype/genotype classification. *Plant Methods*, 14(1), Aug 2018.

[286] V. Tam, N. Patel, M. Turcotte, Y. Bossé, G. Paré, and D. Meyre. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, 20(8):467–484, Aug 2019.

[287] J. Tan, J. H. Hammond, D. A. Hogan, and C. S. Greene. Adage-based integration of publicly available pseudomonas aeruginosa gene expression data with denoising autoencoders illuminates microbe-host interactions. *mSystems*, 1(1), Feb 2016.

[288] Z. Tang, Y. Xu, L. Jin, A. Aibaidula, J. Lu, Z. Jiao, J. Wu, H. Zhang, and D. Shen. Deep learning of imaging phenotype and genotype for predicting overall survival time of glioblastoma patients. *IEEE Transactions on Medical Imaging*, 39(6):2100–2109, 2020.

[289] A. L. Tarca, S. Draghici, G. Bhatti, and R. Romero. Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics*, 13(1), Jun 2012.

[290] J. P. Taylor, R. H. Brown Jr, and D. W. Cleveland. Decoding als: from genes to mechanism. *Nature*, 539(7628):197–206, 2016.

[291] M. B. Taylor and I. M. Ehrenreich. Higher-order genetic interactions and their contribution to complex traits. *Trends in genetics*, 31(1):34–40, 2015.

[292] H. Team. Hail.

[293] T. A. Thornton-Wells, J. H. Moore, and J. L. Haines. Genetics, statistics and human disease: Analytical retooling for complexity. *Trends in Genetics*, 20(12):640–647, Dec 2004.

[294] L. Tiret, P. Ducimetière, A. Bonnardeaux, F. Soubrier, O. Poirier, S. Ricard, F. Cambien, P. Marques-Vidal, A. Evans, F. Kee, et al. Synergistic effects of angiotensin-converting enzyme and angiotensin-IItype 1 receptor gene polymorphisms on risk of myocardial infarction. *The Lancet*, 344(8927):910–913, 1994.

[295] S. Tomida, T. Hanai, N. Koma, Y. Suzuki, T. Kobayashi, and H. Honda. Artificial neural network predictive model for allergic disease using single nucleotide polymorphisms data. *Journal of bioscience and bioengineering*, 93(5):470–478, 2002.

[296] Y. Tomita, S. Tomida, Y. Hasegawa, Y. Suzuki, T. Shirakawa, T. Kobayashi, and H. Honda. Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma. *BMC bioinformatics*, 5(1):1–13, 2004.

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

[298] S. D. Turner, S. M. Dudek, and M. D. Ritchie. ATHENA: A knowledge-based hybrid backpropagation-grammatical evolution neural network algorithm for discovering epistasis among quantitative trait loci. *BioData mining*, 3:1–18, 2010.

[299] H. R. Ueda, S. Hayashi, W. Chen, M. Sano, M. Machida, Y. Shigeyoshi, M. Iino, and S. Hashimoto. System-level identification of transcriptional circuits underlying mammalian circadian clocks. *Nature genetics*, 37(2):187–192, 2005.

[300] A. Van Der Linde and G. Tutz. On association in regression: The coefficient of determination revisited. *Statistics*, 42(1):1–24, Jan 2009.

[301] A. van Hilten, S. A. Kushner, M. Kayser, M. A. Ikram, H. H. Adams, C. C. Klaver, W. J. Niessen, and G. V. Roshchupkin. GenNet framework: Interpretable deep learning for predicting phenotypes from genetic data. *Communications Biology*, 4(1), Sep 2021.

[302] W. van Rheenen, A. Shatunov, A. M. Dekker, R. L. McLaughlin, F. P. Diekstra, S. L. Pulit, R. A. van der Spek, U. Võsa, S. de Jong, M. R. Robinson, and et al. Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nature Genetics*, 48(9):1043–1048, Jul 2016.

[303] J. Venugopalan, L. Tong, H. R. Hassanzadeh, and M. D. Wang. Multimodal deep learning models for early detection of alzheimer's disease stage. *Scientific Reports*, 11(1), Feb 2021.

[304] B. Vilhjálmsson, J. Yang, H. Finucane, A. Gusev, S. Lindström, S. Ripke, G. Genovese, P.-R. Loh, G. Bhatia, R. Do, and et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics*, 97(4):576–592, Oct 2015.

[305] P. Visscher, M. Brown, M. McCarthy, and J. Yang. Five years of GWAS discovery. *The American Journal of Human Genetics*, 90(1):7–24, Jan 2012.

[306] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang. Five years of GWAS discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

[307] P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang. 10 years of GWAS discovery: Biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, Jul 2017.

[308] D. Wallach and B. Goffinet. Mean squared error of prediction as a criterion for evaluating and comparing system models. *Ecological Modelling*, 44(3–4):299–306, Jan 1989.

[309] A. J. Walley, J. E. Asher, and P. Froguel. The genetic contribution to non-syndromic human obesity. *Nature Reviews Genetics*, 10(7):431–442, 2009.

[310] R. G. Walters, S. Jacquemont, A. Valsesia, A. J. de Smith, D. Martinet, J. Andersson, M. Falchi, F. Chen, J. Andrieux, S. Lobbens, and et al. A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. *Nature*, 463(7281):671–675, Feb 2010.

[311] K. Wang, M. Li, and M. Bucan. Pathway-based approaches for analysis of genomewide association studies. *The American Journal of Human Genetics*, 81(6):1278–1283, 2007.

[312] K. Wang, M. Li, and H. Hakonarson. Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics*, 11(12):843–854, Nov 2010.

[313] L. Wang, P. Jia, R. D. Wolfinger, X. Chen, and Z. Zhao. Gene set analysis of genome-wide association studies: Methodological issues and perspectives. *Genomics*, 98(1):1–8, 2011.

[314] Q. Wang, Y. Ma, K. Zhao, and Y. Tian. A comprehensive survey of loss functions in machine learning. *Annals of Data Science*, pages 1–26, 2020.

[315] X. Wang, Y. Zhao, and F. Pourpanah. Recent advances in deep learning. *International Journal of Machine Learning and Cybernetics*, 11:747–750, 2020.

[316] Y. Wang, T. Liu, D. Xu, H. Shi, C. Zhang, Y.-Y. Mo, and Z. Wang. Predicting DNA methylation state of CPG dinucleotide using genome topological features and deep networks. *Scientific Reports*, 6(1), Jan 2016.

[317] Y. Wang, K. Tsuo, M. Kanai, B. M. Neale, and A. R. Martin. Challenges and opportunities for developing more generalizable polygenic risk scores. *Annual Review of Biomedical Data Science*, 5(1):293–320, Jul 2022.

[318] Z. Wang, S. W. Choi, N. Chami, E. Boerwinkle, M. Fornage, S. Redline, J. C. Bis, J. A. Brody, B. M. Psaty, W. Kim, and et al. The value of rare genetic variation in the prediction of common obesity in european ancestry populations. *Frontiers in Endocrinology*, 13, May 2022.

[319] Z. Wang, J. H. Sul, S. Snir, J. A. Lozano, and E. Eskin. Gene–gene interactions detection using a two-stage model. *Journal of Computational Biology*, 22(6):563–576, May 2015.

[320] T. N. Wei. Explaining negative R-squared, Jun 2022.

[321] W. J. WELCH. A mean squared error criterion for the design of experiments. *Biometrika*, 70(1):205–213, 04 1983.

[322] E. Wheeler, N. Huang, E. G. Bochukova, J. M. Keogh, S. Lindsay, S. Garg, E. Henning, H. Blackburn, R. J. Loos, N. J. Wareham, et al. Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity. *Nature genetics*, 45(5):513–517, 2013.

[323] A. Wilkie. The molecular basis of genetic dominance. *Journal of medical genetics*, 31(2):89–98, 1994.

[324] N. R. Wray and M. E. Goddard. Multi-locus models of genetic risk of disease. *Genome Medicine*, 2(2):10, Feb 2010.

[325] N. R. Wray, M. E. Goddard, and P. M. Visscher. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Research*, 17(10):1520–1528, Sep 2007.

[326] N. R. Wray, T. Lin, J. Austin, J. J. McGrath, I. B. Hickie, G. K. Murray, and P. M. Visscher. From basic science to clinical application of polygenic risk scores: a primer. *JAMA psychiatry*, 78(1):101–109, 2021.

[327] R. Xie, J. Wen, A. Quitadamo, J. Cheng, and X. Shi. A deep auto-encoder model for gene expression prediction. *BMC Genomics*, 18(S9), Nov 2017.

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

[329] L. Yengo, J. Sidorenko, K. E. Kemper, Z. Zheng, A. R. Wood, M. N. Weedon, T. M. Frayling, J. Hirschhorn, J. Yang, and P. M. Visscher. Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of european ancestry. *Human Molecular Genetics*, 27(20):3641–3649, Aug 2018.

[330] G. S. Yeo, I. S. Farooqi, S. Aminian, D. J. Halsall, R. G. Stanhope, and S. O'Rahilly. A frameshift mutation in MC4R associated with dominantly inherited human obesity. *Nature genetics*, 20(2):111–112, 1998.

[331] M. Yoshida and A. Koike. SNPInterForest: A new method for detecting epistatic interactions. *BMC Bioinformatics*, 12(1), Dec 2011.

[332] C. You, Z. Zhou, J. Wen, Y. Li, C. H. Pang, H. Du, Z. Wang, X.-H. Zhou, D. A. King, C.-T. Liu, and et al. Polygenic scores and parental predictors: An adult height study based on the United Kingdom biobank and the Framingham heart study. *Frontiers in Genetics*, 12, May 2021.

[333] A. I. Young. Solving the missing heritability problem. *PLoS genetics*, 15(6):e1008222, 2019.

[334] H. Zeng, M. D. Edwards, G. Liu, and D. K. Gifford. Convolutional neural network architectures for predicting DNA–protein binding. *Bioinformatics*, 32(12):i121–i127, 06 2016.

[335] E. Zhang, B. Spronck, J. D. Humphrey, and G. E. Karniadakis. G2net: Relating genotype and biomechanical phenotype of tissues with deep learning. *PLOS Computational Biology*, 18(10), Oct 2022.

[336] Q. Zhang, L. T. Yang, Z. Chen, and P. Li. A survey on deep learning for big data. *Information Fusion*, 42:146–157, 2018.

[337] X. Zhang, S. Huang, F. Zou, and W. Wang. TEAM: efficient two-locus epistasis tests in human genome-wide association study. *Bioinformatics*, 26(12):i217–i227, 06 2010.

[338] J. Zhao, L. Jin, and M. Xiong. Test for interaction between two unlinked loci. *The American Journal of Human Genetics*, 79(5):831–845, Nov 2006.

[339] X. Zhao, Y. Yang, B.-F. Sun, Y.-L. Zhao, and Y.-G. Yang. FTO and obesity: mechanisms of association. *Current diabetes reports*, 14:1–9, 2014.

[340] J. Zhou and O. G. Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931–934, Aug 2015.

[341] Z. Zhuang, X. Shen, and W. Pan. A simple convolutional neural network for prediction of enhancer–promoter interactions with DNA sequence data. *Bioinformatics*, 35(17):2899–2906, 2019.

[342] O. Zuk, S. F. Schaffner, K. Samocha, R. Do, E. Hechter, S. Kathiresan, M. J. Daly, B. M. Neale, S. R. Sunyaev, and E. S. Lander. Searching for missing heritability: Designing rare variant association studies. *Proceedings of the National Academy of Sciences*, 111(4), Jan 2014.