Machine Learning Approaches for the Prioritisation of Cardiovascular Disease Genes Following Genome-wide Association Study

Hannah Nicholls

A thesis submitted to Queen Mary, University of London for the degree of Doctor of Philosophy

Statement of Originality

I, Hannah Nicholls, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Hannah Nicholls

Date: 21/11/22

Abstract

Genome-wide association studies (GWAS) have revealed thousands of genetic loci, establishing itself as a valuable method for unravelling the complex biology of many diseases. As GWAS has grown in size and improved in study design to detect effects, identifying real causal signals, disentangling from other highly correlated markers associated by linkage disequilibrium (LD) remains challenging. This has severely limited GWAS findings and brought the method's value into question. Although thousands of disease susceptibility loci have been reported, causal variants and genes at these loci remain elusive. Post-GWAS analysis aims to dissect the heterogeneity of variant and gene signals. In recent years, machine learning (ML) models have been developed for post-GWAS prioritisation. ML models have ranged from using logistic regression to more complex ensemble models such as random forests and gradient boosting, as well as deep learning models (i.e., neural networks). When combined with functional validation, these methods have shown important translational insights, providing a strong evidence-based approach to direct post-GWAS research. However, ML approaches are in their infancy across biological applications, and as they continue to evolve an evaluation of their robustness for GWAS prioritisation is needed. Here, I investigate the landscape of ML across: selected models, input features, bias risk, and output model performance, with a focus on building a prioritisation framework that is applied to blood pressure GWAS results and tested on re-application to blood lipid traits.

Acknowledgements

I would like to express my deepest gratitude to my supervisors, Dr Claudia Cabrera, Professor Michael Barnes and Professor Sir Mark Caulfield, for their support, knowledge and encouragement throughout my PhD. I could not have wished for better PhD supervisors, providing what would amount to hundreds of meetings (not including the countless ad-hoc meetings, coffees, and conversations), teaching me so much over the course of this project, and ensuring my supervision and learning opportunities were not disrupted during the pandemic. I would like to thank Professor Michael Barnes for providing me with insights that guided the development of my project and taught me how to see the bigger picture for my research. I would like to thank Dr Claudia Cabrera for her guidance, attention to detail, and all the knowledge shared over the past four years. Whether it was talking me through research approaches, explaining complicated technical concepts, providing detailed feedback, or advising on ways to troubleshoot problems, Dr Claudia Cabrera supported my development in every aspect. I am extremely grateful to my supervisors for giving me the opportunity to undertake this project with them.

I would like to thank the British Heart Foundation, for funding my research throughout my MRes and PhD, making this thesis possible.

I would like to thank my institution's blood pressure research group for their guidance. Particularly I would like to thank Dr Fu Liang Ng who provided expert clinical guidance that enabled me to curate training data, and Professor Patricia Munroe and Dr Helen Warren who provided collated BP GWAS data.

I would like to thank Professor Damian Smedley and Dr Julius Jacobson, who provided data from their Exomiser platform.

I would like to thank my computational biology research group, at the centre for translational bioinformatics, for every conversation had and all the thoughtful advice given. I would like to thank Dr Christopher John, who gave me several machine learning resources and writing tips that have been invaluable and that I know will continue to use beyond my PhD.

I would like to thank Dr David Watson, who gave his machine learning expertise that was invaluable to my research.

I would like to thank Dr Konrad Karczewski for the opportunity to complete a research placement with him at the Broad Institute. His insights and guidance throughout my placement refreshed my perspective and expanded my learning.

I would like to thank Dr Kirstie Whitaker and Dr Emma Karoune for the opportunity to work with them on the DECOVID project as part of The Alan Turing Institute. The chance to contribute to a collaborative project on a national scale provided me with invaluable skills and gave me an experience that was one of the highlights of my PhD.

I would like to thank Dr Keat-Eng Ng who took me as an undergraduate placement student. Not only inspiring me to pursue a PhD but also encouraging and supporting my application to my PhD programme, offering me support and guidance in what became one of the most formative parts of my education.

I would like to thank my family and friends whose impact has been unmeasurable. To my mum who has always given unconditional support. To my wife Daniela, who has been by my side for every challenge and every step forward, who has had unrelenting belief in my capability even when I have not believed in myself and who has always motivated me. None of this would have been possible without you.

List of Publications

Peer-reviewed

Olczak KJ*, Taylor-Bateman V*, **Nicholls HL*,** Traylor M, Cabrera CP, Munroe PB. Hypertension genetics past, present and future applications. Journal of Internal Medicine. 2021. PMID: 34166551 ***co-first author**.

Nicholls HL, John CR, Watson DS, Munroe PB, Barnes MR and Cabrera CP. Reaching the End-Game for GWAS: Machine Learning Approaches for the Prioritization of Complex Disease Loci. Frontiers Genetics 2020. PMID: 32351543

Cabrera CP, Ng FL, **Nicholls HL**, Gupta A, Barnes MR, Munroe PB, et al. Over 1,000 genetic loci influencing blood pressure with multiple systems and tissues implicated. Human Molecular Genetics 2019. PMID: 31411675

Submitted / in preparation

The Turing Way Community, Becky Arnold, Louise Bowler, Sarah Gibson, Patricia Herterich, Rosie Higman, Kirstie Whitaker. 2019. The Turing Way: A Handbook for Reproducible Data Science (Version v1.0.2). Zenodo. http://doi.org/10.5281/zenodo.6909298

DECOVID Consortium. DECOVID, a UK two-center harmonized database of acute care electronic health records for COVID-19 research – submitted to Scientific Data.

Nicholls HL, Ng FL, Watson DS, Jacobsen J, Warren HR, Cacheiro P, Smedley D, Munroe PB, Caulfield MJ, Cabrera CP, Barnes MR. Post-GWAS machine learning prioritizes key genes regulating blood pressure – in preparation.

Impact of COVID-19

The pandemic's impact on my PhD primarily affected travel constraints for placements, making all placements become virtual collaborations as opposed to in-person. However, it also enabled me to work on the DECOVID project as part of my placement with The Alan Turing Institute. This project consisted of collaborators from across UK universities and NHS hospitals, who aimed to store detailed and updated health data from hospitals in the UK and allow researchers to investigate the national occurrence of COVID-19 and more effective treatment strategies. During this project, I was responsible for helping to maintain and analyse collated hospital data. I provided support to researchers working across varying project aims, managed the project's database as a superuser, and built a GitLab repository to improve access for all collaborators. The expansion of skills I had on this project then enabled better data management for this thesis.

Table of Contents

List of Figures

List of Tables

1 Introduction

In this introductory chapter, I review the scientific literature and outline the key concepts that set the foundation for this research project. I give an overview of how genetic association studies are investigated in downstream analysis and how for complex traits, specifically cardiovascular diseases and blood pressure, these genetic associations are prime for machine learning prioritisation. I then provide an overview of machine learning concepts and how they have great potential to illuminate patterns in complex genetic data. The potential of this then defines how I conclude the chapter and lay out my research objectives, with the focus of this thesis being on developing a machine learning methodology that can be applied to prioritise genes that are most likely influential to cardiovascular diseases.

1.1 Genome-wide Association Studies

Genome-wide association studies (GWAS) investigate genetic variants across genomes, aiming to identify statistically significant variants associated with a disease or phenotype (Figure 1.1). Variants, also known as single nucleotide polymorphisms (SNPs), are deemed to be associated with a phenotype if they more frequently occur in individuals with the phenotype in comparison to a control population – and these associations are statistically significant if they reach genome-wide significance. However, significant association does not equate to causality alone and SNPs are also

confounded by linkage disequilibrium (LD). Linkage disequilibrium is where variants at different genomic regions (loci) are frequently inherited together more or less often than by random chance. In GWAS, LD creates a crucial obstacle as it clouds the causality of SNPs, for example, a SNP may be inherited with a truly causal SNP, leading to its statistical significance on association testing, yet it is not causal itself. To address this challenge, downstream analysis of associated SNPs is directed by functional investigation of the most likely causal variants driving the genetic association behind a phenotype, which aims to pinpoint molecular functions and pathways of interest for biological and translational investigation. However, the need to decipher causality and increase the certainty that a SNP functionally affects a disease - so that the truly most likely causal variants are being researched in follow-up experimentation - has led to the development of several post-GWAS analysis methods (fine-mapping, mendelian randomisation, network analysis, and most recently machine learning) that are a hot point of interest for determining causality.

For complex diseases, such as cardiovascular diseases (CVD), which affects tens of millions globally[1], the potential of GWAS to guide the translation of genomic findings is promising. CVD traits, such as blood pressure (BP), are one of the most powered examples of GWAS with recent studies genotyping over 1 million individuals[2]. However, as GWAS studies have scaled up to discover ever more disease variants[2-4] it has become impossible to perform a functional investigation on all disease-relevant loci to confirm real signals. With this being further affected by often high signal-to-noise ratios from these genes mapping to loci, which casts doubt on each gene's true

causality or effect size - presenting a challenge for developing clear and streamlined follow-up investigation post-GWAS.

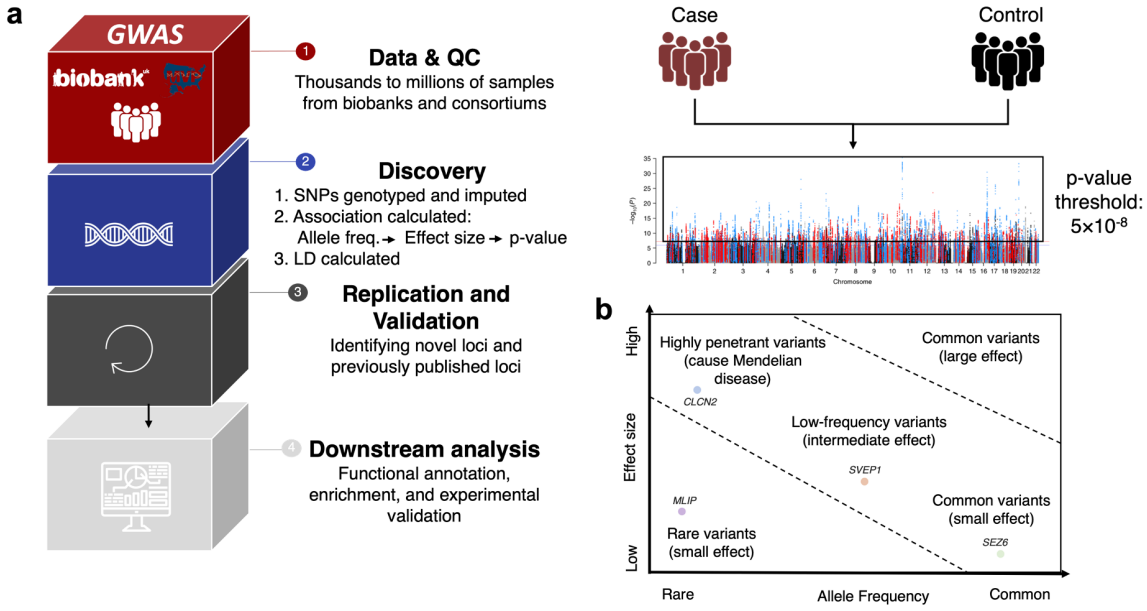


Figure 1.1 An overview of genome-wide association studies. a Genome-wide association studies (GWAS) genotype individuals and analyses their genetic variations to detect variants genetically associated with a trait/disease in a population. Genotyping methods involve either genotype arrays followed by imputation of single nucleotide polymorphisms (SNPs), or whole genome sequencing (WGS), and quality control (QC). Statistical association of genotyped SNPs then identifies regions of variation that associate with a phenotype with significance across the genome. Linkage disequilibrium (LD) is then calculated to identify alleles that frequently occur together more often than by chance, identifying causal variants that are not necessarily genotyped but are in LD with genotyped SNPs. The genetic variants are then functionally characterised, identifying candidates for experimental follow-up research. **b** Most causal variants identified by GWAS are common variants with small

effects, intermediate frequency variants with moderate effects, or highly penetrant mutations[5].

Several GWAS limitations impact post-GWAS analysis. For example, studies have variable GWAS reporting - differing from the gold standard approach of independently replicating results in another cohort – which creates questionable confidence in some discovered loci. Furthermore, as the GWAS sample sizes increase, the ability to corroborate results in matching independent cohorts becomes ever more challenging. A large number of associated genetic markers contribute to a multiple testing problem and call for a balance between stringent *p*-values to correct for false discovery and avoiding overly stringent corrections leading to false negative associations. These challenges are compounded by the need to differentiate causal variants/genes from other genes associated by LD, confounding the detection of causal genes within a locus. The LD then makes it unclear which variants and genes warrant further analysis and functional study.

These issues undermine the robustness of GWAS in the current landscape, and challenge the validity of downstream analyses and biological hypothesis development, critically opposing some of the major motivators for performing GWAS in the first place, such as target validation[6]. Ultimately this highlights the need for bioinformatic solutions to improve the signal-to-noise ratio of GWAS results and to triage variants and genes that are most likely to be causal.

1.2 Post-GWAS Analysis Methods

The definition of causality itself is a challenge. Causality is only truly known on experimental validation of a SNP's functional role in a disease[7], however, reaching this conclusion when associated variants are amassing in GWAS results is difficult and has become exponentially laborious. For example, whilst variants that cause monogenic forms of the disease are causal, polygenic presentations, wherein many variants or genes can have more subtle contributions to a phenotype to provide collective causality, make it hard to isolate individual variant contributions. Meanwhile, getting to the point of experimental validation that is needed to understand such variants is blocked by previously described obstacles (confounding LD, false-positive p-values, and lack of result corroboration in independent cohorts). Functional characterisation of genes and variants cannot overcome these challenges alone to ensure a genetic association has a biological effect on the phenotype. This is a key driver for developing post-GWAS analysis methods, which aim to prioritise variants and genes that are most likely to be causal – as defined by the strength of supporting biological evidence, with each method varying in the use and types of biological data that infer causality. Methods that identify genetic associations with high-quality biological evidence also provide a clearer biological definition of what might make a variant or gene most likely causal for a disease – leading to targeted hypothesis generation that can expedite experimental validation.

From the most commonplace methods downstream of GWAS (Figure 1.2) fine-mapping is often the method of choice. Fine-mapping is the use of statistical methods combined with orthogonal data (e.g., epigenomic data) to identify the causal variant(s) in a locus. Whilst fine-mapping can be a powerful tool due to its in-depth annotation of variants - with several methods, such as PolyFun[8], CAVIARbf[9], fGWAS[10], and PAINTOR[11], developed to take advantage of specific annotations available - it is reliant on LD, sample size and effect size. With the strongest association at a locus capable of being an artefact due to LD correlation, this casts doubt on using only fine-mapping to identify causal variants. The doubt is also amplified by fine-mapping treating each locus independently, and so the orthogonal information the method uses cannot account for biological relatedness between loci.

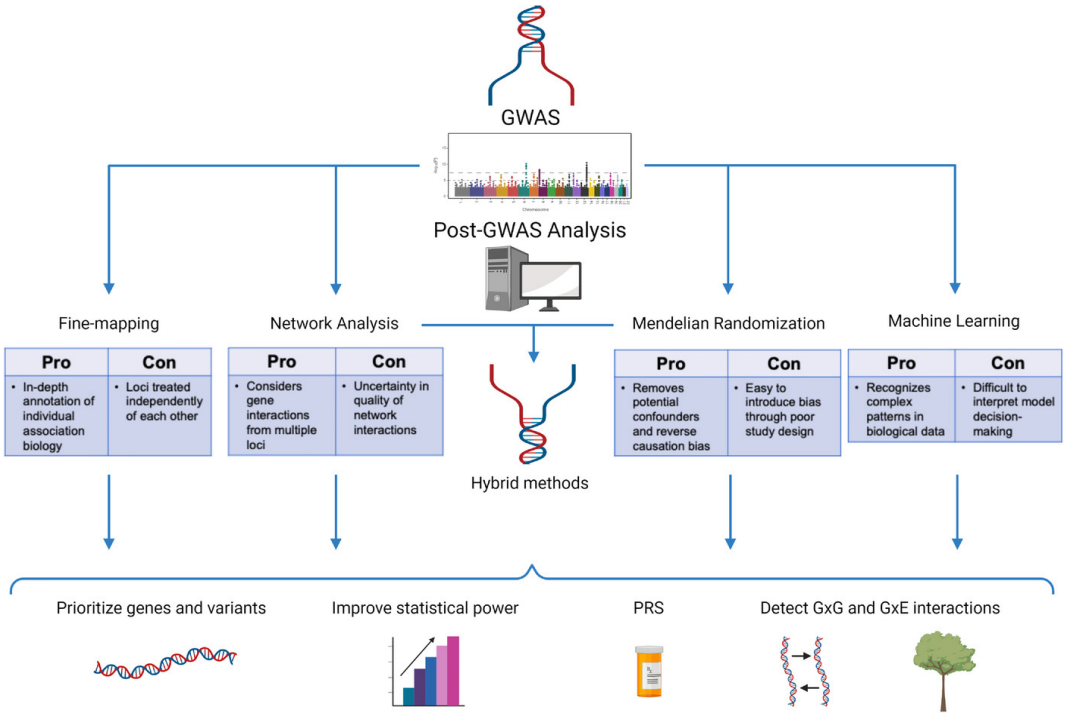


Figure 1.2. Common post-GWAS prioritisation methods[12]. Figure from Olczak et al. (2021) provides an outline of post-GWAS (genome-wide association study) analysis methods and their advantages/disadvantages. Each method can improve varying GWAS insights. For example, fine-mapping and network analysis for

identifying candidate disease genes, Mendelian randomization for PRS (polygenic risk scores), or machine learning for several applications including improving GxG (gene-gene) and GxE (gene-environment) interaction detection[12].

Another frequently used method is network analysis[13, 14], which uses interaction networks of associated genes to link them to known disease-causing genes or genes associated with similar phenotypes[15]. This method is particularly useful to identify genes playing a role in novel disease pathways, due to potentially unexplored network connections and can consider interacting causal variants/genes from multiple loci. However, a crucial caveat to network analysis performance is the quality of gene/protein connections (which can be based on text-mining or unstandardised database curation), where high-quality laboratory-validated data is few and far between – especially for non-coding genes.

Mendelian randomisation (MR), is another statistical method that uses genetic variants as instrumental variables to infer causality and has also been developed for post-GWAS prioritisation[16]. MR tests for causality between its instrumental variables (variants) and the exposure of interest (phenotype), with this in the case of GWAS aggregating variant estimates and applying a regression framework to understand a variant's impact on a phenotype. This method is advantageous due to its ability to control confounders and avoid reverse causation[16]. However, MR also assumes vertical pleiotropy - that the variant affects only the phenotype of interest - and ignores the capability of singular SNPs affecting multiple traits at once, which is common in variants associated with complex traits[17]. This disadvantage implies that MR is limited

when aiming to account for underlying biological relationships indicated by GWAS results.

1.2.1 Machine Learning for Post-GWAS Prioritisation

One method that has been generating attention across areas of research and has great potential in post-GWAS analysis is machine learning (ML). Machine learning algorithms build statistical models from training data to make predictions or decisions. Machine learning consists of supervised, semi-supervised, unsupervised, and reinforcement learning methods (Figure 1.3), with supervised and unsupervised learning being the most commonly implemented with GWAS data. Supervised learning provides ML algorithms with labelled training data and aims to infer a mapping function from the input variables to the output variable - or label for classification tasks. This mapping function may then be used to predict the labels of new 'testing' data. Unsupervised learning, by contrast, has no response variable. Instead, the algorithm must attempt to find patterns in the data, such as clusters or outliers.

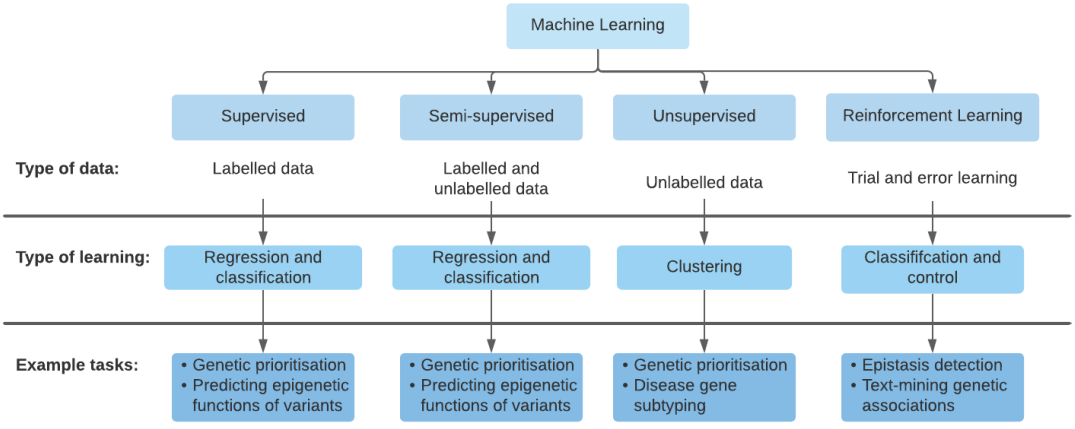


Figure 1.3. Types of machine learning classification. Machine learning is divided into several types (supervised, semi-supervised, unsupervised and reinforcement

learning) depending on the training data and problem. These methods are beginning to have an application to genetic data whether for gene/variant prioritisation[18, 19], prediction functions of non-coding variants[20], disease gene subtyping[21], or more recent approaches for epistasis detection or enhancing text-mining[22, 23].

Supervised learning has been applied to better understand GWAS data, from predicting variant function to prioritising associated genes, however, the models used within this approach have been variable in previous research[18-20]. The overall framework of supervised learning involves algorithmic principles (that vary from model to model) being applied to training data. The algorithm will then initialise its internal parameters – parameters that are defined by a model's algorithmic principles - and then iterate over that training data, making predictions on each iteration that allow the model to update its internal parameters and optimise itself (Figure 1.4). However, an important concept in ML that is prevalent in supervised learning is the risk of overfitting. This is when a model optimises itself too closely to the training data to have a high training performance but cannot then generalise its understanding and replicate its high performance on new test data. Techniques exist to combat this issue, such as cross-validation (dividing the training data into *k* folds with some folds withheld for testing model performance) and performance metrics focused on assessing overfitting (e.g., precision, recall and the F1 score in classification or the predicted r^2 in regression analysis). However, "some amount of overfitting is inevitable, but extreme cases can render a model useless"[24]. Recent studies applying ML to biological problems have also focused on addressing overfitting risk, e.g., by

developing bias-auditing[25], as it is crucial to ensure robust ML that gives reliable biological insights.

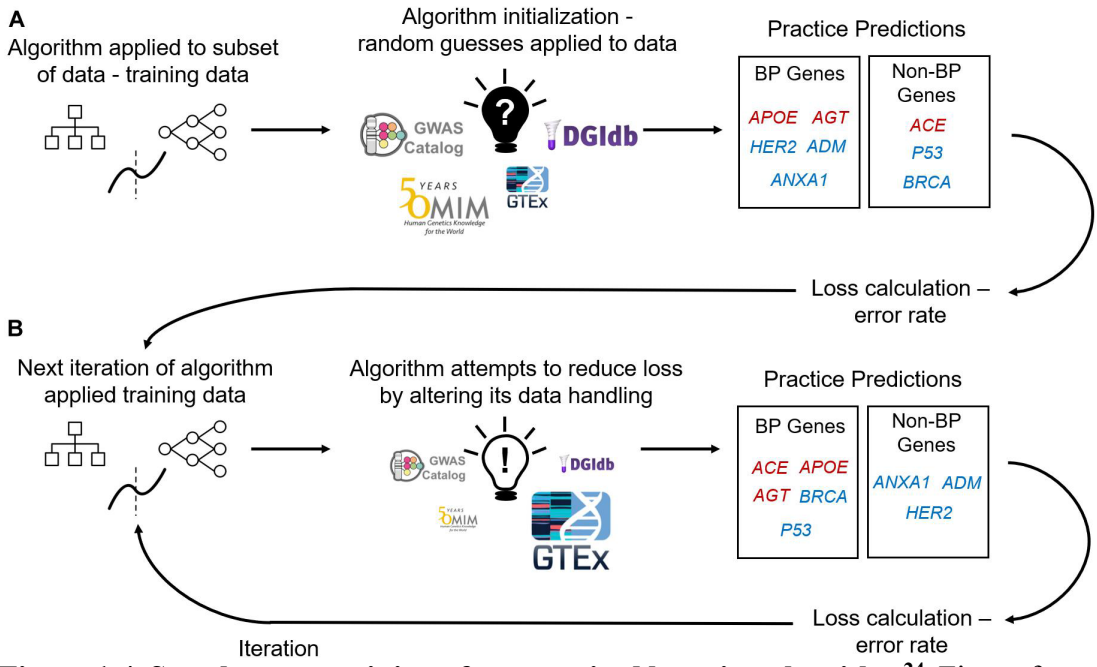


Figure 1.4. Step-by-step training of a supervised learning algorithm[24]. Figure from Nicholls et al. (2020). **a** Labelled training data (e.g., genes labelled as most likely causal or least likely causal for blood pressure – BP) and annotations of those genes are features input into a machine learning algorithm. A machine learning model initialises itself by algorithmic principles that are applied to the training data and its features at random. A model's first iteration can involve assigning feature importance at random (importance denoted by the size of the feature image). Based on this the model will then classify genes into either affecting BP (red genes) or not affecting BP (blue genes). These practice predictions can then be used to calculate loss (an error rate) and iterate over the data again by applying the previous iteration's loss calculation to adjust feature importance (**b**). By using the loss calculations, the model can improve its performance with each training iteration.

Alongside overfitting, a key aspect of ML for biological applications is the degree to which a model's decision-making is transparent and explainable. Whilst some models have an algorithmic design based on interpretability (e.g., explainable boosting machines) and some models have a degree of interpretation via their internal parameters (e.g., calculated internal feature importance or weightings), most models deal in opaque decisions. Explanability tools have been developed that use a model's inputs and its output predictions to interpret the model's decision-making - such as SHAP (SHapley Additive exPlanations) or LIME (Local interpretable model-agnostic explanations). Most recently SHAP has gained popularity among ML researchers, due to its ability to provide both local and global model understanding alongside its efficient application as a package with many visualisation options to view under the hood of a model (Figure 1.5). SHAP is based on game theory using Shapley values[26]. Game theory aims to understand the interactions of two or more players (ML features) that are involved in a strategy to achieve the desired outcome (model prediction). Shapley values are the average expected marginal contribution of one player (one feature) after all possible combinations (of features) have been considered. Using these values SHAP can calculate overall feature importance and each feature's influence on model-decision making, also doing so for every single data point[26].

Overall, SHAP's ability to give both global and local understanding of predictions makes it an ideal tool for interpreting ML applied to biological problems – which is necessary to justify a further functional investigation of any biological predictions that come from ML. This conclusion is also supported by recent research that has begun

to adapt SHAP to biological data, such as PoSHAP (positional SHAP) which modifies the explainability tool to consider peptide sequence positions in the ML predictions of protein binding[27].

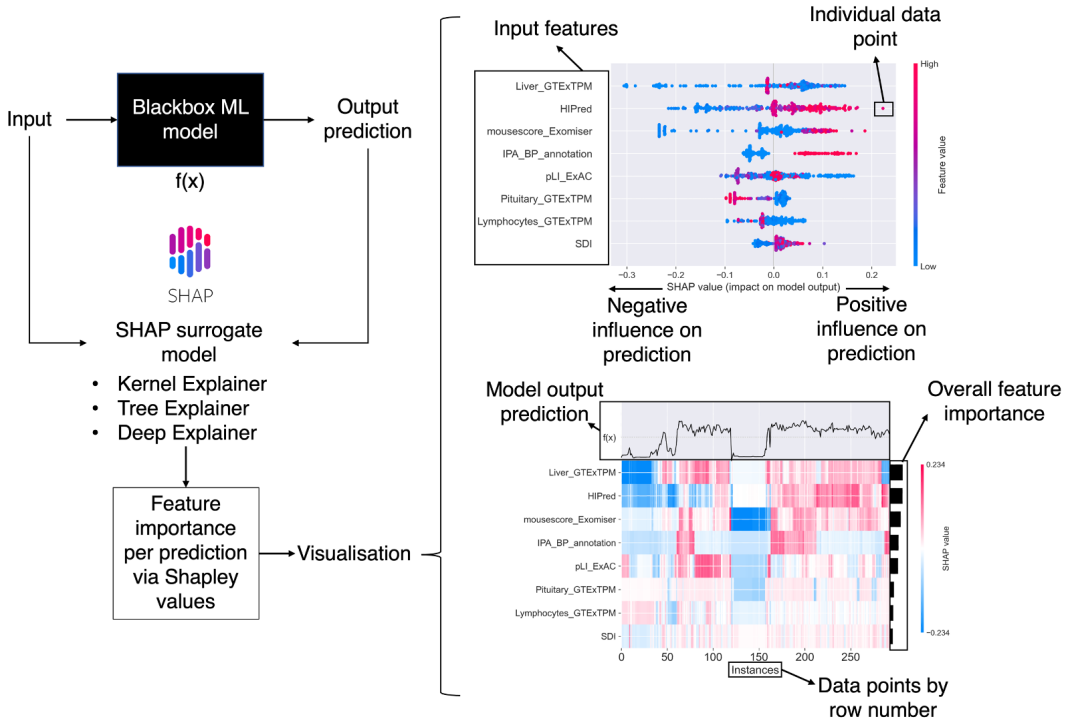


Figure 1.5. SHapley Additive exPlanation overview. A machine learning (ML) model's inputs and output predictions are used in a SHAP (SHapley Additive exPlanation) surrogate model that then uses Shapley values to calculate feature importance per each data point/prediction made by the model. SHAP then allows for visualisations giving both a local understanding of individual data points and a global understanding of overall feature importance. SHAP plots are colour coded by feature values or SHAP values, with both plotted in relation to the other to show how each feature influences decision-making for each data point.

Machine learning has had applications built in recent years to enhance GWAS performance and downstream interpretation[28, 29]. When tailored for understanding

GWAS data, ML predictions can provide an improved statistical foundation of evidence to support or improve GWAS results. For instance, ML in GWAS has been applied to identify loci, increase the statistical power of GWAS[30], detect epistatic interactions[31], improve polygenic risk scoring produced from GWAS[32], and prioritise genes and variants on post-GWAS analysis[33].

The development of systematic prioritisation post-GWAS using ML has been researched as early as 2007[34]. Since then, several computational methods for prioritising GWAS-associated loci have been developed with growing attention on ML applications[13, 29, 35]. Machine learning for prioritising GWAS results has used common models (Figure 1.6) such as logistic regression (LR), decision tree (DT) classifiers – e.g. gradient boosting machines (GBM) and random forests (RF)[36, 37] - and support vector machines (SVM)[33], and recent advances also including deep learning models[38, 39]. However, unlike other methods, whilst interest in ML applied to GWAS data is growing there has been little evaluation between methods, requiring further assessment of the reliability of ML for enhancing GWAS results.

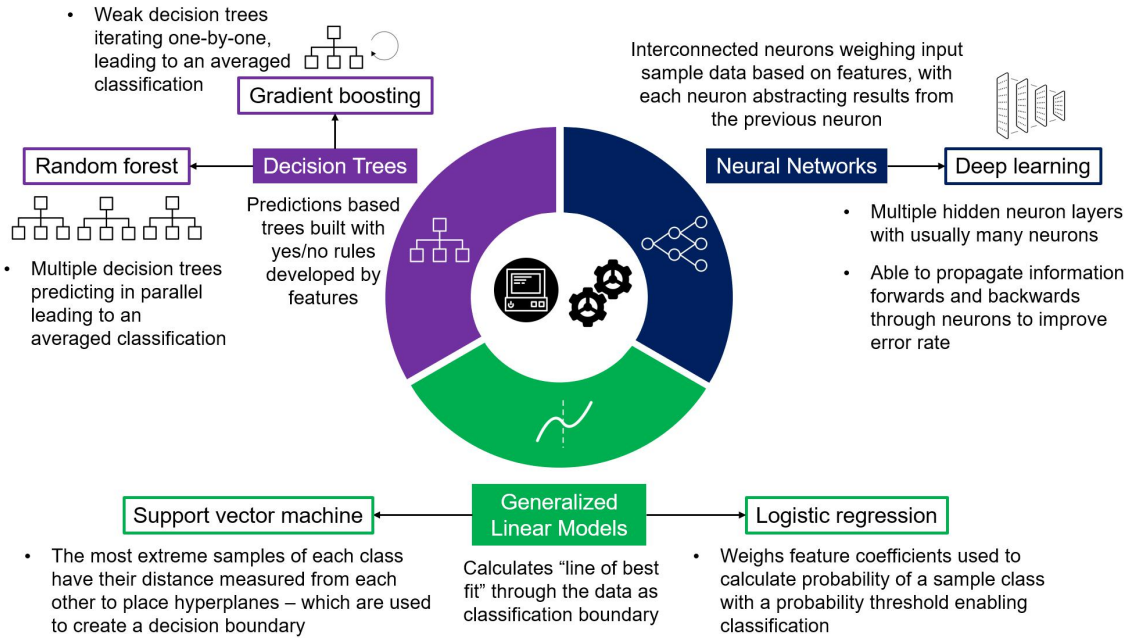


Figure 1.6. Commonly used supervised machine learning models[24]. Figure from Nicholls et al. (2020) showing the most common supervised learning approaches, each category has had varying algorithms applied to post-GWAS prioritisation[24].

1.3 Post-GWAS Prioritisation for Complex Diseases

The growth of GWAS over the past decade has identified thousands of associated loci, in November 2022 the NHGRI-EBI GWAS catalog contained 434,351 variant-trait associations from 6,096 publications[1]. Thousands of variant associations can now be found within a single complex disease or trait; such is the case for inflammatory bowel diseases (IBD) with 2,020 variant associations, schizophrenia with 4,988 variant associations, and lipid measurements with 53,236 variant associations[1]. Although, it

[1] https://www.ebi.ac.uk/gwas/

should be noted that there is a potential for overlap between these large association numbers, with multiple reports of variants from various studies. However, complex diseases with large GWAS investigations such as these are ideal candidates for ML application, as they potentially offer a large number of training samples and need methods like ML that can illuminate underlying novel pathways.

Individual GWAS loci have already shown the potential for large-scale prioritisation by providing novel biological insights and potential drug targets and drug re-purposing opportunities[40]. Evangelou et al. (2018) discussed 7 loci from their blood pressure (BP) GWAS that had genes with re-purposing potential, and the idea of modulating BP by re-purposing drugs has been discussed in other literature[41] but not further investigated. For example, Evangelou et al. (2018) found associations in the *SLC5A1* gene, which is also a drug target of a type 2 diabetes drug, canagliflozin, highlighting an opportunity for drug re-purposing to treat hypertension[3]. Researching re-purposing opportunities from GWAS associations such as this is ideal for complex conditions such as hypertension, where effective treatment frequently requires prescribing multiple antihypertensive drugs[42]. However, drug re-purposing for BP from genetic associations has had little research, with more generally only 12 of the associated genes by Evangelou et al. (2018) in BP clinical trials and none focusing on re-purposing[43].

Similarly, GWAS for lipid traits have amassed results with ample opportunity for analysis. Most recently Ramdas et al. (2021) developed a multi-functional analysis for a large multi-ancestry GWAS of five blood lipid traits - high-density lipoprotein

(HDL), non-high-density lipoprotein (nonHDL) low-density lipoprotein (LDL), total cholesterol (TC), and triglycerides (TG). This work incorporated information from several biological layers (gene expression, chromatin structure, and cell and tissue enrichment) alongside a variant prioritisation framework to not only identify the most likely causal disease variants and genes but use supporting information to suggest their underlying biological mechanisms at play in lipid metabolism[44]. For example, Ramdas et al. (2021) identified *RRBP1* (ribosomal binding protein 1) as having evidence at each layer of their functional analysis that validated the gene's potential role in lipid metabolism – from expression quantitative trait loci (eQTL) colocalisation in the liver to its variant having an open chromatin structure that interacts with the *RRBP1* promoter in adipose tissue. These results then converged with another study finding that *RRBP1* affects lipid homeostasis in mice, emphasizing how integrated evidence can streamline post-GWAS analysis. Another study by Kanoni et al. (2021) used the same multi-ancestry meta-analysis and combined established gene prioritisation methods: Polygenic Priority Score (PoPs), Data-driven Expression Prioritized Integration for Complex Traits (DEPICT), closest gene to the sentinel SNP, genes with coding variants in credible sets, eQTL localisation, and transcriptome-wide association study (TWAS). By combining prioritisation methods Kanoni et al. (2021) add another further integration of information to inform downstream analysis and increase the confidence in their output gene ranking. In comparison Ramdas et al. (2021) also prioritised *RRBP1* with high confidence. However, these results then require an in-depth functional study of such highly prioritised genes to truly validate their disease impact and push forward genetic research with translational benefits.

Understanding the functional impact of associated variants for complex traits is a challenge, with most studies also varying in their downstream approach post-GWAS, adding another layer of difficulty to assess best practices. However, the downstream functional analysis is subsumed by a greater problem in that differentiating variants and inferring causality is very challenging without further laboratory investigation. For example, BP associations have been found in several *SMAD* family genes and the *TGFβ* gene, which collectively participate in the TGFβ pathway, leading to the suggestion that these may affect sodium transport in the kidney and ventricular remodelling[3]. However, multiple genes impacting the same pathway raise the question of which gene should be functionally investigated first. Usually, the evidence is not strong enough to warrant laboratory investigation of all the associated genes in a particular pathway. The follow-up laboratory studies to date have developed without a standardised method for selecting causal genes consequently, and they are likely to be susceptible to personal or "cherry picking" bias. These issues highlight the need for a pipeline that methodically triages variants and genes based on their likelihood of affecting a trait. Only then will there be consistency in the follow-up of genetic results using functional analysis with minimised risk of investigating false positives or low-impact genes. The standardised *in silico* identification of the most likely causal genes at a genome-wide scale may be an opportunity to gain higher-level systems insights into complex trait biology. This in turn may help to fine-tune ML prioritisation algorithms, as seen with research using ML variant prioritisation as a feature fed into gene prioritisation[46].

1.4 Post-GWAS Prioritisation for Blood Pressure

Hypertension serves as a common denominating risk factor for complex conditions such as coronary heart disease and stroke. High BP with unknown aetiology (essential hypertension) dominates 90-95% of high BP cases and involves multiple organ systems contributing to the phenotype (Figure 1.7), yet 8-12% of hypertensive individuals show resistance to current treatments[47]. This need for improved understanding and treatment presents an opportunity for GWAS, which is unlocking novel insights into the genomic regions associated with BP at an unprecedented speed. However, the largely unexplored BP GWAS associations have untapped potential for functional insights via an optimised post-GWAS analysis approach.

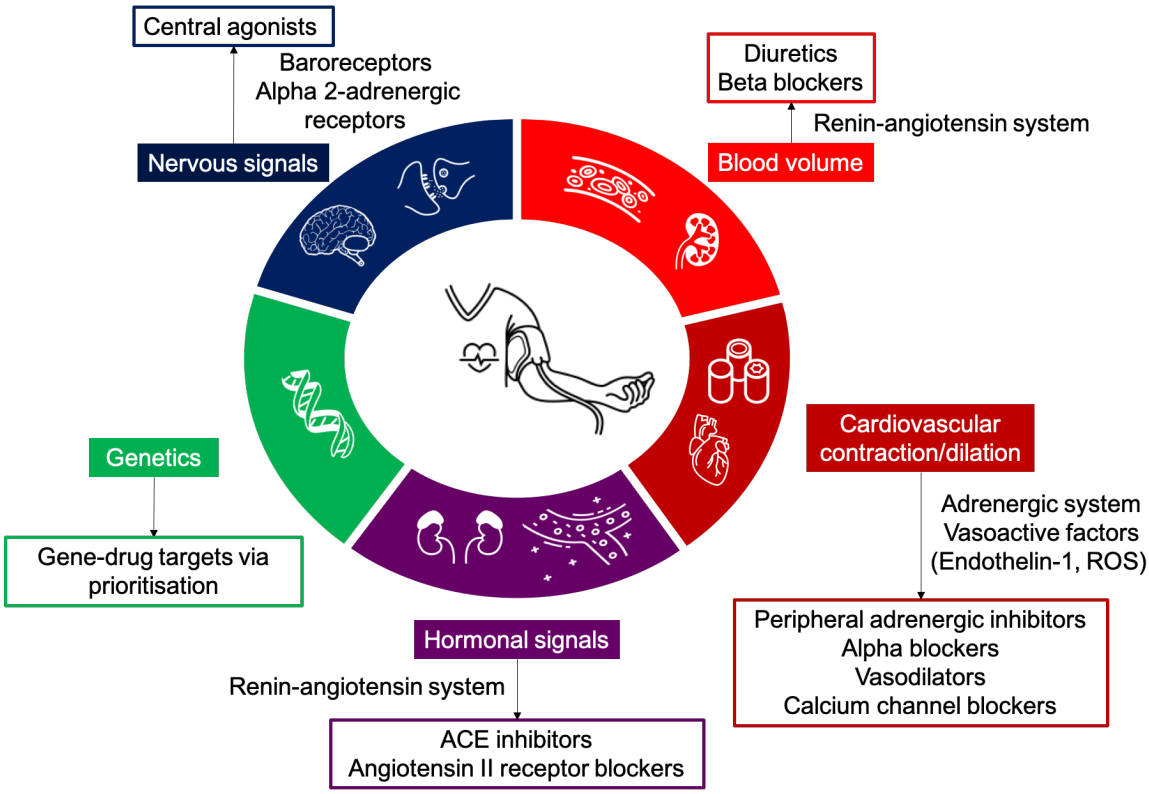


Figure 1.7. The biology of blood pressure. Regulation of blood pressure (BP) involves the interactions of several organ systems, which are predominantly cardiovascular, renal and neurological. These systems each have drug targets to regulate BP, however, the genetic component of BP is poorly understood in comparison and has the potential to develop more targeted treatments.

At present, the GWAS catalog reports 7,982 BP associations (with some possibly overlapping and not all being independent associations)[48]. BP is measured by three quantitative traits in GWAS - systolic BP (SBP), diastolic BP (DBP), and pulse pressure (PP). This growing number of associations holds great potential for biological insights and has already begun to bear fruit. For example, most recently a rare variant BP-GWAS in over 1.3 million individuals identified a missense variant (rs45573936)

in *SLC29A1*, affecting the function of equilibrative nucleoside transporter (ENT1), for which inhibitors have anti-cancer, cardio- and neuro-protective properties[2]. These findings suggest a potential for investigating ENT1 inhibitors for BP regulation. However, due to the signal-to-noise ratio from genes mapping to loci derived from variants in high LD, only a few genes have been identified as causal BP genes, providing an incomplete picture of the genetic role in BP that is needed to better treat hypertension.

The most commonplace approaches to post-GWAS analysis (such as fine-mapping or network analysis) provide an ability to prioritise associated genes and variants and guide functional research. However, their application to highly polygenic traits can be challenged by the assumption that most fine-mapping methods assume only one variant contributes to the trait per loci, or treat loci as independent, and can be affected by factors like effect size and LD. For traits such as BP surpassing these disadvantages is crucial, as finding novel drug targets and BP pathways requires nuanced investigation of associations across organ systems and an understanding of their underlying gene interactions. ML offers an alternative approach that has already begun to show promise for advancing the current understanding of BP. Mishra Manoj et al. (2020) provide a proof-of-concept for the use of ML and highlight its potential as a post-GWAS analysis tool for complex traits. They applied a deep learning model to gain an understanding of the functionality of non-coding BP SNPs – finding a higher than expected frequency acting in CTCF-binding regions and predicting that rs9337951 affected the secondary structure of mRNA from *JCAD*[20]. They followed this ML result with experiments transfecting the rs9337951 allele in kidney cell lines,

which showed decreased JCAD expression in cells with the allele, validating their model's prediction[20].

An increasing number of studies are investigating how ML can be tailored to prioritise genetic associations across diseases[19, 33], but the ML pipelines for GWAS prioritisation are mainly non-benchmarked and case-specific applications[49, 50]. In this thesis, I investigate the application of ML for post-GWAS analysis. I aim to benchmark an optimised approach to BP GWAS, based on testing ML approaches (multiclass classification and regression analysis) and ML models (tree-based ensemble models, generalised linear models, and deep learning). With the further objective of creating a ML framework that is re-applicable to other phenotypes to prioritise genes most likely influencing cardiovascular traits. Additionally, whilst I focus on gene-level prioritisation, I also aim to investigate variant-level prioritisation. Overall exploring how ML can aid in reaching the endgame for GWAS by developing an evidence based post-GWAS analysis approach.

2 Exploratory Analysis of Blood Pressure GWAS Data

2.1 Introduction

To investigate how ML can be applied to the findings of BP GWAS, careful data curation and feature quality control are needed to achieve the best possible performance. In this chapter, I explore the integration of a range of multi-omic databases and their potential as features to be used in ML, alongside the curation of a subset of training genes from an individual GWAS. This exploratory data analysis provides an overview of the training data that is then used in chapters 3 and 4. The data is curated for ML to prioritise genes that are most likely to be causal in BP traits, with causal genes being defined as those that are likely to contribute to BP - as recognised by ML outputs that are trained on curated genes with known BP roles.

2.1.1 Biological Annotations

GWAS associations are typically annotated to a wide range of biological annotations from growing multi-omic databases, which contain data that can be applied to ML. Biological features range from eQTL (expression quantitative trait loci), RNA, epigenetic, and protein data to describe a variant or gene's functionality. The growing integration of related biological features suggests they will provide a clearer insight for models to be able to pinpoint the most likely disease-causing genes in a locus[51, 52]. In the past decade, this integration has been built upon by databases such as the Genotype-Tissue Expression (GTEx) project[53]. GTEx collects genetic variation data, with several annotations such as transcripts per million (TPM) - provided at the gene-level as a median expression measure - for genes across 53 human tissues and has been

used in data for GWAS prioritisation models[19, 33]. Well-curated databases such as GTEx benefit ML as they offer a point of direct comparison between studies taken from the same resource, and the range of tissue types available also adds to the ability of models to identify genetic roles in tissues that are not a known site of action for a phenotype of interest. GTEx is also consistently updated and as it continues to grow, it will provide a standardised source of expression data to optimise how models use gene expression in their predictions.

The use of other biological features, e.g., RNA and epigenomic features, has also grown in recent years. These features may provide further insights into associated loci located in non-coding regions. For example, Mishra Manoj et al. (2020) used deep learning to predict functional annotations of BP-associated SNPs, using several features: CpG islands, CTFC binding sites and conservation, enhancer sites, regulatory element activity, microRNAs, mRNA splicing sites, and lncRNAs among others. From collecting these datatypes, the researchers gained an understanding of the SNP's relationships to each feature in turn offering insight into the binding of transcriptional factors, enhancer function, chromatin conformation, mRNA splicing, or the function or expression of regulatory RNA[20]. These researchers found from their model that BP-associated SNPs were more likely to be located in CTCF-binding regions and that the SNPs were more likely to affect CTFC binding[20]. Such results show how ML combined with optimal feature collection can be used to illuminate the functions of non-coding SNPs. As understanding of the non-coding genome improves, this type of information can fit into a prioritisation pipeline to streamline GWAS results in non-coding regions.

There are also several underused features in genetic prioritisation such as gene essentiality scores, loss-of-function (LoF), and haploinsufficiency measures. Gene essentiality measures aim to capture a gene's importance to reproductive success, creating metrics that impact evolutionary and systems biology that can aid in drug development by identifying genes less likely to cause drug resistance[54]. However, to the best of our knowledge, only the Mantis-ml model by Vitsios et al. (2020) uses gene essentiality, only looking at this datatype in mice. This highlights, that whilst many emerging biological datatypes are becoming commonly used in ML, there is still room for improvement. The lack of gene essentiality features is also likely in part due to fewer databases with small-scale analysis, and most databases only offering measurements for protein-coding genes. Jia et al. (2020) recently developed a subcellular diversity index (SDI) to quantify gene essentiality, which focused on protein localisation to understand gene essentiality via Gene Ontology (GO) terms[55]. Beyond developing the SDI score they also used that measure to calculate a drug interaction probability per gene[55], which could potentially be used to develop clinical insights from ML prioritisation. However, this SDI score is limited to their annotated genes that are mostly protein-coding.

Loss-of-function mutations are also crucial to understand as they illuminate molecular functions directly impacted by the mutated genes, holding great potential for new drug targets[56]. When focusing on LoF measures, metrics often provide variant-level scores[57]. However, a probability of loss-of-function (pLI) score was developed by Lek et al. (2016) that calculated the measure for protein-coding genes in humans using

exomes from the Exome Aggregation Consortium (ExAC). Lek et al. (2016) found their pLI metric outperformed other measures[58]. However, it should be noted that a high probability or predicted measure of loss-of-function does not guarantee the variant is lethal and this needs to be considered when analysing how a ML model interprets LoF scores. Recent research has analysed the human LoF in exomes with supporting whole-genome information to gain a better understanding of LoF variants in the context of the whole-genome[56]. This work has also been followed by research focusing on LoF impacts on untranslated regions of protein-coding genes[59], providing novel insight into how LoF variants may impact the regulation of protein-coding genes to alter their function. Such work provides a roadmap for future human knockout investigation[56] that is beneficial to population studies dedicated to understanding LoF variants (such as East London Genes & Health, which focuses on British Bangladeshi and British Pakistani people in east London[60]) and the identification of new drug targets.

Another example that investigated both human and mouse LoF is by Cacheiro et al. (2020) who developed a comprehensive gene essentiality measure - Full Spectrum of Intolerance to Loss-of-function[61] - that has subcategories depending on the organism and cellular viabilities (based on mouse phenotyping screens and human cell lines). This measure provides greater opportunities for ML with both human and mouse data; however, the analysis was only for 4,000 genes requiring further investigation across the genome. As studies such as these increase in scale, gene essentiality is likely to become more frequently used in ML and may offer insight into how essential genes

that do not tolerate loss-of-function interact with other biological characteristics to affect complex traits.

Haploinsufficiency scores (a measure of gene function when there is only one viable gene copy) have had little to no use as a feature in ML. This may be due to their previously limited curation that relied on protein-protein interactions (PPI) networks[62]. Shihab et al. (2017) developed HIPred to measure haploinsufficiency by ML, aiming to score both well-studied and lesser-studied genes with no bias towards those in PPI networks[63]. They benchmarked HIPred against other haploinsufficiency measures (RVIS, EvoTol, GHIS, HIS and IS) finding an improved performance on several human and mouse model datasets, and that HIPred outperformed methods incorporating PPI data when scoring less-studied genes[63]. These results suggest scores provided by methods such as HIPred can enable more comprehensive features that can aid models prioritising genes across the whole genome.

2.1.2 Pathogenic Features

Alongside general biological characterisation, disease-specific data is gradually increasing. Vitsios et al. (2020) for example prioritised chronic kidney disease genes, using annotations from the Chronic Kidney Disease database among their features to improve stratification. This implies the potential of future models to take advantage of similar disease databases, with databases such as DisGeNET[64] or phenotype-specific resources (e.g. AutismKB [65] or CardioGenBase[66]) currently being untapped for ML prioritisation in published work beyond Mantis-ml[33]. There are also text-

mining tools that will provide data on gene significance in publications related to a specific disease. However, this is subject to bias within the text-mining, depending on the methods used and the reliability of the source material being text-mined. A stronger phenotypic measure, however, is from curating gene-drug interactions directly, which can be curated from databases such as the Drug Gene Interaction Database and the British National Formulary – although, this requires time and effort with clinical validation to ensure correct drug interactions are being recorded. Non-public databases such as Ingenuity Pathway Analysis provide disease-specific data (such as molecule and pathway associations to a disease) which is stringently curated and could provide powerful insight for a model, but the lack of accessibility to such databases then creates a re-usability problem. There are also public disease-specific prioritisation tools that provide variant-level information that can be abstracted to the gene-level (e.g. Exomiser, which prioritises disease-causing variants using clinical and animal model data, offering a disease-specific gene score with curation[67]). Furthermore, the GWAS catalog provides publicly available disease-specific information (such as p-values for phenotypic associations), however, this returns to the caveats of GWAS with potential false-positive risk in collected data.

2.1.3 Variant-level Annotations

Variant annotations are also used for prioritisation (e.g., algorithmic scores such as Eigen, CADD, DANN, GWAVA, REVEL, DeepSea). These scorings predict the pathogenicity of variants based on their expected functional consequences and have been used as ML features in variant prioritisation[46]. Such measures are also supported by variant annotation databases that are improving data quality. For instance,

ENCODE which has investigated ~98 % of the non-coding genome[68], or the Exome Aggregation Consortium (ExAC) which has analysed 125,748 exomes and 15,708 whole genomes in diverse populations[69]. Databases such as these create a wealth of data from which offshoot tools are also able to provide more potential features (e.g. HaploReg which identifies SNPs of epigenetic interest from ENCODE data[70]). This presents a lot of descriptive features for variant prioritisation but also suggests these scores could be useful when collapsed at the gene-level. However, collapsing variant information has been performed by varying methods with no standardised approach. For example, Kolosov et al. (2021) developed a disease-agnostic gene prioritisation method (GPrior) and they used variant annotations (functional annotations and GERP pathogenic scores). To use these features, they took both mean and median scores per gene, selecting either summary measure depending on the feature – e.g., for GTEx tissues median is preferred due to potential data skewing of gene expression. In comparison Khan et al. (2018) do not define how they condense variant annotations to their genes for prioritisation, however, GPrior provides a standardised procedure that can add reliability to future work. Furthermore, collated scores are being developed, such as seen with the GenePy tool which amalgamates 16 pathogenic variant scores for one gene-level score[71], which could be re-purposed as a ML feature.

2.1.4 Feature Importance and Feature Selection

Beyond data collection, studies also need to consider feature importance and feature selection to gain an understanding of model decision-making – ensuring that the features provide relevant information and preferably meet i.i.d (independent and identically distributed) assumptions. This is in part due to an encompassing problem

for all of ML of having many features and few samples (the “curse of dimensionality”) and the need to reduce features to increase computational efficiency. This is often a part of why researchers choose L1 regularised logistic regression, which automatically performs feature selection. Several prioritisation studies have used logistic regression, such as Isakov et al. (2017) using the elastic net, who found positive feature coefficients (predicting causal genes) were highest for immune and inflammatory response features from GO. Gettler et al. (2019) also used logistic regression – as part of their gene prioritisation regression model (GPRM) - to prioritise genes for Crohn's disease. While Gettler et al. (2019) do not discuss the impact of feature importance, they note that GO enrichment analysis showed immune and inflammatory genes were significantly enriched. This enrichment is to be expected from an autoimmune disease, however, it also suggests validation for the feature importance found by Isakov et al. (2017). Maciukiewicz et al. (2018) applied L1 logistic regression to identify significant features and followed up with SVM for predicting causal variants for duloxetine response in major depressive disorder. They found a non-coding RNA annotation had the largest positive coefficient. However, unlike the study of IBDs[50], Maciukiewicz et al. (2018) is the first prioritisation study to focus on their drug response phenotype, requiring further work to validate feature importance and to suggest how these important annotations may fit into the biological understanding of GWAS results.

Additionally, besides using a model’s internal feature weightings, several other methods can provide feature importance. Permutation is also able to provide feature importance, doing so for any model by shuffling feature values and viewing the model

error rate. Vitsios et al. (2020) use permutation via the boruta algorithm, which creates synthetic features from a random permutation to weigh the importance of original features and remove any unimportant annotations. However, permutation has been noted as disadvantageous for dealing with correlated features[73]. There are also explainability tools that give feature importance that can be used for selection, such as SHAP. Using SHAP alone for feature selection, however, has a risk of selection being influenced by the model that the SHAP's inputs came from (incorporating the model's bias into feature selection). However, SHAP has been combined with the boruta algorithm for selection (in a tool named BorutaShap), creating an optimised feature selection method where SHAP feature importance is considered in combination with testing randomised features[74].

2.1.5 Bias in Biological Characterisation

Bias within features and their selection, and throughout a ML pipeline, is also a crucial area that needs exploration relating to GWAS prioritisation. Due to the nature of biological experiments, their collected features are highly likely to contain artefacts or experimental noise that needs to be factored into ML. Also, a lot of databases develop annotations from a restricted number of cell lines (e.g. ENCODE[75]) or predominantly male model organisms, and specifically for GWAS, the majority of individuals genotyped are white Europeans[76]. These examples highlight just some of the biases in biomedical data that will be amplified on ML and risk harm further down the pipeline as false-negative or less impactful genes may be prioritised. Database curators are starting to account for biased data. In 2020 GTEx published sex-biased data, identifying "*that 37% of genes in at least one of the 44 tissues studied exhibit a tissue-*

specific, sex-biased gene expression"[77]. For ML applications, this publication offers one example of bias that can be tracked and audited on output prioritisation, as several prioritisation methods use GTEx as a source for features[19, 33].

Some elements of feature bias, such as noisy features can be initially addressed with data cleaning (e.g., removing features that are heavily missing values for most genes/variants or removing correlating features with redundant information). However, the exact protocol for data cleaning is often case-specific, requiring a user to set thresholds (e.g., the percentage of missingness before a feature is removed, or a correlation threshold to remove highly correlating features) and risking either including bias or a loss of useful information. In the case of correlation, Guyon et al. (2003) note that if features are not perfectly correlating they may provide non-redundant information that a model can use, suggesting only extremely highly correlating features should not be removed. On the other hand Darst et al. (2018) show correlating features providing redundant information can be particularly harmful to certain models, requiring feature pre-processing that is model-dependent. Random forests for example are affected by correlation as correlation can mask the interactions of other features[79]. In contrast, models that assume feature independence, such as Naïve Bayes, can use correlated features with less risk[80]. Overall, this highlights that methods for cleaning correlated features are case-dependent and will likely incorporate biasing features if one method is used for several models without care. Another risk with data cleaning is the degree of missingness allowed in the data. Vitsios et al. (2020) set a 25% missingness threshold before removing features[33]. This enables the model to predominantly learn from real-world data. However, for several

biological features missingness is prevalent, e.g., many features account for only protein-coding genes and so non-coding features have missing data that cannot be given to an ML model for reliable prediction. Vitsios et al. (2020) address this by prioritising only the protein-coding exome. Overall, the risk of biased features requires investigation in any ML pipeline, with cleaning methods tailored to the specific problem, and needs to be developed with standardised practices for use cases of ML in biology.

How biological characteristics interact with one another is also a key bias point that is regularly unaddressed in ML studies. One example of this is gene length, which has been used as a standalone input feature in models[33], or as a factor in calculating new biological features (e.g., DNA scores such as GenePy[71]). Gene length is known to create artefact correlation where longer genes have more opportunities for measurement signals due to their increased length[81]. It has been established for techniques such as RNA-sequencing that adjusting for gene length is necessary to avoid bias[53], yet this is unexplored in ML and how model decision-making may be impacted. Lopes et al. (2021) investigated the impacts of gene length on gene function, finding gene size correlates with several traits (transcript length, intron count, protein size, the type of tissue the gene is likely to be expressed in, natural selection suppression, gene co-expression, and PPIs). Each of these traits has been used in ML models prioritising genes[19, 33], however, the relationship between such features and gene length is not always accounted for.

Studies that do investigate gene length relationships, and whether to use gene length as a feature or in creating other features, also vary in their methods. The development of HIPred measurements accounted for gene length bias, following a method laid out by previous work which identified other haploinsufficiency scores, namely RVIS, as biased by gene length[62]. This method involved comparing HIPred analysis with an equal number of random genes based on matched coding sequence length[62]. Vitsios et al. (2020) use gene length as an input feature, regulating it by prioritising protein-coding genes and taking the median length of exomes per gene. This approach reduces bias and also by having gene length as a feature this enables the model to identify the relationship gene length may have with other features, with models being able to identify correlating features and reduce their weight in prediction. Clark et al. (2007) accounted for gene length bias when investigating patterns of SNPs in *Arabidopsis thaliana*[83]. They found no significant relationship between gene size and observed variation patterns. However, this study was conducted in 2007, when SNP research was in its infancy, suggesting the relationship between gene length and SNPs would likely be different on re-analysis. The inclusion of gene length correction in studies as early as Clark et al. (2007) do however provide a precedent for gene length correction in post-GWAS analysis as a standard practice that will allow for better interpretation of prioritisation results.

Gene length also needs to be considered not only for features but in the curation of training genes. This is particularly true for studies which develop gold and negative standard genes using biological information influenced by gene length, such as OpenTargets using PPIs. Gene length has a relationship with PPIs which has been

shown by Lopes et al. (2021) who found shorter genes more frequently had zero recorded PPIs. Meanwhile, OpenTargets identified PPIs, using them as a sign of gold standard positive or negative genes depending on if a direct PPI with a disease-causing gene was present, presenting an opportunity for bias with smaller genes being more likely to be labelled as negative. However, whilst Mountjoy et al. (2020) publicly provide OpenTargets gold standard positives (with the smallest gene being *OR10G3* at 941 base pairs and the largest gene being *RBFOX1* at 1,694,245 base pairs, and all genes being protein-coding with a mean size of 85,65 base pairs) they did not publish their list of their gold standard negatives for comparison. Additionally, Lopes et al. (2021) note short genes may become more refined in their definitions in the future (e.g., some short genes may be due to annotation errors), showing it is important to repeatedly consider the impacts of biological characteristics, such as gene length, as data quality increases.

Protein-protein interactions themselves are another biological factor that needs to be regulated. Machine learning studies using PPIs vary in their criteria for what constitutes a trustworthy interaction. For example, OpenTargets counted PPIs with a confidence >0.7 in the STRING database[84]. Setting high confidence is a double-edged sword as it allows assurance in your interactions, but true interactions may be missed if they were measured at lower confidence – risking training genes that are false positives/negatives. Mountjoy et al. (2020) note that they removed negatively labelled genes from their training data if they had a >0.7 confidence PPI with a positive labelled gene, with this removing 229 out of a total of 9,171 negatively labelled genes[84]. This was, alongside genes not being within 500kb of loci, the only defining criteria for

negatively labelled genes, creating a great risk of bias in part due to the large class imbalance versus 445 positive genes, but also due to the nature of STRINGdb's protein data. STRINGdb offers several interaction measures (co-expression, gene fusion, gene neighbourhood, text-mining, experimental data, and pathway database knowledge) which it gives in a combined default score[85]. However, the inclusion of several of these datatypes risks false positive/negative interactions, especially for text-mining-based interactions collected from PubMed abstracts[85]. A possible way to minimise bias by using PPIs for data labelling, or also for its use as a feature in other studies, would be to collect interactions using the least biased measures available – e.g., using only the experimentally recorded interactions – or implementing bias-reducing methods tailored to protein interaction data as they develop[86].

An optimal ML model hinges on data size and quality for reliability and performance, but for genomic applications, extra considerations are needed to avoid any biological biases impacting model learning. In this chapter, I focus on feature curation and exploratory data analysis of BP GWAS data to curate bespoke training data for post-BP GWAS prioritisation using ML.

2.2 Methods

2.2.1 GWAS Description

The BP GWAS analysed in this research project was performed by Evangelou et al. (2018) for three BP traits (SBP, DBP and PP) in individuals of European ancestry. They used cohort data from the UKBiobank (n=458,577), International Consortium

for Blood Pressure (ICBP) (n=299,024), Million Veteran Program (MVP) (n=220,520) and the Estonian Genome Center, University of Tartu (EGCUT (n=28,742)). This study used the UKBiobank and ICBP data (with a combined n of 757,601 used in meta-analysis) for discovery. In replication, they had a one-stage analysis (using combined UKBiobank and ICBP data) and a two-stage analysis using the MVP and EGCUT data (in total providing a combined n of 1,006,863 in meta-analysis). For the UKBiobank GWAS, the data underwent imputation from the Haplotype Reference Consortium (HRC) panel followed by an additive genetic model for all three BP traits. For the ICBP GWAS, the SNPs were imputed from the 1000 Genomes project or HRC panel and quality control previously reported in 72 combined cohorts was used. Post-quality control and summary effect sizes were calculated for ~7 million SNPs. By identifying genome-wide significance at p-values $< 5x10^{-8}$ this analysis found 535 novel BP loci, confirmed 92 previously reported loci and found support for all 274 BP loci previously published before this study.

2.2.2 Data Collection

In total 20 databases were downloaded, providing 114 features that entered data pre-processing (Appendix A Table 1). Overall, these databases aimed to comprehensively describe the genes across various categories, such as gene expression, epigenetics, genetic functionality, pathogenicity, gene essentiality, and phenotypic measures specific to BP.

The GWAS summary statistics for three BP phenotypes (SBP, DBP and PP) were collected to be used as feature input[47]. To collect gene-level data bedtools (v2.28.0)

was used to map variants to the hg19/GRCh37 reference genome from Ensembl (release 92, Homo sapiens.GRCh37.87). A gene was assigned to a variant if the variant was within a 5kb window distance from the start and end of transcription of the gene. All variants in the GWAS (n=7 million variants) were annotated to genes within 5kb. These genes were then divided into two groups:

1. *BP-genes*: collated from the genes annotated to the variants in high LD (r^2> 0.8) with a lead BP-associated variant (from the 901 reported loci by Evangelou et al. (2018)) and a curated list of previously reported loci in other studies (including various ancestries, rare variants and gene-environment/lifestyle interaction GWAS (n=2,004))[2, 4, 87-89].
2. *Non-BP genes*: not in LD, p-value >0.15, not within 500kb +/- loci and no direct or secondary protein-protein interactions with BP genes.

The *BP-gene* and *non-BP* gene groups underwent annotation from all collected databases, with their grouping divisions used at a later stage to curate the training data (further details in section 2.2.3).

Variant annotation was performed using ANNOVAR (2018Apr16)[90] – providing 10 features detailing pathogenic scorings and epigenetic information. Variant annotations from other databases (UCSC and DeepSEA) were also collected, alongside the variant-level GWAS summary statistics (beta values for DBP, SBP and PP). These variant annotations were then collapsed to the gene-level, grouping variant annotations by their genes and selecting the most significant variant scores per each gene for each feature. For the beta values of each variant, the maximum absolute value between all

three BP phenotypes' beta values (SBP, DBP and PP) was selected while also retaining a positive or negative direction of effect for model interpretation, creating one final and singular beta feature. This beta value pre-processing step was enabled by Evangelou et al. (2018) using three sets of GWAS summary statistics, one for each of the three BP traits assessed by their study.

All other databases provided gene-level annotations that could be directly merged into the main dataset. Whilst Exomiser usually provides a variant-level prioritisation, gene-level scores were curated by experts – identifying Exomiser scores for genes under increased blood pressure and hypertension HPO terms that were divided into three features (human, mouse and fish Exomiser scores).

2.2.3 Training Data

Genes used in the training data were assigned into one of four categories for label-based classification from most to least likely to impact BP. The previously identified *BP-genes* group had three gene groupings subset from it with each gene group assigned a label:

1) Genes were labelled as *most likely* BP genes if they were known to interact with and be involved in BP mechanisms, having BP-regulatory roles, curated by an expert in the field (Appendix A Table 4).
2) Genes were assigned a *probable* label if they were evaluated by a text-mining tool (Génie[91]) as having an adjusted p-value <0.01 in relation to its significance

in BP in publications; or genes were also labelled as probable if there were known to interact with drugs that had BP side-effects as identified by an expert.

3) Genes were labelled as *possible* BP genes if they had annotation in Ingenuity Pathway Analysis (IPA) relating them to BP from experimental analysis within the IPA database.

Any of the *BP-genes* that did not meet any of the criteria to enter the training data in these three categories were reserved as genes to be predicted by the trained model (n=1,804).

2.2.3.1 Least Likely Blood Pressure Gene Curation

From the *non-BP genes* group, the fourth and final group of training genes were labelled as least likely to affect BP. These genes underwent several curation steps to be identified:

1) The genes had no variants with a p-value <0.15 across the gene in the whole BP GWAS consisting of all 7 million variants (with multiple p-value thresholds tested on ML performance to select this 0.15 threshold, detailed further in section 2.2.3.2).
2) The total gene length was outside a +/-500kb window of any BP loci (across several collated BP GWAS with >1000 loci[2, 4, 87-89]).
3) The genes had no evidence of LD ($r^2 < 0.1$) with BP SNPs (SNPs defined as having $r^2 > 0.8$ LD with a sentinel SNP across the collated BP GWAS[2, 4, 87-89]).

4) The genes had no direct or secondary PPIs with *BP-genes* across collated BP GWAS[2, 4, 87-89] (3,786 genes with variants in high LD, r^2> 0.8, across studies). The PPIs were counted from STRINGdb, using only PPIs measured from experimental data only with a >0.15 confidence threshold (multiple thresholds tested alongside the p-value threshold in step 1) to filter genes that give the best ML performance, detailed further in section 2.2.3.2.

After PPI filtering this resulted in 93 genes that passed all five steps and ML testing to be assigned as genes *least likely* to affect BP.

These four groupings provided 377 training genes in total (51 genes labelled *most likely*, 149 genes labelled *probable*, 84 genes labelled *possible*, and 93 genes labelled *least likely* to impact BP) (Appendix A Tables 5 and 6). The training data categories were further investigated by the possible gene group being either kept in the training data to give a 4-label dataset (n=377) or removed from the training data creating a 3-label training dataset (n=293). The 3-label gene group tested model performance with the IPA annotation being included as a feature when it is not being used as labelling criteria to identify possible training BP genes.

2.2.3.2 Machine Learning to Test Training Gene Curation

As mentioned in the *least likely* BP gene curation, ML tests were run to decide on the best gene size for this least likely gene group to be used in the training data. The ML application of these test runs involved testing training datasets with 51 *most likely*

genes, 149 *probable* genes, and a variable number of *least likely* genes (depending on the p-value threshold and the PPI confidence threshold tested in the iteration). 4-label testing of least likely gene size on multiple ML test iterations was not included due to poor ML performance, with the 4-label group using the best least likely gene curation identified by the 3-label ML test iterations. In total four p-value thresholds were tested on ML iterations to filter least likely genes (p-values between 0.1-0.25, chosen by the filtered gene sizes being comparable to the other groups in the training data for class balance). Also, only 0.15 or 0.4 STRINGdb confidence levels for only experimentally measured PPIs were chosen for the least likely gene group's PPI filtering – to capture as many PPIs as possible and to filter out potentially false negative least likely genes by using only experimental evidence in STRINGdb.

Each ML application for these test runs used features that passed data pre-processing (detailed in section 2.2.4). Due to the data size of iterations, only tree-based models were benchmarked for computational efficiency: random forest (RF), gradient boosting (GB), extreme gradient boosting (XGB), CatBoost (CB), light gradient boosting (LGBM), decision tree (DT), extratrees. All models were applied using scikit-learn (v0.23.2) except extreme gradient boosting (xgboost package v1.2.0), light gradient boosting (lightgbm package v3.3.2), and CatBoost (CatBoost package v1.0.6).

The ML test runs had multiclass iterations and regression analysis iterations to assess least likely gene curation that applied to the differing ML approaches detailed in chapters 3 and 4 respectively. The results of the top-performing model for each

iteration of these test runs were recorded for comparison, to select the optimal least likely gene group that entered the final training data curation (n=93).

2.2.3.3 Assessing Genomic Bias

To assess biological bias, all four training groups had their genomic characteristics visualised. For feature cleaning and feature selection, only the 3-label training data (n=293 genes) was visualised due to higher ML performance (further explored in chapter 3), excluding a comparative visualisation of selected feature pairwise distributions for the 3-label versus 4-label training data (Figure 2.7). All further exploratory data analysis of features for the 4-label training data can be found in: https://github.com/hlnicholls/PhD-Thesis/tree/main/Chapter2/BP%20GWAS%20EDA/4%20label

After the training genes were grouped, the distance between the genes along the genome was recorded to avoid genomic distance as a confounder. This test showed whether the labelled gene groupings were likely to have similar annotations due to close genomic distance - identifying genes dispersed across chromosomes in the genome for each group.

2.2.4 Data Pre-processing

Machine learning, from exploratory data analysis to output model scoring of the trained top-performing model, was conducted using Python (v3.8.5). All 114 features were assessed for missingness and correlation (Appendix A Table 2).

All variant-level features that were collapsed to the gene-level had their missingness and correlation with gene length assessed (Appendix A Table 3) to avoid gene length bias and clean these features before entering ML. Sixteen variant-level features had a high proportion of data missing (> 25%) and so were removed. Variant-level features correlating with gene length measured by an absolute value >0.3 correlation were then explored in sensitivity analysis and removed in data pre-processing (9 of the variant-level features were removed, which were all counts of epigenetic sites per each gene, Appendix A Table 2). The probability-loss-of-function (pLI ExAC) whilst provided by ExAC as a gene-level score was also assessed for gene length correlation, finding that it had correlations of 0.31 (Pearson correlation) and 0.29 (spearman correlation) for the training genes (Appendix A Table 3). As the study curating pLI scores notes that it has minimal gene length correlation (0.17) in coding sequences[58], this led to the feature still being included to pass on to ML. For all other variant-level features, only beta value, GWAS catalog p-value, and chromatin state segmentation counts were complete enough variant-level features - with < 0.3 absolute correlation with gene length - to enter further feature pre-processing to be used in ML (Appendix A Table 2).

All gene-level features were removed if found to be missing for all genes by >25%. For correlation between all feature-feature relationships, a sensitivity analysis was performed using correlation coefficient and testing thresholds >0.9, >0.99 and > 0.85. Features were removed for each threshold and tested on ML benchmarking identifying the best correlation threshold at >0.9 for model performance (Appendix A Table 7). After removing the variant-level features highly correlated with gene length, any

features >25% missing, and any features with >0.9 correlation between all feature-feature relationships were removed. This led to 21/114 collected features that were imputed using random forest imputation (using the missingpy package, v0.2.0).

The feature distributions were compared between the training and test folds used in ML k-fold cross-validation (defined further in Chapter 3 Methods section 3.2.2), identifying features with any significant differences (<0.01 p-value) using Kolmogorov-Smirnov testing. Features with significant differences in training versus testing data were also removed to maintain i.i.d assumptions (i.i.d testing found in https://github.com/hlnicholls/PhD-Thesis/blob/main/Chapter3/3%20label/correlation09/Kfolds/iid%20assumption%20testing.ipynb). For the best performing 0.9 correlation threshold, 15/114 features remained to enter feature selection.

The 15 cleaned features underwent feature selection using the BorutaShap package (v1.0.13). BorutaShap selects features by comparing their importance to that of the randomised copies of each feature; known as shadow features. Feature importance from BortuaShap is derived from an input tree-based model. To select an optimised model to run the feature selection, a preliminary ML model benchmarking was performed, giving seven tree-based models (extreme gradient boosting, light gradient boosting, CatBoost, gradient boosting, random forest, decision tree and extra trees) all 15 cleaned features to undergo nested cross-validation with parameter-tuning. To run ML, scikit-learn (v0.23.2), xgboost (v1.2.0), CatBoost (v1.0.6) and lightGBM (v3.3.2) packages were used (further details can be found in Chapter 3 Methods). The top-

performing tuned model (extreme gradient boosting) then served as the input model for BortuaShap, which selected 6/15 features. All feature cleaning and feature selection steps were repeated for training data curated at 0.99 and 0.85 correlation thresholds for comparison (correlation comparison for ML performances found in: https://github.com/hlnicholls/PhD-Thesis/tree/main/Chapter3/3%20label).

The selected features then underwent further visualisation of their relationships, plotting their correlation and mutual information using the seaborn package (v0.11.2). Pairwise distributions of the selected features for the 3-label dataset and the 4-label dataset were visualised. For the 3-label data, the training dataset versus the dataset of all other genes to be predicted by the trained model were visualised in R using the ggplot2 package (v3.3.5). The distributional differences for all features between the training genes and predicted genes also underwent Kolmogorov-Smirnov significance testing with adjusted p-values to assess any distributional differences that may affect the i.i.d assumptions and ML generalisation to new data.

2.3 Results

2.3.1 Least Likely Blood Pressure Gene Analysis

To curate robust training genes, the *least likely* BP gene group underwent multi-layered filtering to ensure that the group contained genes that are as unlikely as possible to have an influence on BP and reduce the chance of false negative training examples. This led to the testing of multiple p-value thresholds (selecting genes containing only variants with high p-values on BP GWAS) and PPI thresholds

(filtering out genes with direct and secondary interactions with *BP-genes*) (Table 2.1). Four ML benchmarking iterations were tested against these thresholds finding 93 *least likely* genes (filtered with a p-value > 0.15 and PPIs > 0.15 experimental confidence) gave the best ML performance across metrics for both multiclass and regression analysis (Table 2.1).

a

p-value threshold	STRINGdb PPI confidence filter	Number of least likely genes passing thresholds	Balanced Accuracy	F1
0.1	0.15	159	0.66	0.77
0.15	0.15	93	0.65	0.73
0.2	0.4	98	0.57	0.64
0.25	0.4	68	0.6	0.67

b

p-value threshold	STRINGdb PPI confidence filter	Number of least likely genes passing thresholds	r^2	predicted r^2	MAE
0.1	0.15	159	0.78	-89834.63	0.13
0.15	0.15	93	0.74	0.897	0.12
0.2	0.4	98	0.81	-0.63	0.05
0.25	0.4	68	0.54	-0.7	0.16

Table 2.1. Least likely blood pressure gene group testing on machine learning. a shows the machine learning performance on multiclass classification (with the results recorded for the top-performing model per each iteration, measured by balanced accuracy and F1 score. **b** shows the machine learning performance on regression

analysis – also with the results recorded for the top-performing model per each iteration, measured by r^2, predicted r2 and mean absolute error (MAE). Each table shows least likely gene groups filtered by p-value thresholds and protein-protein interactions (PPIs) with blood pressure genes measured by STRINGdb at different confidence thresholds.

2.3.2 Genomic Characteristics

I analysed the genome coverage of all sentinel BP SNPs and their LD SNPs (>0.8 r^2), collated across all BP GWAS as of 2020[2, 4, 87-89], finding they cover 8.3% of the genome, representing almost 10% of the known gene complement. These results represent an important insight into the complex systems regulating BP and offer a basis for a better understanding of BP biology and the personalisation of hypertension treatment.

The training data curated from genes within the Evangelou et al. (2018) GWAS comprised of 293 rows of genes by 114 total annotations. The genomic characteristics of these genes were investigated, focusing on their genomic distance from one another, their distributions in gene lengths and the gene types per each group in the training data. Comparing the genes distributions across chromosomes showed each labelled group had genes varying in their positions across the genome (Figure 2.1). Furthermore, 17/181 loci in the training data contained multiple genes, leaving minor positional relatedness that could affect correlated genetic annotations.

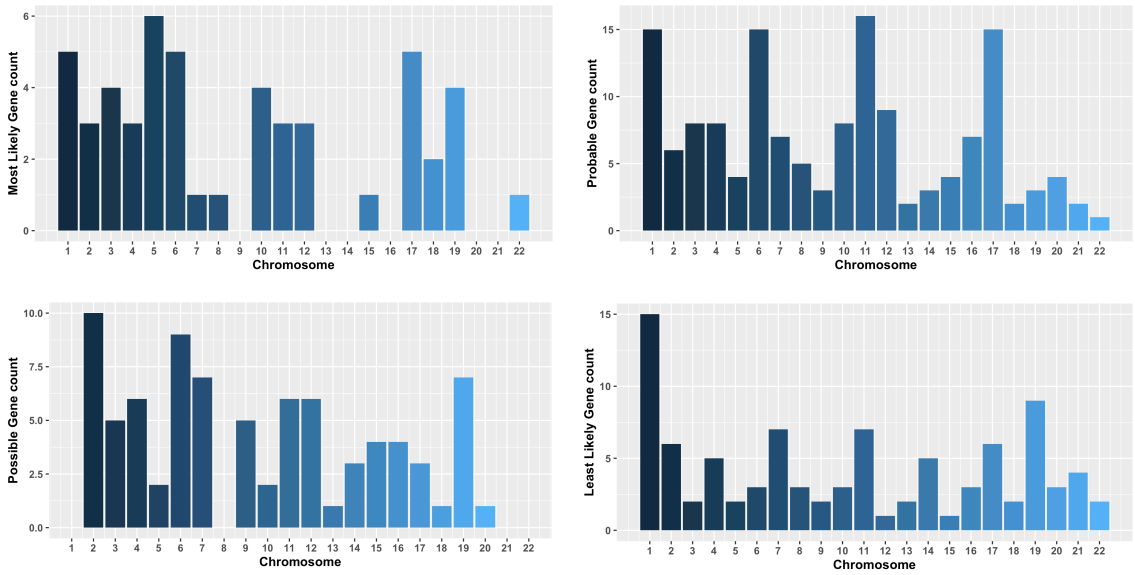


Figure 2.1. Training genes distributions across chromosomes. Counts of training gene groups across the four developed labels (*most likely, probable, possible* and *least likely*) at each chromosome across the genome.

On investigating gene lengths in the training data, all gene groups have genes predominantly shorter than 500,000 bp (Table 2.2). The *least likely* gene group was found to have the shortest genes of any group, with the shortest gene being 206 bp (Table 2.2). Furthermore, the correlation gene length had with the variant-level features found all features counting epigenetic sites per gene to have high positive correlations with gene length in the training data (Appendix A Table 3), showing these annotations should not be used as features. However, variant-level features for beta values, log p-values from the GWAS catalog, methylation site signal values, and counts of chromatin segregation states from ENCODE were found to have a minimal absolute positive correlation (<0.3), allowing for these features to pass further into the ML pipeline.

Gene label	Minimum gene length (bp)	Maximum gene length (bp)	Median gene length (bp)
Most likely	1467	722156	45142
Probable	2410	1125646	36394
Possible	2052	2298477	91717
Least likely	206	95203	1415

Table 2.2. Gene length per labelled group in the training data. Gene length summary statistics measured by base pair (bp) for each training gene group.

The final biological characteristic visualised was gene types, showing all the *most likely*, *probable* and *possible* gene groups are protein-coding or processed transcripts (Figure 2.2), reflecting the higher likelihood of these genes having more study and therefore more complete data to be useable in ML. The *least likely* gene group was also curated by analysing only protein-coding genes to avoid developing a gene group with less annotation, affecting which features would enter ML and how much the *least likely* gene group consisted of heavily imputed data. In comparison, the genes to be predicted (n=1,804) were also predominantly protein-coding, however other gene types (pseudogene and antisense) are also present, highlighting that these genes will need further analysis on output prioritisation by ML.

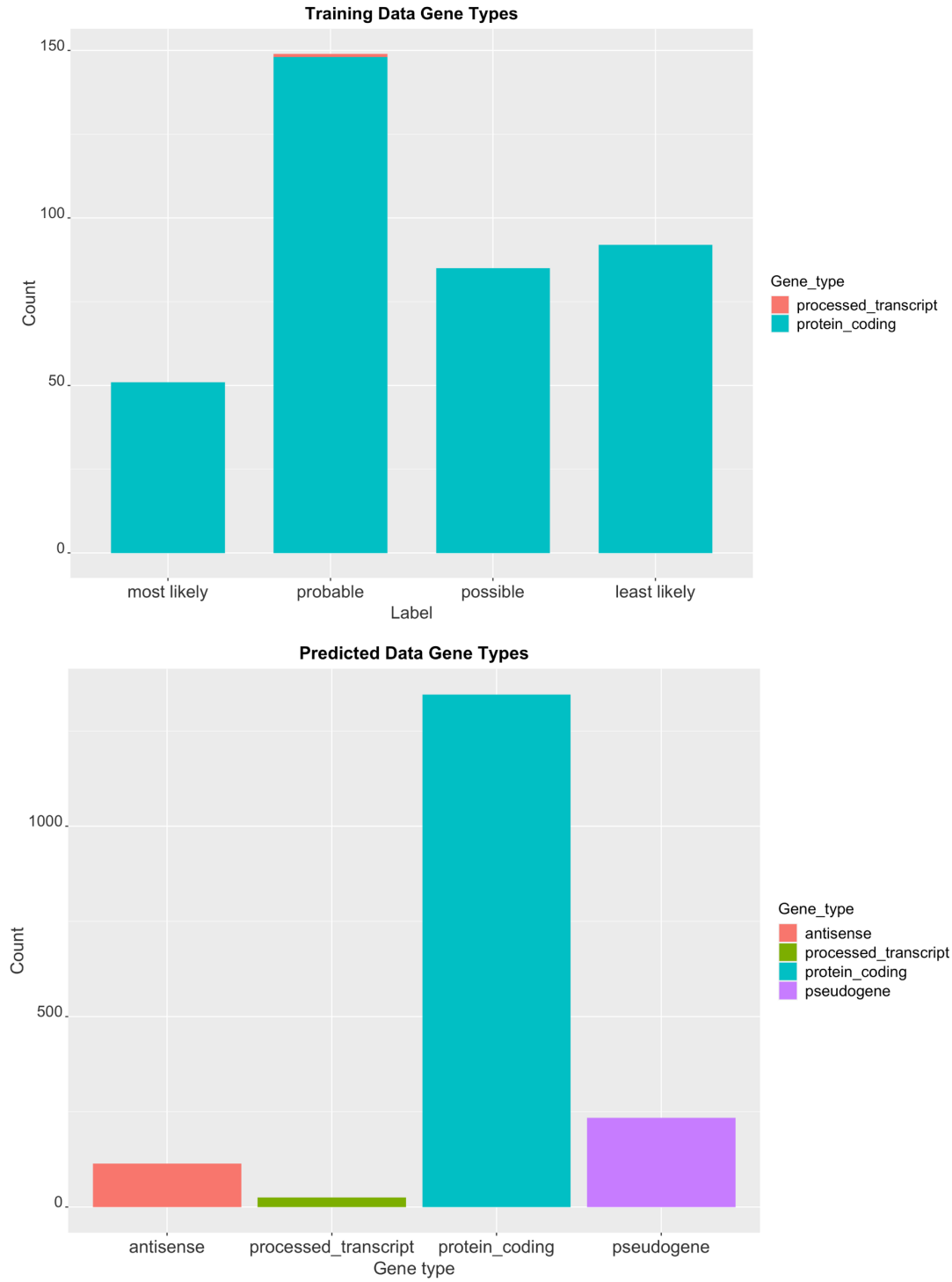


Figure 2.2. The proportion of gene types in the training data and predicted data.

The training data (n=377) contains only protein-coding gene groups, whilst the genes to be predicted (n=1,804) contain predominantly protein-coding genes with also processed transcripts, antisense and pseudogenes.

2.3.3 Feature Cleaning and Feature Selection

Following gene curation, collected annotations were then assessed for their potential use as ML features. From the 114 total annotations collected, 66 had less than 25% missingness in the training data (Figure 2.3, Appendix A Table 2). This was followed by 45 features having a greater than 0.9 r^2 correlation, and therefore removing these 45 correlated features (correlation threshold filtering detailed earlier in section 2.2.4, Appendix A Table 8). On testing the remaining 21 features and their i.i.d assumptions on cross-validation folds, a further 6 features were removed, leaving 15 features to undergo BorutaShap feature selection for multiclass classification.

BorutaShap feature selection identified six accepted features (HIPred, Heart - Atrial Appendage TPM, Pituitary TPM, Exomiser mouse score, SDI, pLI ExAC) and two tentative features (Fallopian Tube TPM and EBV-transformed lymphocytes TPM) (Figure 2.4). Tentative features are due to the Boruta algorithm reaching its designated number of iterations (default=100) without assigning importance to that feature in comparison to the most important shadow feature. This result requires a fix to be applied by comparing the median feature importance of the feature and the maximum shadow feature. For this analysis, running the additional step for tentative features then rejected both the Fallopian Tube TPM and EBV-transformed lymphocytes TPM from feature selection.

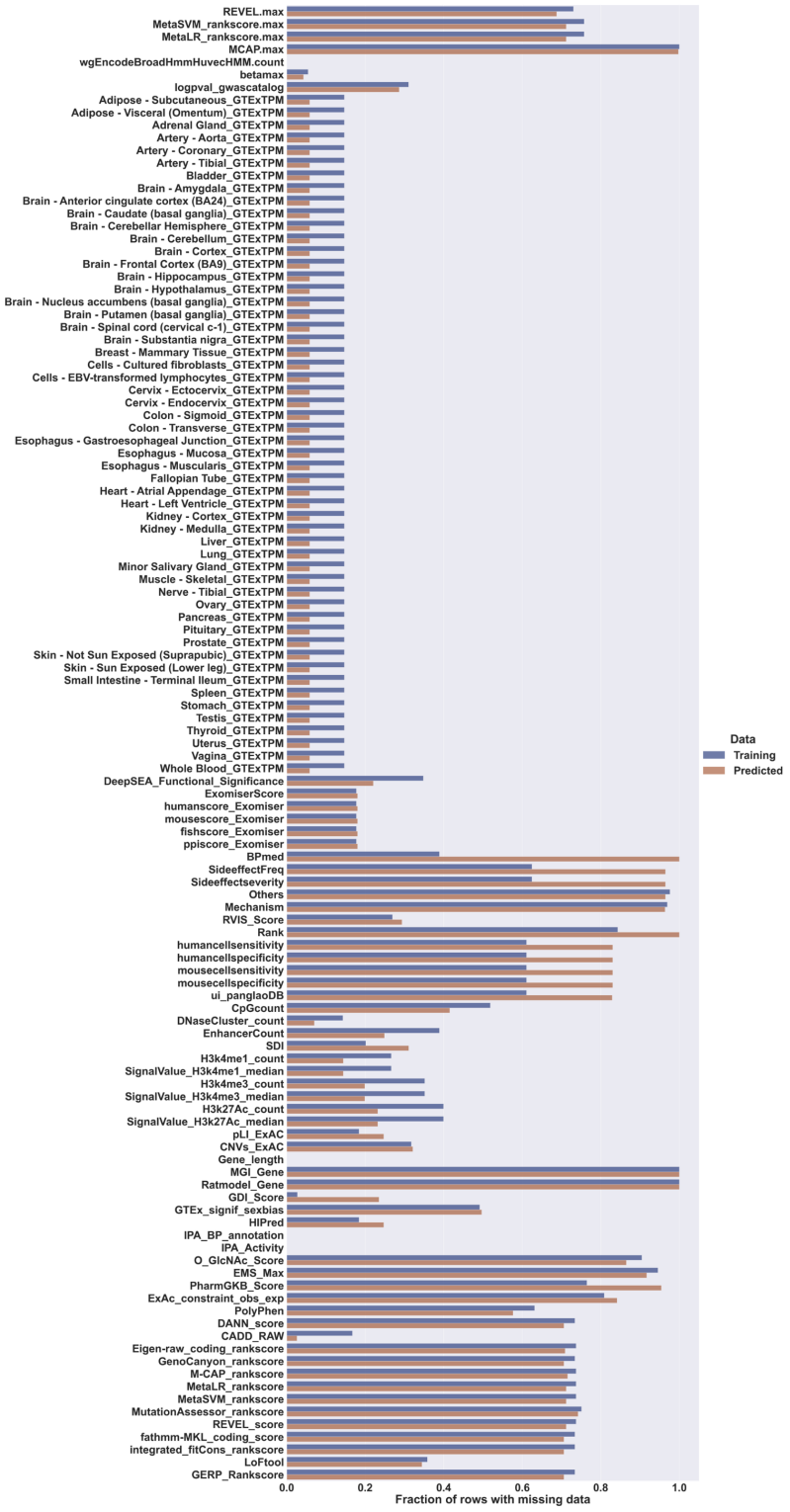


Figure 2.3. Feature missingness for the training and predicted data. Bar plots of the training data missingness (n=293, blue) and the predicted data missingness (n=1.804, orange).

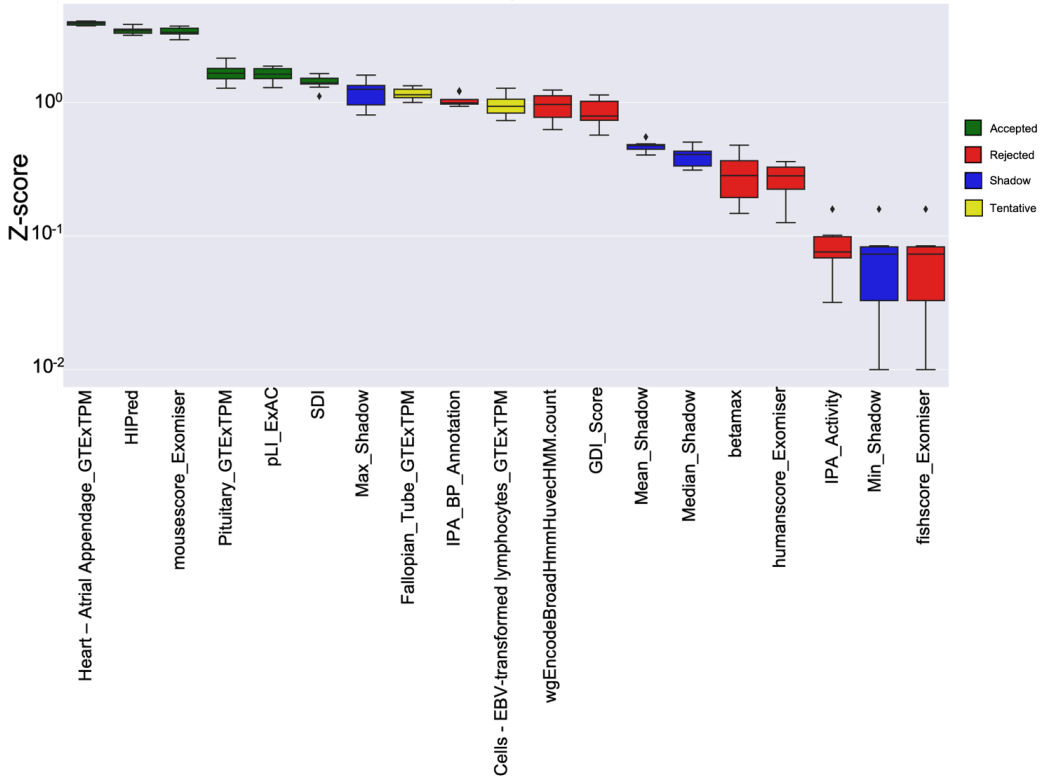


Figure 2.4. BorutaShap Feature Importance and Selection. BorutaShap identifies feature importance by comparing features against their randomised shadow features – with features calculated to be more important than the most important shadow feature ('max shadow') being selected. The selected features are coloured in green, summary shadow feature importances in blue, and tentative features (requiring comparison against the tentative feature's median importance against the maximum shadow feature) are in yellow. Feature importance is calculated via Shapley values and normalised using the z-score.

On examining the relationships between selected features, their correlation, mutual information gain and distributions were visualised (Figures 2.5-2.7). The correlation matrix identified that all features had minimal to negative correlation excluding

HIPred and pLI EXaC – with the two features having a positive correlation (0.63 r^2). To test the influence of the correlation between HIPred and pLI EXaC, I investigated partial dependence plotting between the two features (Figure 2.6). Partial dependence plots showed both features interacting affect model prediction with different directionality per each class (Figure 2.6), with linearity for the *probable* and *least likely* classes (the higher both features are in value, the more likely a probable prediction, and vice versa with lower feature values for a least likely prediction). Their interaction's influence on most likely BP gene prediction is less linear, with a 0.3 probability of predicting *most likely* genes when pLI has a value (approximately between 0.1-0.2) and HIPred has a value between 0.55-0.7 (Figure 2.6). The probability then drops to 0.24 when pLI remains between the same but HIPred either increases or decreases in value. The mutual information gain identified all features as offering information about the target variable (the gene labels scored between 0-1), with pLI ExAC providing the least information and Exomiser mouse scores providing the most information (Figure 2.5b).

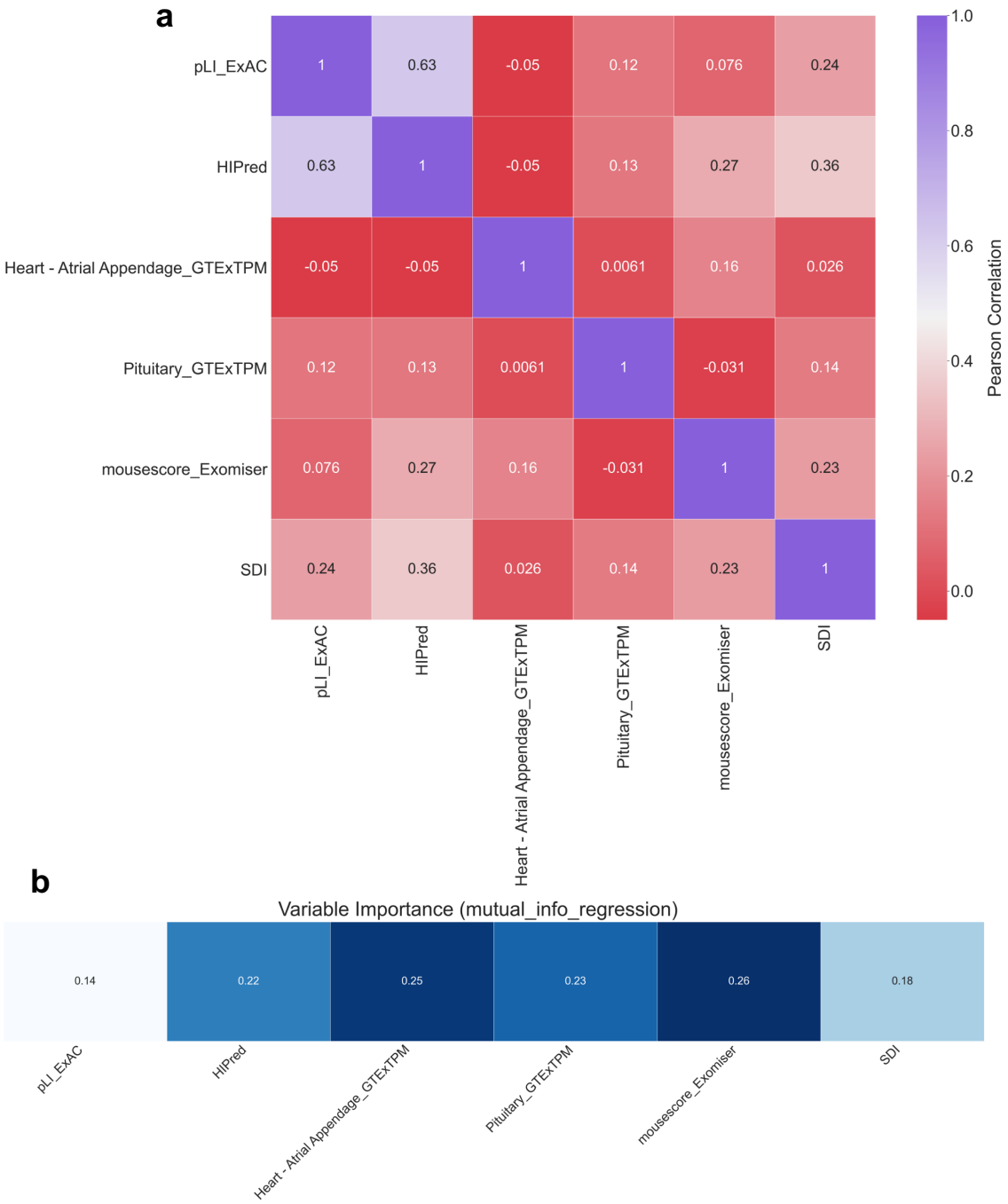


Figure 2.5. Relationships between selected features. The correlation matrix (a) shows all selected features and their correlation coefficients with a high correlation denoted in purple and a low correlation denoted in red. The mutual information shows the average information each feature conveys about the target variable.

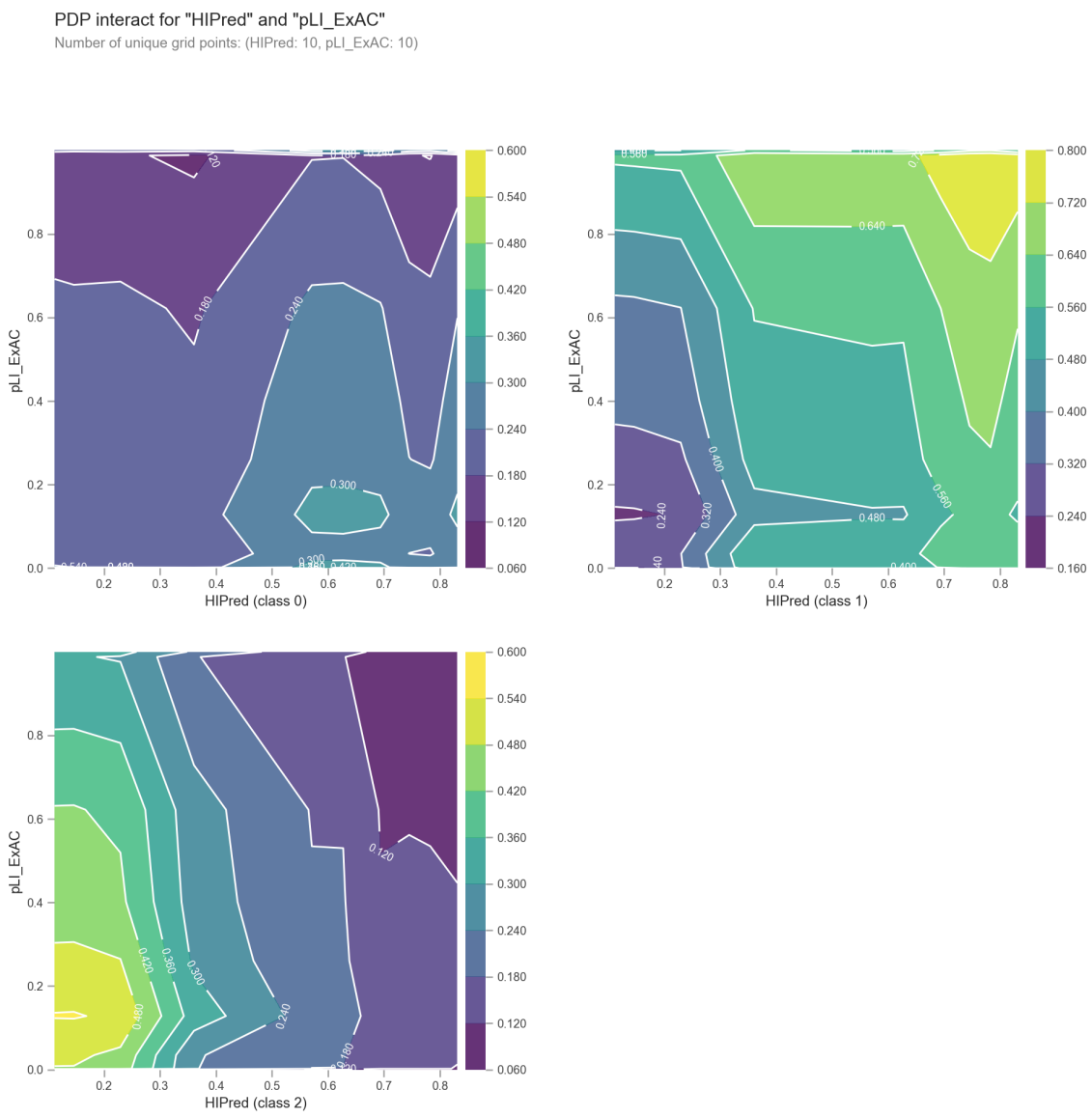
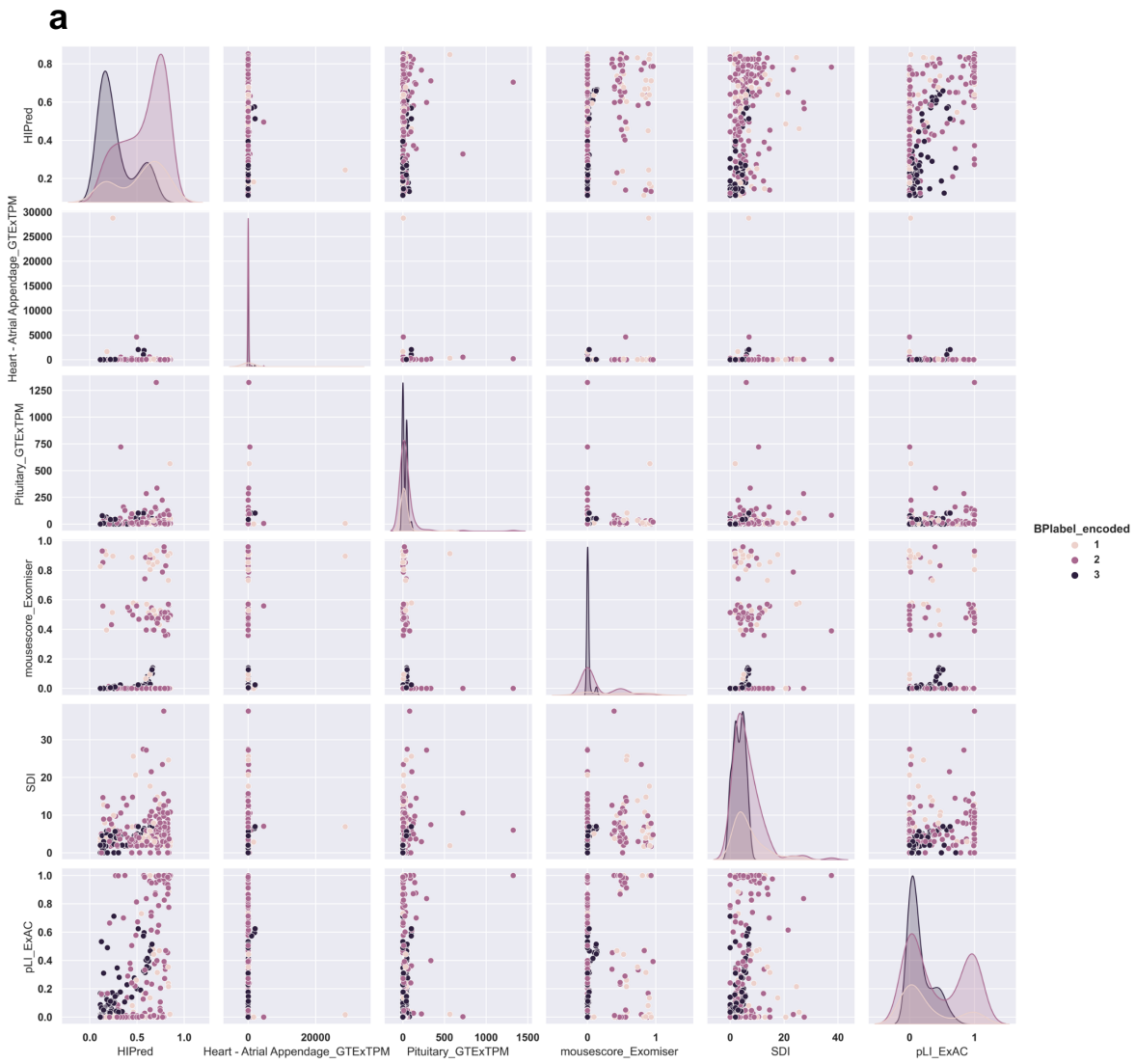


Figure 2.6. Partial dependence plotting between HIPred and pLI ExAC features. Class 0 denotes *most likely* blood pressure genes; class 1 denotes *probable* genes and class 2 denotes *least likely* genes. Each plot presents a colour scale from blue to yellow indicating the expected probability a gene is classified in that group depending on the values of the two features.

Pairwise plotting of selected features also showed that within each feature there are varying distributions between each of the 3 labelled gene groups (Figure 2.7a). This was plotted in comparison with the 4-labelled data, identifying that the fourth additional training gene group and their selected features offered less distinctive distributions in comparison to the 3-label groupings (Figure 2.7b).



b

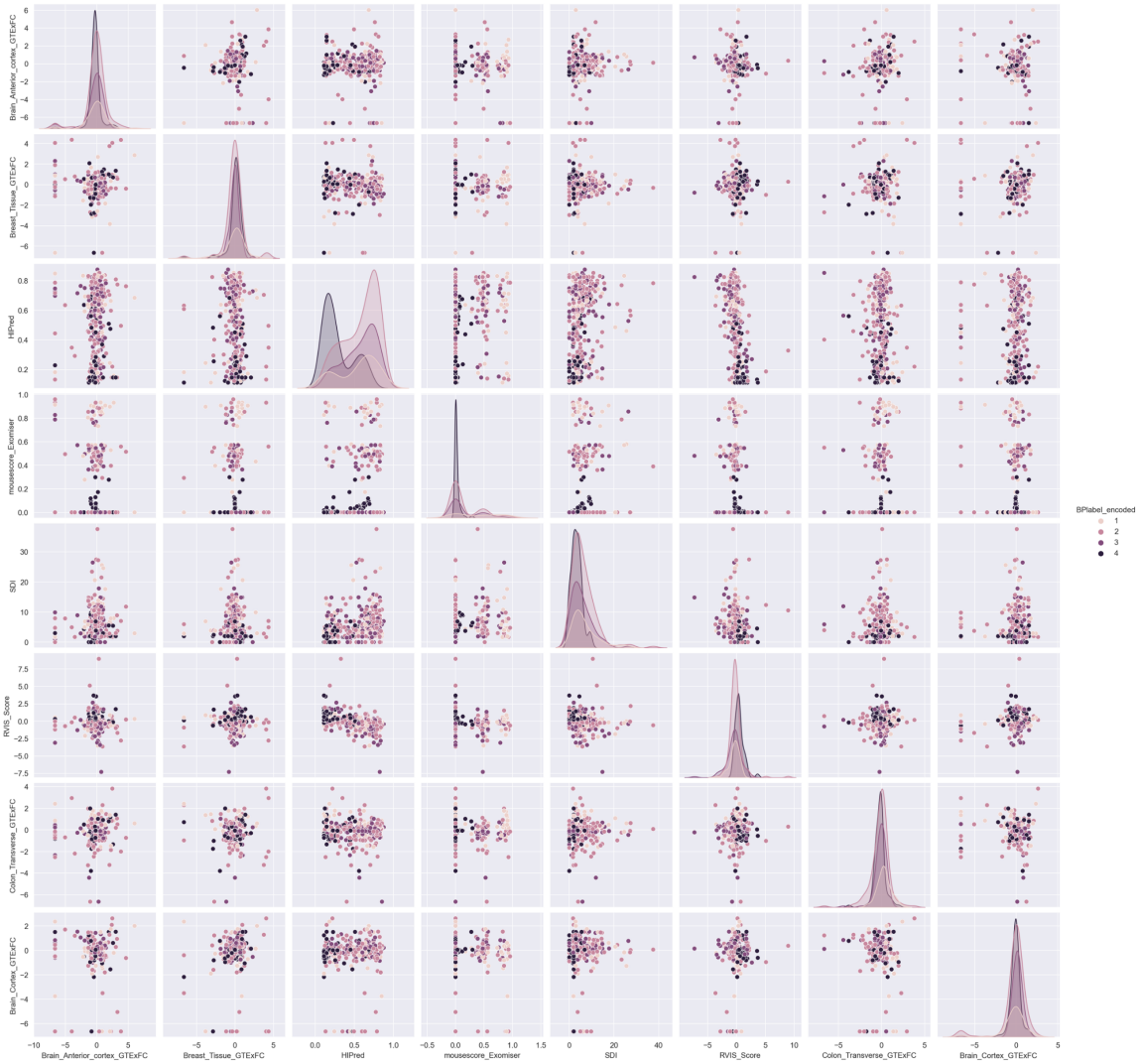


Figure 2.7. Pairwise distribution plots of the selected features. a) Distributions for the 3-labelled training data's selected features. Each selected feature is plotted against one another, with each blood pressure (BP) gene group identified (using the 3-label dataset). Most likely BP genes are scored at 1, probable genes scored at 2, and least likely genes scored at 3. **b)** Distributions for the 4-labelled training data's selected features. Most likely BP genes are scored at 1, probable genes scored at 2, possible genes scored at 3 and least likely genes scored at 4.

The distributions of the selected features showed both GTEx features (pituitary and atrial appendage TPMs) have distributions skewed to the left for both the training and predicted data (Figure 2.8). When comparing the training data distribution versus the predicted data, all differences have significant adjusted p-values excluding pLI EXaC.

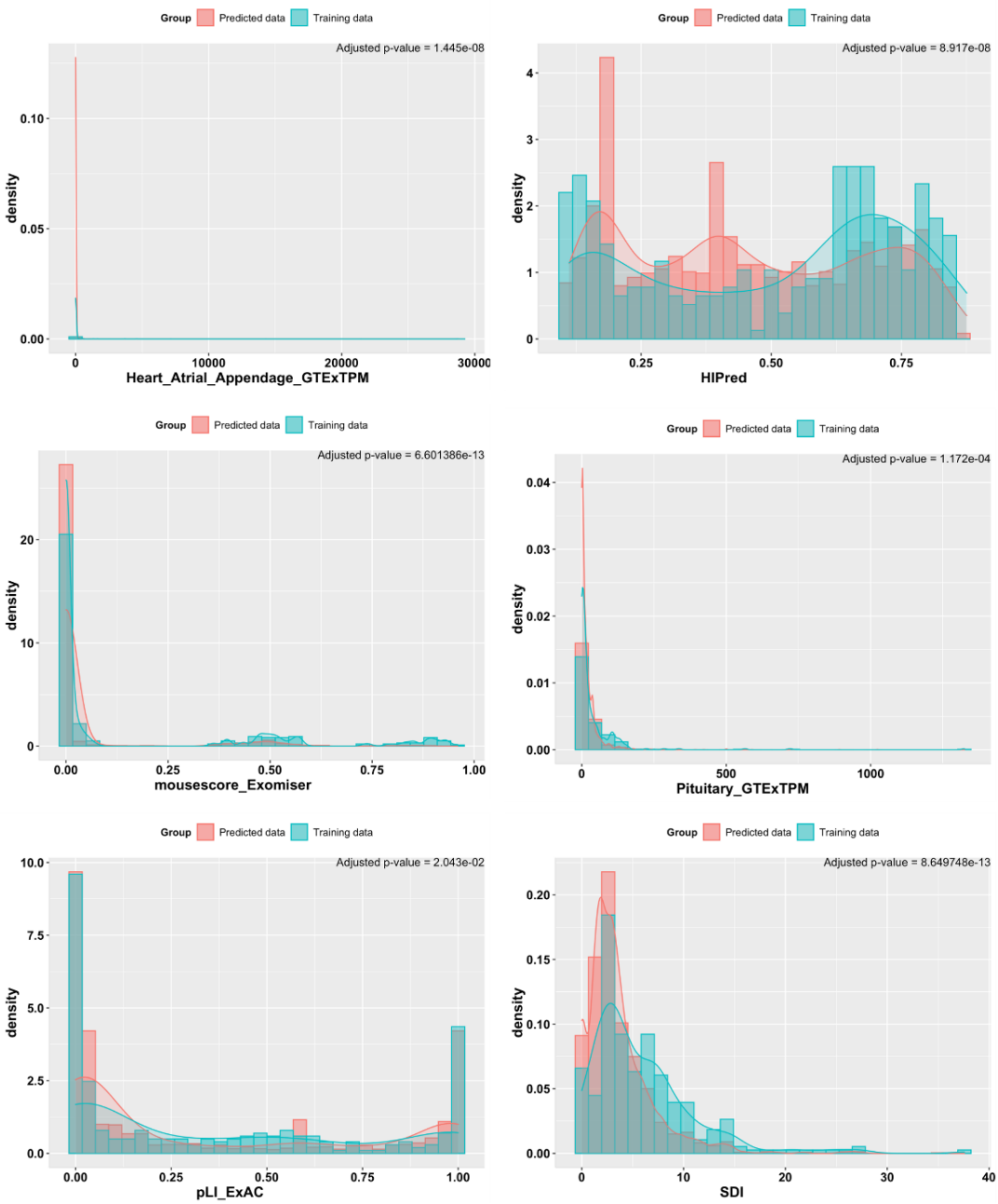


Figure 2.8. Distributions of selected features in training and predicted data. The annotations from the selected features for all training genes (n=293) were plotted in

blue against the data to be predicted (n=1,804) in red, with any significant differences calculated by Kolmogorov-Smirnov tests with false discovery rate adjusted p-values.

2.4 Discussion

A key focus of this chapter was the curation of exemplar genes that were of high enough quality to act as ML training data. To ensure this aim was met, expert clinical curation was used to define the *most likely* BP genes, providing 51 genes with established BP-related functions, alongside *probable* and *possible* gene group curation – which are supported by relevant literature and experimental evidence that justifies their labelling. However, the *least likely* BP gene curation posed the greatest challenge. Defining genes as non-causal is not a common research focus and there is great potential for false negative genes to be included that have unexplored impacts on BP. Furthermore, as BP is a complex trait that involves multiple organ systems this also increases the likelihood that a given gene may have even a small additive impact on BP, making *least likely* gene curation especially challenging. To address this obstacle as thoroughly as possible, I expanded on negative gene filtering seen in previous work (namely OpenTargets which identifies negative genes by only filtering out genes that had direct PPIs with disease-causing genes[84]), by using additional p-value filtering and also PPI filtering that considering not only direct but any secondary interactions with BP-genes associated across several GWAS[2, 4, 87-89].

However, the threshold filters themselves also needed evidence supporting their selection – otherwise cherry-picked settings of p-value and PPI thresholds may bias

ML performance. I tested four threshold settings, based on their output least likely gene size having as much class balance as possible with 51 *most likel*y and 149 *probable* BP training genes (Table 2.1). These four ML performances on classification and regression had *least likely* gene sizes ranging from 159-68. The highest ML balanced accuracy and r^2 performances were using the 0.1 p-value and 0.15 PPI confidence (giving 159 least likely genes). However, notably for the regression analysis, the predicted r^2 (which performs a leave-one-out cross-validation and gives insight into model overfitting) was heavily negative (-89834.63) indicating that model is fitting to noise added into the data by the larger least likely gene group and causing overfitting. On the other hand, the classification showed a positive balanced accuracy and F1 scoring (0.66 and 0.77) respectively, suggesting the overfitting is not reflected in classification. However, the 0.15 p-value and 0.15 PPI threshold (n=93 *least likely* genes) gave good ML performances on both classification and regression iterations, with a slightly lower classification performance (0.65 balanced accuracy and 0.73 F1 score) alongside an encouraging predicted r^2 of 0.897. This performance, alongside the gene size being 93 – which makes the group become a more balanced minority class, unlike the 159 gene group that would make the *least likely* BP genes the majority gene group in the training data – led to the 0.15 p-value and PPI thresholds and this 93 gene size being selected for further use in the BP training datasets. Although it should be noted that future work will need to track these curated *least likely* genes, ensuring that as BP research advances these genes are still reliable training examples, and not false negatives, which in turn will also validate the use of the p-value and PPI thresholds set here.

Analysis of genetic characteristics in the curated training data shows that bias risks due to genomic distance and gene length are minimised (with genes spread across the genome and gene length assessed to remove any biasing variant-level features including gene length itself). However, most genes being protein-coding highlights an unavoidable issue, as genes with greater completeness in their features are more likely to be protein-coding and will be more beneficial to ensure a trustworthy ML model. In the future, as non-coding genes become more frequently annotated, these genes can be incorporated into the ML pipeline as databases improve in size and quality for non-coding regions. However, for the time being as ML needs robust input data for its outputs to be reliable, this limitation led to training data with only protein-coding genes, falling in line with other gene prioritisation methods that also predominantly focus on protein-coding genes[19, 33].

Annotations were collected across molecular scales that aimed to provide a range of data for ML. In this collection, it was key to ensure annotations did not come from overlapping sources. For example, Exomiser has PPI data that is taken from STRINGdb that would overlap with the PPI filtering of least likely genes that also used STRINGdb, and so this was not an included feature. Fortunately, many databases have varying methodologies and data sources which means overlapping information can be avoided. For example, the Exomiser mouse score uses the Mouse Genome Database (MGD) and International Mouse Phenotyping Consortium (IMPC) data[67], SDI uses semantic similarity of GO terms[55], HIPred uses Ensembl and ENCODE[62], ExAC uses exome sequence data[58], and GTEx uses genotyping of tissues[53]. However, whilst most databases use tailored formulas or experimental methods to collect their

data, HIPred is the only feature that comes from a ML calculation itself. HIPred's use of ML suggests that this feature is exposed to ML biases such as overfitting, which could compound when the measure is used within further ML. Although, HIPred's development focused on minimising the study bias risks that are present in other haploinsufficiency measures that use biological networks[62]. Furthermore, Steinberg et al. (2015) tested HIPred's application on multiple datasets and used cross-validation, indicating that it has undergone assessments to minimise overfitting and can provide reliable information for gene prioritisation.

Also, it should be noted that for GTEx several datatypes exist that could have been explored, such as gene expression fold change. Initially fold change, a measure of cis-eQTL effect size, was collected but found to be heavily missing (Appendix A Table 9). Furthermore, recent work by Mostafavi et al. (2022) focusing on systematic differences between GWAS and eQTL signals has shown that eQTLs are skewed to unimportant genes[92], suggesting fold change data may provide less reliable information than other gene expression measures. In contrast, TPM provides a direct interpretation of the transcripts per gene that is normalised for gene length in GTEx, giving less room for it to convey biased information. However, fold change and TPM are still different understandings of gene expression, which suggests both datatypes could be included in the same training data supporting each other to inform a model. In this way, they would not provide overlapping information but different interpretations of gene expression, and this should be explored in future work (with a more complete fold change in the training data enabling its use as a feature in other phenotypes).

Overall, the range of features collected here are from publicly available databases (excluding IPA) that are comprehensive and still expanding. As databases develop and increase the feature optionality available to ML researchers, it will be important for future work to employ standardised feature collections (e.g., making sure to consider how the database collects its data, which data type to collect as a feature, and how that data might overlap with other features) to have not only robust methodology but also transparency for comparisons between ML and other methods that use orthogonal data such as fine-mapping.

Several features that would have provided more detailed information were heavily missing from the training data, such as cell-type data. Cell-type expression provides granularity into a gene's functions in different cellular systems, which would likely better inform a ML model. However, whilst single-cell RNA sequencing is growing in popularity, the access to the deposits of such data is variable and challenging[93], with few databases focused on this datatype across a wide variety of cell-types. In this chapter I used PangloaDB[93], which pre-processes single-cell RNA sequencing data from hundreds of studies in humans and mice, providing a ubiquitous index – measuring how often a gene is expressed across cell-types. However, this feature was 61% missing in the training data, showing the need for further expansion of such single-cell databases before they can reliably inform ML applications. Databases such as the Expression Atlas[94] and the Human Cell Atlas[95] are devoted to developing cellular data, indicating that it is only a matter of time until this datatype becomes a common contender as a feature in ML.

Feature missingness also showed variant-level features particularly suffer from missing annotations. This is partly due to a lack of experimental variant data (especially for the majority that are in non-coding regions) and a lack of standardisation in methods that are annotating potentially pathogenic variants[96]. This lack of variant-level information also reduces the granularity of information available to the model and highlights a need for research to focus on understanding pathogenic variants, which can then better inform the understanding of gene functions in diseases.

However, overall features from several omics databases met the <25% complete criteria (gene expression, gene essentiality, haploinsufficiency, phenotypic scores, and methylation data). This was followed by 45 GTEx tissues being removed for being highly correlated (>0.9 r^2, chosen after testing multiple correlation thresholds on ML performance). Highly correlated GTEx tissues may be unavoidable due to there being correlations between related tissues (e.g., all arterial tissues had > 0.98 correlation with each other in the training data, Appendix A Table 2) which has already been established in previous work[97]. However, highly correlating features may still hold useful information, which is why (alongside best ML performance) the 0.9 correlation was chosen as the threshold to remove only extremely correlating features. Highly correlating information may still be of use due to various models being able to differentiate between a feature's correlative noise and its impactful values – such as ensemble models that test different hypotheses. However, depending on the model type correlation can still bias an algorithm – for example, random forest is known to decrease the importance of features that are highly correlated[98], losing out on valuing

features that individually may provide important information about the target variable - and so this was accounted for with the testing of lower correlation thresholds (0.85 and 0.99) on ML (with ML performance on different correlation tests further discussed in chapter 3).

BorutaShap feature selection then identified 6 features to enter multiclass classification. All selected features have similar importance calculated by BorutaShap (Figure 2.4), with heart (atrial appendage) gene expression, HIPred and Exomiser mouse scores being the most important features. The importance of mouse Exomiser scores is to be expected as the scores convey the phenotypic impact that a gene has on BP (calculated from mouse phenotype data for increased BP) and validates the use of BorutaShap for feature selection. However, the high importance put on mouse Exomiser scores also flags a risk as genes with higher scores will already have known relationships with BP, therefore reducing a ML model's ability to prioritise novel genes if it is heavily reliant on this score. Furthermore, whilst mouse Exomiser scores are less likely to overpower model decision-making, as it is the third most important feature, the heart gene expression being the most important feature also indicates that the ML may be susceptible to circular pattern recognition. As only GTEx tissues are selected that have established BP relationships (heart and pituitary tissues) and so a model cannot leverage information in novel tissues about BP.

Ultimately, the selection of features related to BP is a double-edged sword, as whilst they may encourage circular pattern recognition, they also validate the BorutaShap feature selection method and the ML decision-making, ensuring that the model will

use relevant information and is less likely to be valuing noise in the data. Furthermore, the selection of BP-related features may identify genes in established BP pathways that have not yet been investigated but may hold potential as novel therapeutic targets. Additionally, the BP-related features are also accompanied by more general genetic information that describes gene functionality (haploinsufficiency from HIPred, gene essentiality from SDI and probability of LoF from pLI). These functionality features offer the opportunity for a model to stratify genes with greater functional impacts, potentially then enabling the prioritisation of genes that have a greater functional effect on BP regulation. These features were also shown to interact with each other to inform ML, giving a model information that would lead ML away from circular BP pattern recognition. This was shown by the partial dependence plotting between HIPred and pLI, which suggests their relationship is informative for each labelled gene group in the training data (Figure 2.6).

Meanwhile, whilst the IPA BP annotation (the other phenotypic annotation collected alongside Exomiser measures) was ranked 7th important by BorutaShap it was not more important than its shadow feature and therefore not selected, despite this feature being a BP-specific phenotypic annotation. IPA annotations of genes were based on identifying which genes had experimental analysis relating the genes to BP in the IPA database. This IPA annotation was curated as a categorical feature (scoring genes at 1 for those that have a direct link to BP and 0 for those that do not) suggesting it may not be as informative as continuous variables that can offer the decision tree used in BorutaShap several values for splitting nodes within a tree. From a domain biology perspective, however, the strength of the IPA phenotypic feature to provide more BP

information from functional evidence suggests that it may still serve not as a feature but as an additional criterion for creating another gene group to expand the training data. This reasoning developed the *possible* gene group category (for which ML is tested further in chapter 3 in comparison to the 3-label training data focused on here). Creating a fourth gene group is also one of the few ways to test improving the training data size, as the *most likely* gene group of 51 BP genes could only be expanded on with further expertly validated BP-drug interacting genes being published.

The selected features passed feature cleaning partly due to them meeting i.i.d assumptions within the training data (investigating each feature's distributional differences within k-folds and whether they were statistically significant). However, on comparing these features in the training data (n=293) with the data to be predicted by the fitted model (n=1,804) the comparison showed distributional differences. Kolmogorov-Smirnov testing found significant differences between all features excluding pLI ExAC. For the phenotypic features, such as mouse Exomiser scores this is expected as genes with BP annotations are more likely to appear in the curated BP training data. However, the significance of all other features indicates that it will be important to explore the model interpretation of the predicted data in comparison to the training data (as doing so may ascertain whether a model has been impacted by these distributional differences).

In conclusion, the training data curated here was shown to have minimised genetic bias risks, with this bias being minimised for both training data approaches with both 3-labels and 4-labels. The features collected also contain a variety of information from

genomic, epigenetic, and phenotypic data, arming a ML model with a range of datatypes to better understand the BP training genes. However, the exploratory data analysis of the selected features highlights the need to follow how they are being interpreted by ML models, with thorough benchmarking of model performance and decision-making to ensure a robust ML pipeline from start to finish.

3 Multiclass Classification for Prioritising Blood Pressure Genes

3.1 Introduction

In this chapter, I apply ML benchmarking to the training data curated in chapter 2, posing gene prioritisation as a multiclass supervised learning problem. This approach allows for a thorough assessment of model performance across several metrics, with various methods (curation of training data, models, and class balancing approaches) tested to optimise the ML framework.

3.1.1 Machine Learning Models

The landscape of ML for prioritisation of genes post-GWAS has focused primarily on labelled supervised learning approaches. Involving both simplistic and complex models in applications for post-GWAS prioritisation (Appendix B Table 1), across studies ML performance varies depending on the problem requirements and data available. Most commonly, five types of models have been implemented: logistic regression, support vector machine, random forest, gradient boosting, and deep neural networks. Logistic regression is a frequently applied statistical method that can be contemplated as a generalised linear model. In logistic regression, a regularisation term is usually applied - e.g., L1 (the sum of the absolute value of feature weights) and L2 (the sum of squared feature weights) – that introduces some bias while reducing variance, thereby improving predictive ability[99]. Isakov et al. (2017) used

elastic net logistic regression[100] which combines L1 and L2 penalties to prioritise IBD genes. This method performs both feature selection (L1) and shrinks coefficient sizes to reduce variance (L2)[101]. Regularised logistic regression with elastic net aims to minimise the 'curse of dimensionality - where data has a larger number of features than samples – which is a particular blight on genomic datasets such as GWAS data. For example, Isakov et al. (2017) used data consisting of 314 positive genes and 1,736 negative genes each annotated with 1,027 features. By applying logistic regression with elastic net, they could then select the best data for their models (309 features selected that were predominantly from biological ontologies). However, due to the growing size of genetic data, and the broader range of features becoming available to describe genes and variants, the increased computational demand requires more advanced models.

Nine out of 23 ML models for post-GWAS prioritisation reviewed in this thesis (Appendix B Table 1) are ensemble models, namely random forests, and gradient boosting. Ensemble methods combine multiple models to improve performance and are ideal for heterogeneous GWAS data. Deo et al. (2014) developed a GBM (OPEN - Objective Prioritisation for Enhanced Novelty) for prioritising causal genes in multiple diseases. They used data comprising more than 40,000 genomic features from public databases - Gene ontology (GO), Mouse Phenotype database, Human Phenotype Ontology (HPO), and Online Mendelian Inheritance in Man (OMIM) - aiming to benefit from unbiased features. GBM is a tree-based model, with tree branches performing yes/no decisions based on feature value thresholds that lead to a sample's classification[102]. GBM operates one tree at a time, attempting to optimise

with each tree. Deo et al. (2014) made accurate predictions with GBM identifying genes affecting CVD-related traits. Performance was measured by the area under the receiver operating characteristic curve (AUROC), with values ranging between 0.75-0.9 across traits[18]. The model's consistently high scores are due in part to ensemble methods providing the opportunity for predictive mistakes to be removed in aggregate, due to multiple models testing different hypotheses and taking an average, expanding the representational space of a classification problem[103]. This is seen with gradient boosting across research, with the model known for reducing bias and variance and offering improved accuracy[102]. However, there is also a need to benchmark model performance, as whilst ensemble models are reliable, a singular approach to a novel classification problem provides a risk of unnoticed overfitting - when a model performs well on training data but poorly on new/unseen datasets that do not exactly match the patterns present in the training data. Some amount of overfitting is inevitable, but extreme cases can render a model useless. Overfitting is also a known issue for gradient boosting depending on the regularisation techniques used.

Vitsios et al. (2020) built a semi-supervised learning (models using both labelled and unlabelled data during training) framework in which they benchmarked seven models (random forest, extra trees, GBM, extreme gradient boosting, SVM, deep neural networks and a stacking classifier using all models) to prioritise genes for three diseases - amyotrophic lateral sclerosis, chronic kidney disease and epilepsy[33]. In total, they used data containing more than 1,200 features describing tens of thousands of genes for each disease. They found that random forest was the top-performing classifier, with this ensemble model consisting of multiple decision trees predicting in

parallel[104]. Gradient boosting was the second most accurate, showing the high performance of tree-based ensemble classification. However, the AUCs between all algorithms were deemed too similar to conclude one model outperformed all others across datasets. These results were also supported by comparison with a combined framework using all models in prioritisation, the stacking classifier, ensuring the highest reliability in the chosen classifier for each disease[33]. Meanwhile, Kafaie et al. (2019) aimed to prioritise genes associated with colorectal cancer by comparing various models (SVM, random forest, logistic regression with stochastic gradient descent and K-nearest neighbours). They found that logistic regression was the highest-performing ML model. The contrast in these results emphasizes that a classification problem may require simpler solutions and that GWAS prioritisation for all traits may not be encompassed by a one-size-fits-all model.

Besides ensemble learning and logistic regression, SVM is also consistently used within studies performing benchmark comparisons[50, 72, 106, 107]. SVM aims to plot a decision boundary between groups by measuring hyperplanes - based on the distances between the most extreme samples of each classification group[108] (Figure 1.3). SVM is regularly compared due to its effectiveness in high dimensional spaces and computational efficiency. However, within benchmarking studies, SVM has not shown itself to be the highest performing model. For example, Vitsios et al. (2020) found it had the lowest AUC (0.83, only slightly lower than the top-performing random forest at 0.85) of their seven models, while Kafaie et al. (2019) found SVM performed better than random forest yet worse than logistic regression. The varying performance of SVM also highlights the importance of input data, as Kafaie et al.

(2019) were one of the only studies to focus on comparing feature selection methods as well as models. Kafaie et al. (2019) found SVM performed well given certain features, whilst in comparison logistic regression had a more stable high performance regardless of the external feature selection, emphasizing the value of logistic regression's internal feature selection via regularisation.

Deep learning has also been explored for prioritisation. This method can increase sensitivity in larger datasets due to the method's ability to incrementally capture abstract representations of high-level information. In general, this is beneficial for GWAS prioritisation where the data is growing dramatically in size and heterogeneity with increasing annotations post-GWAS and currently few labelled samples (known disease-causing variants/genes) for supervised learning. Deep learning becomes advantageous in this scenario as it identifies complex patterns via supervised and unsupervised learning from large datasets[109] and can be applied for further insights into GWAS data. However, whilst deep learning enables the consideration of millions of parameters, its application to date has mostly flourished in image classification and natural language processing[110-112], requiring an investment in its development and benchmarking with traditional models for developing GWAS applications. A deep neural network (ExPecto) applied by Zhou et al. (2018) used natural language processing to prioritise causal variants for immune-related diseases using sequence-based features. This dataset contained more than 140 million promoter-proximal mutations and allowed for the unidirectional flow of information from base-sequence to functional predictions which enabled variant prioritisation. To approach this large dataset ExPecto applies a spatial transformation to the data, weighting transformations

based on transcription start site distances. This was performed on a tissue-specific basis of over 200 tissues[39], providing hundreds of features for the model to process. ExPecto is also able to perform pattern recognition and prioritisation of rare and unobserved variants. However, whilst models such as deep learning are selected based on their suitability to the data, performance can also be dependent on class balance and data quality available.

Another example using deep learning is Bao et al. (2020) developing a deep learning kernel method to infer gene causality within loci for gastric cancer, colorectal cancer, lung cancer, and psychiatric disorders[113]. In this method, the neural network layers encode raw SNPs as abstracted information, which is followed by a kernel regression layer that tests the SNP's significance in disease-associated pathways - with kernel methods being particularly useful in identifying non-linear relationships[114]. Deep learning is able to augment itself by incorporating other ML methods into its network and this shows how the method is uniquely advantageous with increased computational power. The method also highlights how ML has the potential to break free from circular pattern recognition, as the disease-pathway data used allowed for the model to identify SNPs as significant in disease pathways that they had not been associated with previously. For example, the model by Bao et al. (2020) found a link between SNPs acting on dilated cardiomyopathy and schizophrenia[113], suggesting they have shared biological pathway(s) that are yet to be explored in functional work. On further analysis, Bao et al. (2020) did find relevance in clinical studies with schizophrenia patients shown to develop dilated cardiomyopathy. These results differ from other studies where ML faces issues with prioritising genes in shared pathways

of known causal disease genes (due to input data describing the genes usually being circular in nature) providing less novel biological insight that can be translated to new drug targets.

3.1.2 The Ideal Machine Learning Method

Applying a reliably optimal model is difficult to ascertain for any ML problem. An ideal ML model for post-GWAS prioritisation would have thousands of positive and negative examples to learn from in training. However, in GWAS this is far from the case and there are varying definitions of positive and negative genes/variants for diseases, with most diseases having a great class imbalance on ML as there are minimal positive or negative disease-causing genes. One approach to address this class imbalance and quality of labelling has been developed by OpenTargets who have been curating gold standard positive and negative cases for their extreme gradient boosting model to prioritise genes post-GWAS with a 'locus 2 gene' (L2G) score[84]. This method focuses on prioritising variants and genes within an individual GWAS, providing an interface for researchers to view locus-level prioritisation for published GWAS research. In comparison, other studies address class imbalance by developing positive-unlabelled learning frameworks with bagging techniques[19, 33]. Positive-unlabelled learning allows for models to learn from equal sample sizes and for the training data to not need negative/non-disease-causing examples, however, overfitting is also a risk in this approach as the model will learn specific patterns from a bagged sample that it then generalises to the rest of the input data.

Another important aspect of an ideal model applied to biological problems is its ability to recognise novel patterns and not get trapped within circular predictions. For example, the OpenTargets researchers note their positively labelled loci are biased towards nonsynonymous variants[115], making the model less likely to prioritise novel variants with a smaller effect size. To improve their output, they present the ML score with several other metrics (fine-mapping, disease-disease colocalisation analysis and disease-molecular trait colocalisation analysis across 92 tissues and cell types, phenome-wide association study analysis, and enriched trait evidence), creating a stronger evidence-base to support their ML outputs. Whilst this approach overall is a slow-growing and labour-intensive way to define many gold standards and curate additional biological metrics, it is a clear-cut way to develop high-quality training data and reliable output prioritisations of new data. In contrast, many other studies do not define their criteria for positive/negative examples or provide further analysis of prioritised genes, with other studies following gene rankings with only discussion of complementary studies to their findings[113]. This lack of comparison between studies then creates difficulty in knowing the true performance of a model for a given disease, indicating the need for improved detail and clarity in training data collection.

Overall, there is a need for benchmarking to select the model best suited to the data and a particular prediction problem. This, in combination with a focus on the size and quality of the training data curated will enable robust optimisation of a ML framework.

3.2 Methods

3.2.1 Training Data

From exploratory data analysis to output model scoring of the trained top-performing model, all steps of the machine learning pipeline were conducted using Python (v3.8.5). The BP GWAS data that underwent pre-processing and feature selection was fully described in chapter 2. From the 7 million variants that were annotated to genes, these genes had 114 features collected that were assessed for missingness and correlation, undergoing feature cleaning and selection described in chapter 2. Features were removed if found to be missing for all genes by >25%. Features with a person's correlation coefficient >0.9 and not meeting in i.i.d assumptions between train and test data were also removed (i.i.d testing found in https://github.com/hlnicholls/PhD-Thesis/blob/main/Chapter3/3%20label/correlation09/Kfolds/iid%20assumption%20testing.ipynb). Features passing data cleaning were then imputed using random forest imputation (using the missingpy package, v0.2.0) and went into BorutaShap (v1.0.13) feature selection.

The feature cleaning and selection steps were completed for two training datasets – one dataset with three labels and another with four labels (defined in chapter 2 Methods 2.2.2 and 2.2.3). The 3-label dataset (n=293) consisted of three gene groups (51 *most likely* genes identified as BP-regulators, 149 *probable* genes identified from text-mining, and 93 *least likely* genes to affect BP identified through their lack of GWAS significant or PPIs with BP genes, Appendix A Table 6), and the 4-label (n=377) included an extra fourth category of 84 *possible* genes identified by their

annotation to BP in IPA (Appendix A Table 5). The four gene groups and how they were curated were defined in chapter 2 Methods section 2.2.3.

Any of the *BP-genes* group (group defined in chapter 2 Methods section 2.2.2) that did not meet any of the criteria to enter the training data were reserved as genes to be predicted by the training model (n=1,804 for genes that did not enter the 3-label training dataset, and n=1,720 for the genes that did not enter the 4-label dataset).

3.2.2 Machine Learning Model Benchmarking Methods

Fourteen multiclass models were benchmarked: random forest (RF), gradient boosting (GB), extreme gradient boosting (XGB), CatBoost (CB), light gradient boosting (LGBM), decision tree (DT), extratrees (ET), k-nearest neighbours (KNN), support vector machine (SVM), a sequential neural network (NN), logistic regression (LR), a voting model and a stacking model both consisting of all performing models (excluding k-nearest neighbours in the voting model due to incompatibility within scikit-learn) and finally a bagging model using the top-performing model (CatBoost). All models were applied using scikit-learn (v0.23.2) except extreme gradient boosting (xgboost package v1.2.0), light gradient boosting (lightgbm package v3.3.2), CatBoost (CatBoost package v1.0.6), and the neural network (TensorFlow v2.9.1 and Keras v2.9.0). Multiclass classification was chosen over binary classification due to the small sample size of most likely BP genes. These models were benchmarked on both the 3-label and 4-label datasets, providing a performing comparison between the two curations of training data.

Each model underwent hyper-parameter tuning using Bayesian optimisation (using scikit-optimize v0.8.1) and nested 10-fold stratified nested cross-validation. Nested cross-validation involves two cross-validation loops performed in parallel (outer and inner loops), minimising the risk of overfitting and enabling hyper-parameter tuning. For every iteration of an outer cross-validation fold, all inner cross-validation folds are also performed. Running on all inner folds finds the optimal model parameters that are further tested on that outer cross-validation fold. This was performed with ten k-folds of the training data for each outer and inner cross-validation. Parallel computer processing was not enabled for model assessment (as this invalidates the nested aspect of the cross-validation). Model performance was evaluated with accuracy, balanced accuracy, F1 score, precision and recall selecting the top-performing model for further analysis.

Hyper-parameter tuning and benchmarking performance for all models excluding the neural network took 6-8 hours per iteration (depending on training dataset size). However, due to the higher level of complexity of tuning neural network hyper-parameters, the tuning for the sequential neural network via Bayesian optimisation was significantly more time-consuming (>1 day run time). To address this time-inefficiency a hyperband tuner within the Keras package was first implemented – identifying tuned hyper-parameters within minutes. However, due to the hyperband's incompatibility with scikit-learn's nested cross-validation function, it was not possible to tune the neural network this way and have directly comparable results to the other models. These hyperband-tuned parameters were instead used to set smaller ranges to test within scikit-learn's Bayesian optimisation for the neural network's tuning.

Overall, allowing for hyper-parameters to be tested that decreased the runtime for the neural network to a few hours.

To combat class imbalance, an oversampling iteration and a balanced class weights iteration of model benchmarking were tested using the imblearn package (v0.9.0). Each of these tests were performed with the top-performing training dataset only (the 3-label dataset). Within imblearn, Synthetic Minority Oversampling Technique (SMOTE) was used to oversample all minority classes, giving matching numbers of genes in each group as the majority class. This created a training dataset with 149 genes in each group (oversampling to match the 'probable' majority gene group size, n=447 for the total 3-label training data). Meanwhile, the balanced class weights were applied using scikit-learn's class weight computation on model fitting, which penalises misclassifications of the minority class with greater penalty weights. Furthermore, after using class balancing, probability calibration was performed on the top-performing model's output classifications using scikit-learn's sigmoid calibration. This method fits a regressor to calibrate the probabilities predicted by a model fitted to the training data, supporting probability estimates so that they can be interpreted as confidence level for the classification.

Model performances were assessed using scikit-learn's metrics and confusion matrix functions, and these were further plotted in R (v4.1.2) using ggplot (v3.3.6). The top-performing class weight balanced CB model underwent interpretation using the python SHAP package (v0.36.0), providing feature importance values both globally for overall model performance and individually for each gene. Plots of the feature

importance (for both overall predicting and individual predictions) were created alongside feature-feature interactions. All ML code can be found in: https://github.com/hlnicholls/PhD-Thesis/tree/main/Chapter3

3.2.3 Gene Prioritisation Analysis

I used Enrichr[116] to compare Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways in three groups: 1) the most likely predicted genes, 2) the genes containing sentinel SNPs from the GWAS by Evangelou et al. (2018), and 3) the training genes related to BP (genes labelled most likely and probable respectively, curated as described in Chapter 2). KEGG gene enrichment analyses were visualised using ComplexHeatMap (v2.6.2) in R.

3.3 Results

3.3.1 Multiclass Machine Learning Framework

Two iterations of the training data were devised: a 3-label dataset consisting of the *most likely*, *probable* and *least likely* genes, and a 4-label dataset consisting of all three groups plus the fourth *possible* labelled group. This resulted in 293 training genes in the 3-label dataset and 377 genes in the 4-label dataset. From the 114 collected features for the 3-label data (focused on due having higher ML performances detailed in section 1.3.2), 48 were removed due to missingness, and a further 45 were removed due to being highly correlating. I then used BorutaShap to perform feature selection on the 20 features that remained after cleaning (with all feature cleaning and selection detailed in chapter 2). Six features (HIPred, Heart - Atrial Appendage TPM, Pituitary TPM, Exomiser mouse score, SDI, pLI ExAC) were selected and used as model input, with benchmarking fourteen models on repeated nested cross-validation. After data pre-processing, model benchmarking was performed on both datasets followed by class balancing approaches being tested (oversampling versus class-weighting). The top-performing model was fitted to the top-performing training data (3-label). This left 1,804 *BP-genes* that were taken from one GWAS[47] and were not included in the training data that were then prioritised by that top-performing model (Figure 3.1).

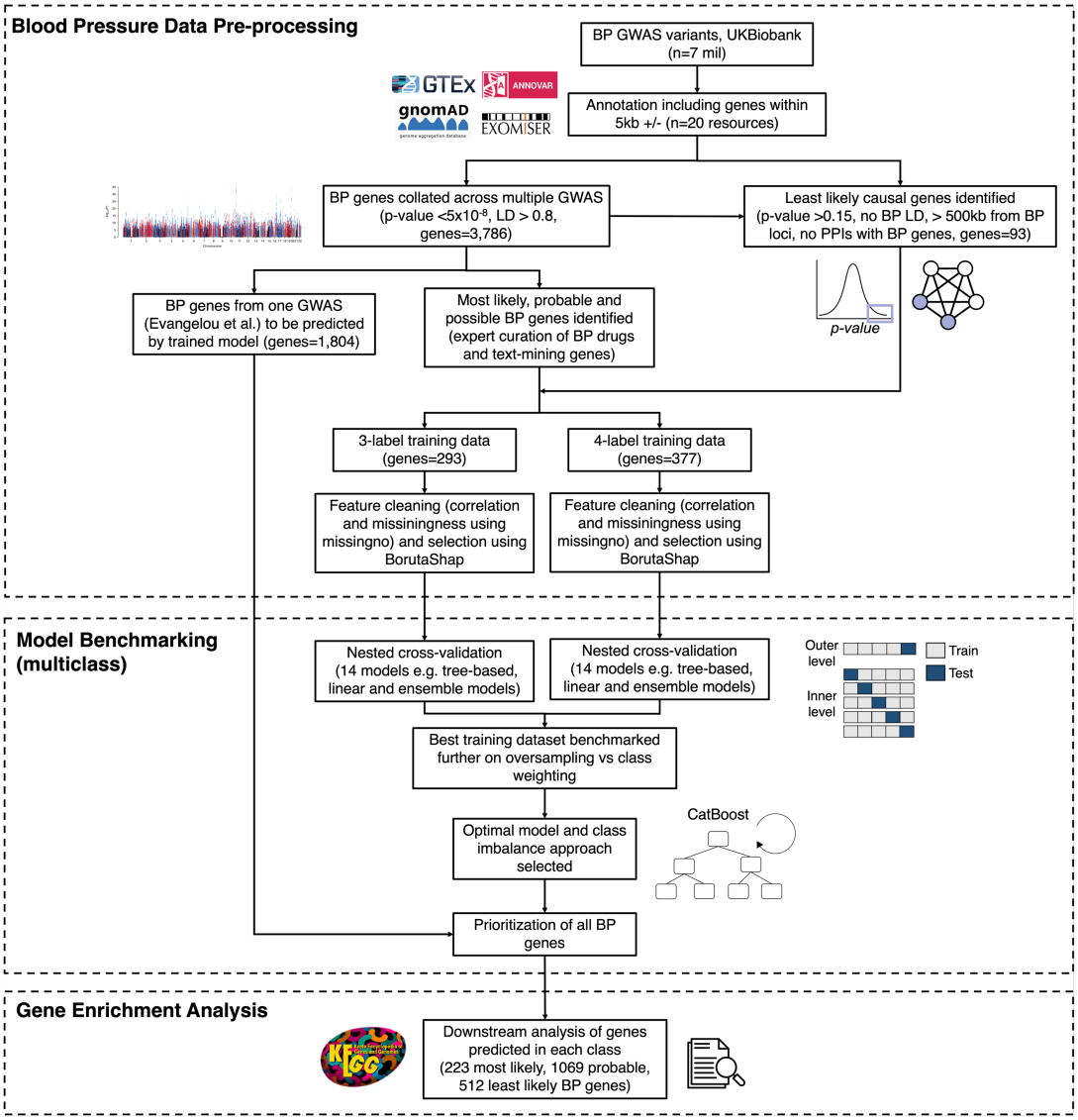


Figure 3.1. Multiclass classification framework overview. Blood pressure genome-wide association study (BP-GWAS) variants from Evangelou et al. (2018) were annotated to genes and evaluated by benchmarked machine learning. Data pre-processing involved annotating variants to genes from the whole GWAS and collecting gene-level annotations from several databases. The genes were then filtered to identify *BP-genes* (with linkage disequilibrium, LD, $r^2 > 0.8$ and a p-value $<5x10^{-8}$) and *non-BP genes* that are genes *least likely* to affect BP (selected by meeting criteria of: not in LD, p-value >0.15, not within 500kb +/- loci and no direct or

secondary protein-protein interactions with BP genes). The *BP-genes* were categorized into three groups: BP-regulator genes (clinically evidenced regulators of BP, labelled *most likely*), text-mining BP genes (genes with published evidence of BP interactions, labelled *probable*) and BP genes labelled as *possible* due to having experimental analysis relating to BP in IPA. The genes, alongside a *least likely* BP gene group, created two training datasets (3-label and 4-label depending on the presence of the *possible* gene group) and unlabelled genes are those to be predicted by the top-performing model with regression. Model benchmarking was then applied to compare 14 models using 6 selected features, testing the two training datasets and then oversampling versus class balancing approaches. The top-performing trained model (CatBoost) was then used for gene prioritisation, with the genes undergoing downstream analyses in which each predicted class grouping was compared.

3.3.2 Three Label versus Four Label Performance

On comparing model performances between the two datasets curated to identify most likely to least likely BP genes, the training data with three labels (n=293) had consistently better performance for all models across all metrics versus the training data with four labels (n=377) (Table 3.1). The performance per each labelled gene group showed the models all have low F1 scores for predicting the most likely gene group in both training datasets, and the four-labelled gene group also had low F1 scores for predicting the possible gene group (Figure 3.2).

	Three Label					Four Label				
Model	**Accuracy**	**Balanced Accuracy**	**F1**	**Precision**	**Recall**	**Accuracy**	**Balanced Accuracy**	**F1**	**Precision**	**Recall**
XGB	0.75	0.64	0.72	0.74	0.75	0.6	0.53	0.56	0.58	0.6
LGBM	0.71	0.63	0.69	0.71	0.71	0.59	0.51	0.53	0.52	0.59
CB	0.76	0.68	0.73	0.74	0.76	0.54	0.46	0.46	0.43	0.54
GBM	0.72	0.62	0.7	0.74	0.72	0.59	0.48	0.51	0.53	0.59
RF	0.7	0.6	0.68	0.65	0.7	0.56	0.48	0.49	0.5	0.56
DT	0.69	0.61	0.68	0.69	0.69	0.49	0.39	0.43	0.39	0.49
ET	0.72	0.6	0.68	0.64	0.72	0.53	0.43	0.43	0.4	0.53
KNN	0.72	0.66	0.71	0.72	0.72	0.5	0.44	0.47	0.5	0.5
SVM	0.73	0.67	0.73	0.7	0.73	0.54	0.49	0.48	0.45	0.54
LR	0.59	0.48	0.55	0.57	0.59	0.49	0.42	0.42	0.43	0.49
NN	0.66	0.58	0.62	0.62	0.67	0.5	0.39	0.39	0.36	0.5
Stacking	0.72	0.64	0.7	0.69	0.72	0.57	0.5	0.53	0.54	0.57
Voting	0.72	0.6	0.69	0.68	0.72	0.6	0.51	0.52	0.53	0.6
Bagging	0.76	0.67	0.72	0.72	0.76	0.59	0.5	0.55	0.52	0.59

Table 3.1. Model performance comparison between training data with three or four labels. Each model was benchmarked on two training datasets, one with three labels (n=293) and one with four labels (n=377), with each model assessed across accuracy, balanced accuracy, F1, precision and recall. The fourteen models were benchmarked: extreme gradient boosting (XGB), gradient boosting (GBM), CatBoost (CB), LightGBM (LGBM), random forest (RF), decision tree (DT), Extratrees (ET), K-nearest neighbours (KNN), support vector machine (SVM), logistic regression (LR),

neural network (NN), and three meta-ensemble methods – stacking, bagging, and voting models.

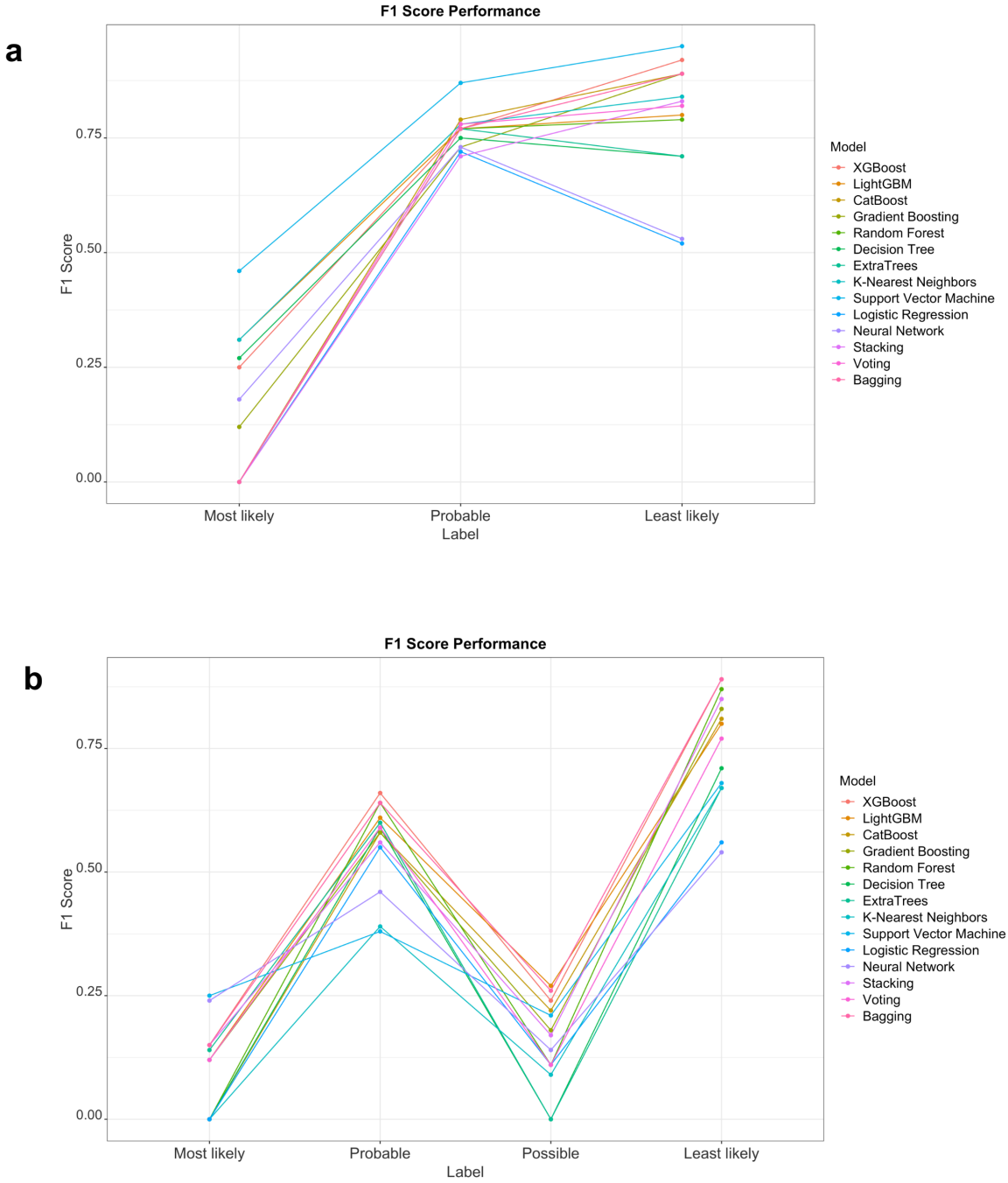


Figure 3.2. F1 Score Performances for all models on 3-labelled and 4-labelled training data. a) F1 score performance on the 3-labelled training data (n=293), b) F1 score

performance on the 4-labelled training data (n=377). Lines connect the labels for each model to visualise the model's ability to predict each class.

3.3.3 Three Label Oversampling versus Class Weighting Performance

From the 3-label versus 4-label comparison, the better-performing 3-label training data was explored in further analysis. SMOTE oversampling compared with class weighting found SMOTE to give a higher performance for all models (Table 3.2, Figure 3.3). However, class weighting shows more conservative benchmarking results (e.g., 71% median balanced accuracy of the top-performing model CB, Table 3.2).

	Three Label Oversampling					Three Label Class Weighting				
Model	**Accuracy**	**Balanced Accuracy**	**F1**	**Precision**	**Recall**	**Accuracy**	**Balanced Accuracy**	**F1**	**Precision**	**Recall**
XGB	0.82	0.82	0.82	0.83	0.82	0.72	0.69	0.72	0.72	0.72
LGBM	0.82	0.82	0.82	0.84	0.82	0.73	0.7	0.73	0.75	0.73
CB	0.79	0.79	0.79	0.8	0.79	0.73	0.71	0.72	0.73	0.73
GBM	0.84	0.84	0.84	0.85	0.84	0.72	0.68	0.72	0.3	0.72
RF	0.78	0.77	0.77	0.77	0.78	0.69	0.66	0.69	0.7	0.69
DT	0.69	0.69	0.69	0.71	0.69	0.71	0.62	0.69	0.71	0.71
ET	0.7	0.7	0.68	0.72	0.7	0.65	0.67	0.63	0.7	0.66
KNN	0.74	0.74	0.73	0.74	0.74	NA	NA	NA	NA	NA
SVM	0.71	0.71	0.71	0.71	0.71	0.68	0.67	0.68	0.73	0.68
LR	0.55	0.55	0.55	0.55	0.55	0.5	0.51	0.49	0.51	0.5
NN	0.71	0.7	0.69	0.7	0.71	0.64	0.68	0.62	0.69	0.64
Stacking	0.82	0.82	0.82	0.84	0.82	0.72	0.68	0.72	0.74	0.72
Voting	0.82	0.82	0.82	0.82	0.82	0.69	0.65	0.68	0.7	0.69
Bagging	0.83	0.83	0.83	0.83	0.83	0.75	0.69	0.74	0.75	0.75

Table 3.2. Model performance using oversampling on training data with three labels. Each model was benchmarked on the 3-label training dataset, testing performance on oversampling (n=447) or class weight adjustment (n=293). Each model was assessed across accuracy, balanced accuracy, F1, precision and recall metrics. The fourteen models were benchmarked: extreme gradient boosting (XGB), gradient boosting (GBM), CatBoost (CB), LightGBM (LGBM), random forest (RF), decision tree (DT), Extratrees (ET), K-nearest neighbours (KNN), support vector machine (SVM), logistic regression (LR), neural network (NN), and three meta-ensemble methods – stacking, bagging, and voting models.

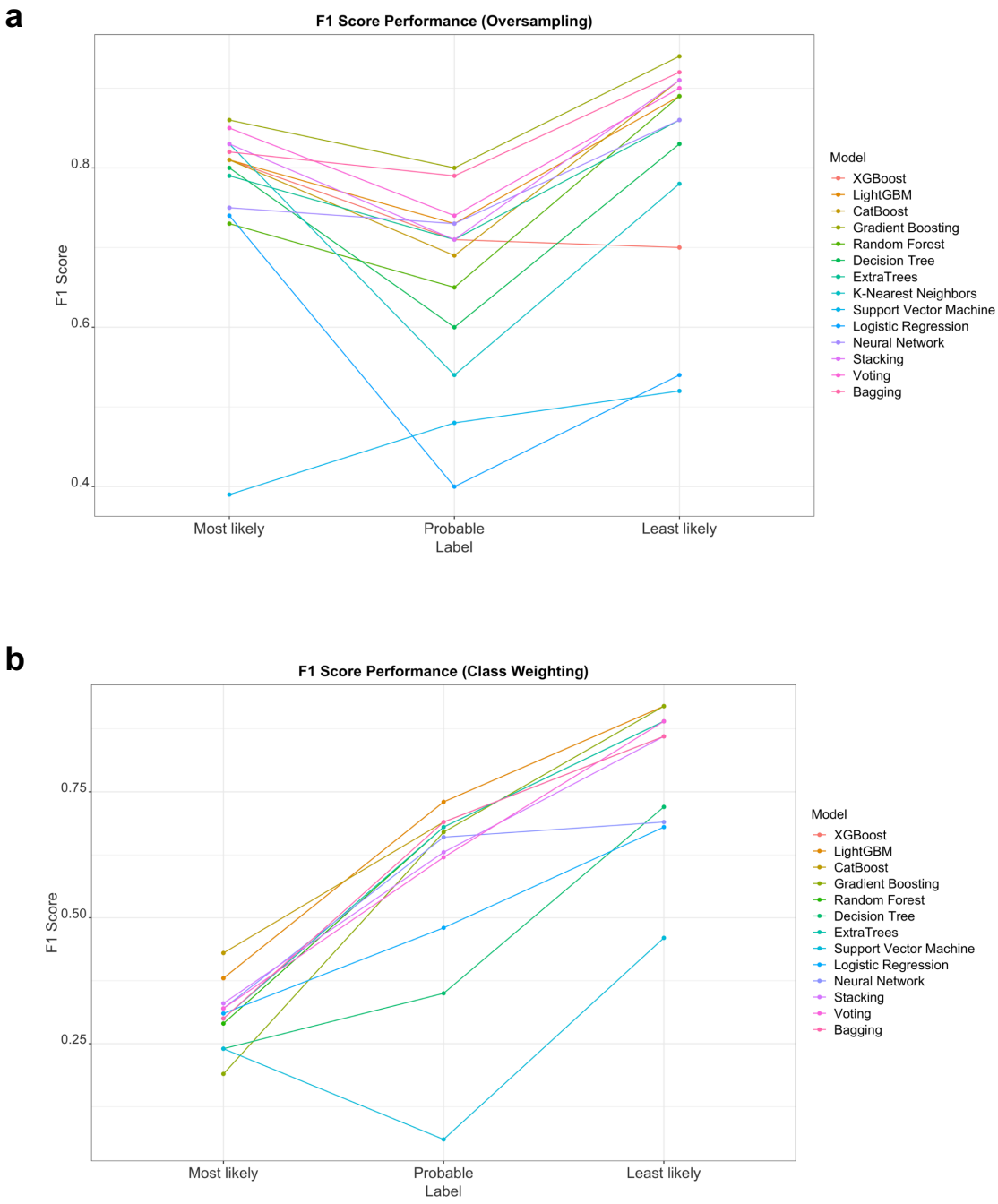


Figure 3.3. Comparison of F1 scores for all models predicting three labels after oversampling. a) F1 score performance on the oversampled training data (n=447), b) F1 score performance on the class weighted training data (n=293). Lines connecting the four labels for each model visualise the model's ability to predict each label.

Despite oversampling's improved model performance, a more conservative choice was made to select the class weighting approach to take on to downstream analysis. Class weights adjust model regularisation whilst SMOTE adjusts the data directly (thereby directly adjusting the generalisability of the model to value-specific data patterns that may not be representative of the minority class the sampling was taken from). Class weighted model performances identified CatBoost as the top-performing model (71% median balanced accuracy on 10-fold cross-validation) (Table 3.2, Figure 3.4), which was closely followed by LGBM and XGB (70% and 69% median balanced accuracy respectively).

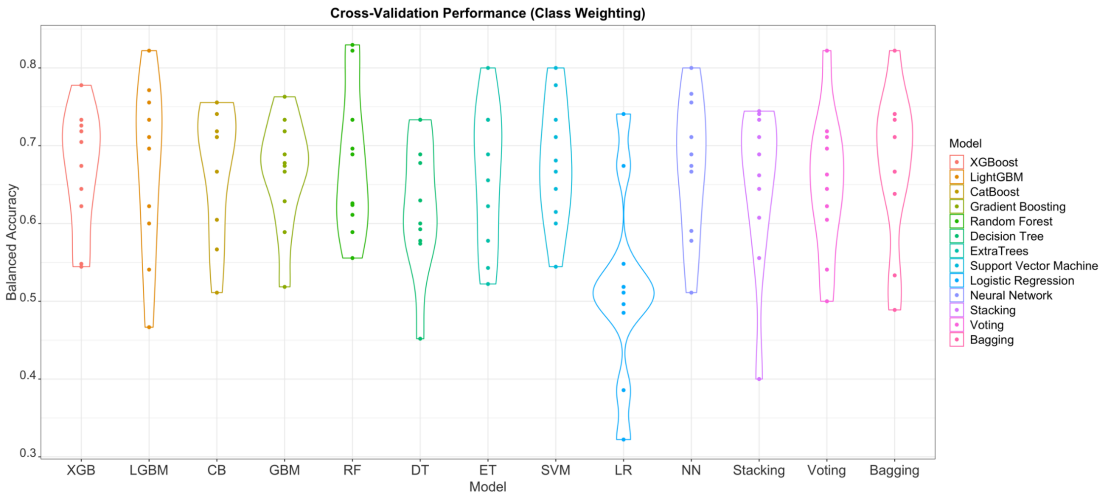


Figure 3.4. Model benchmarking performance on 10-fold stratified cross-validation. Fourteen models were benchmarked: extreme gradient boosting (XGB), gradient boosting (GBM), CatBoost (CB), LightGBM (LGBM), random forest (RF), decision tree (DT), Extratrees (ET), K-nearest neighbours (KNN), support vector machine (SVM), logistic regression (LR), neural network (NN), and three meta-ensemble methods – stacking, bagging, and voting models. The model performance was assessed on stratified 10-fold nested cross-validation.

3.3.4 CatBoost Model Interpretation

The CB model was further investigated to interpret model decision-making. Class weight balancing was applied followed by probability calibration of CB's predictions, improving performance (0.509 log loss on the uncalibrated probabilities versus 0.452 log loss on calibrated probabilities) (Figure 3.5). Confusion matrices identified the model's strongest ability to predict least likely genes (aligned with the F1 scores also being highest for this group), followed by probable and most likely genes (Figure 3.6). The model's performance on the test data showed its conservative approach to predicting most likely genes (Figure 3.7b). CB also provides internal feature importance interpretation, showing HIPred as the most important feature overall (Appendix B Table 2). SHAP plots were used to show feature interpretation for each gene class, finding the mouse Exomiser score was most important to identify most likely genes, and HIPred was most important for probable and least likely genes (however with HIPred values having opposite directionality for its importance in each gene group).

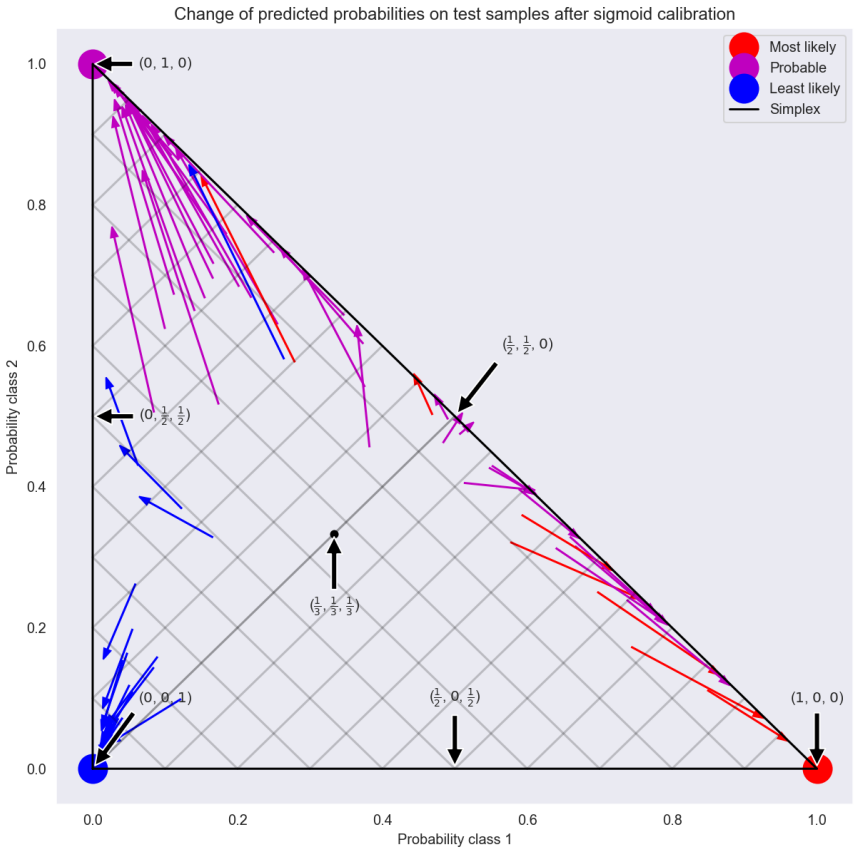


Figure 3.5. Probability calibration of the fitted CatBoost model. Simplex plot shows each class as a vertex of the simplex with perfectly predicted classes represented by one of three dots (e.g., the red dot is the perfect most likely prediction with probabilities of 1, 0 and 0 for each respective class). Each vector starts at the uncalibrated probability for that gene made by CatBoost and the end point of the vector/the arrowhead is at the calibrated probability. The colour of the arrows represents the classes (red for most likely, purple for probable and blue for least likely genes).

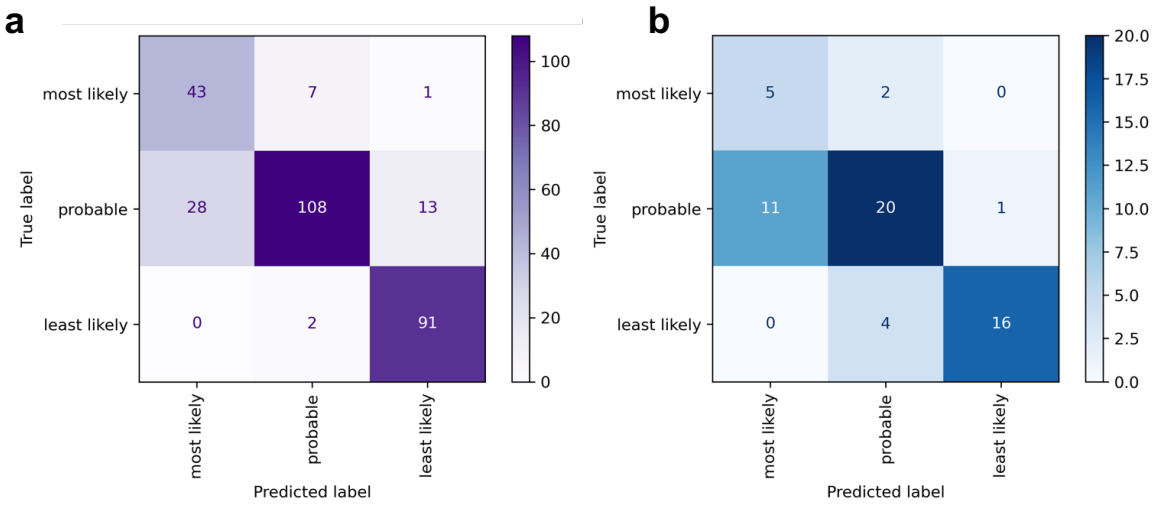


Figure 3.6. Training and test data predictions by CatBoost. Confusion matrices of the model's predictions for all the training data (a) and the test dataset (b).

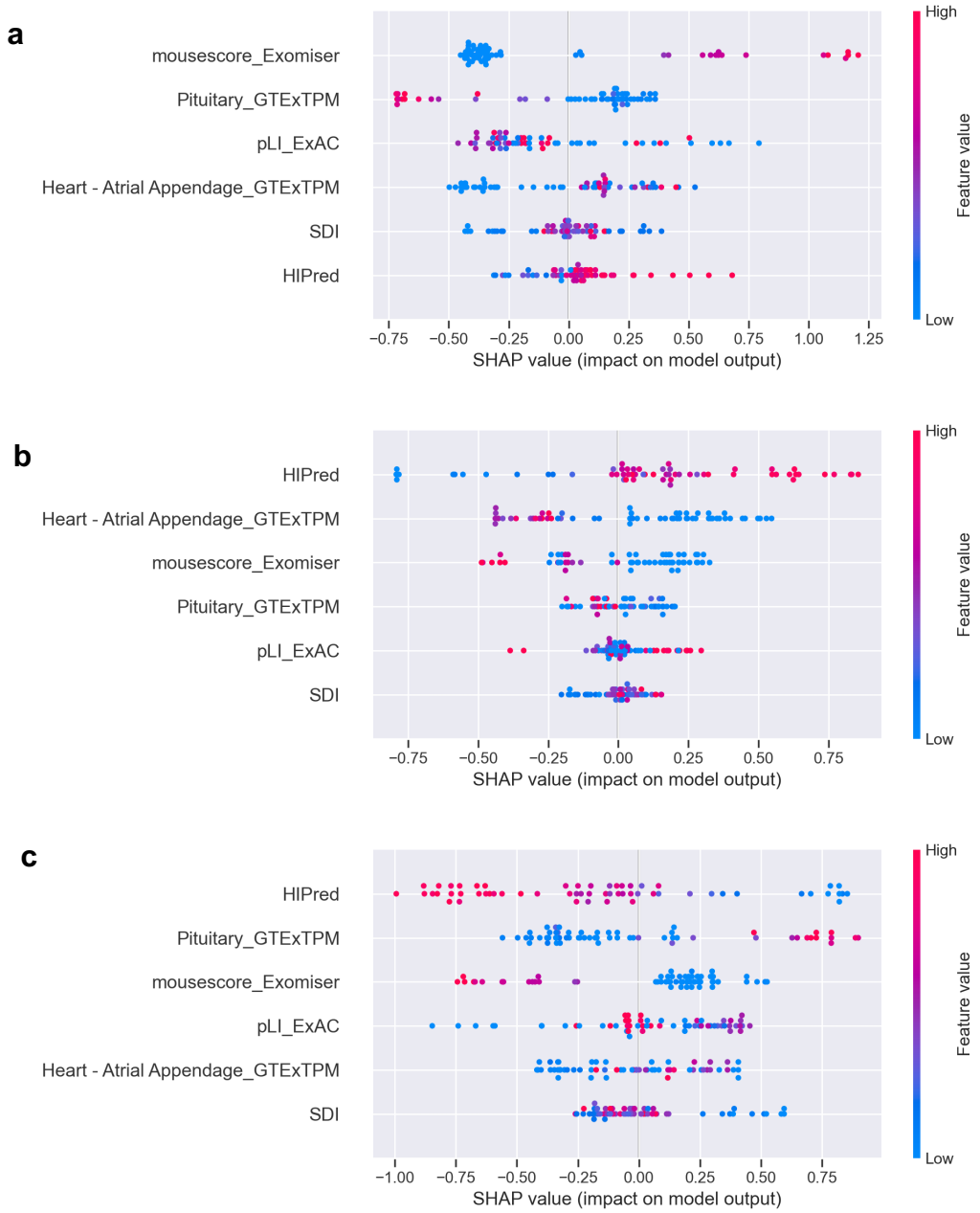


Figure 3.7. SHAP summary plots of each label by XGB. Feature interpretations by CatBoost were measured via SHAP values. Summary plots of each feature interpretation are for most likely BP genes (**a**), probable genes (**b**), and least likely genes (**c**). The SHAP value on x-axis indicates the direction of model influence from that feature for each gene (e.g., a higher SHAP value indicates a more positive output model score). The colour-coding of points (genes) indicates whether their feature

value was high (red) or low (blue), and the ordering of features on the y-axis is by descending feature importance.

3.3.5 CatBoost Gene Prioritisation

From selecting CB as the top-performing model, it was fitted to 293 training genes to then classify the 1,804 *BP-genes* that were unlabelled – with the fitted model using class weight balancing and calibrated probabilities. In total, CB classified 223 genes as *most likely*, 1069 genes as *probable*, and 512 genes as *least likely* to affect BP (Appendix B Table 3). Individual investigation of the top prioritised genes showed *most likely* and *probable* predicted genes had publications and animal models with cardiovascular and renal phenotypes (Table 3). The top three *least likely* predicted were olfactory genes. However, due to heavy missingness in the predicted *least likely* genes and 172 genes having almost all the exact same imputed values (Appendix B Table 4), 172 genes had the same probability of 0.9397 to be classed as *least likely* with the fourth highest probability in that class.

Most Likely Classified Genes				
Gene	**Most likely Prediction Probability**	**Probable Prediction Probability**	**Least likely Prediction Probability**	**Gene Description**
MLIP	0.926	0.058	0.017	Muscular LMNA Interacting Protein; unknown protein function and associated with cardiac abnormality in mouse models[117].
JPH2	0.847	0.134	0.019	Junctophilin 2; component of junctional complexes and associated with impaired cardiac contractility in mouse models[118].
NOTCH3	0.836	0.149	0.0149	Notch receptor 3; receptor for ligands that regulate cell-fate determination[119] and associated with the adaptive response of vasculature in mouse models[120]
MRVI1	0.83	0.159	0.032	Murine Retrovirus Integration Site 1; lymphoid-restricted protein associated with acting as a tumour suppressor gene[121].
CSRP3	0.814	0.153	0.016	Cysteine and glycine rich protein 3; cytoskeletal protein and associated with cardiomyopathy in mice[122].
Probable Classified Genes				
Gene	**Most likely Prediction Probability**	**Probable Prediction Probability**	**Least likely**	**Gene Description**

			Prediction Probability	
LEF1	0.06	0.92	0.01	Lymphoid enhancer binding factor 1; transcription factor associated with Wnt signalling and several mouse phenotypes including abnormal heart morphology (MGI:96770).
SYT1	0.056	0.919	0.025	Synaptotagmin 1; membrane protein of synaptic vesicles associated with preweaning lethality (MGI:99667) in mice and carcinogenesis[123]
GRM7	0.48	0.916	0.368	Glutamate metabotropic receptor 7; G protein-coupled receptor associated with embryonic neurogenesis[124].
GRM4	0.059	0.915	0.026	Glutamate metabotropic receptor 4; G protein-coupled receptor associated with major depressive disorder[125].
MPPED2	0.063	0.913	0.023	Metallophosphoesterase Domain Containing 2; associated with tumourgenesis[126].
Least Likely Classified Genes				
Gene	**Most likely Prediction Probability**	**Probable Prediction Probability**	**Least likely Prediction Probability**	**Gene Description**

OR5AS1	0.0145	0.0376	0.948	Olfactory receptor family 5 subfamily AS member 1; involved in olfactory signalling.
OR5I1	0.0145	0.076	0.948	Olfactory receptor family 5 subfamily member 1; involved in olfactory signalling.
OR5B12	0.016	0.037	0.947	Olfactory receptor family 5 subfamily B member 12; involved in olfactory signalling.

Table 3.3. Top genes for each class predicted by CatBoost. Each gene has a predicted probability per class made by CatBoost for *most likely, probable* and *least likely* classes. The highest probability out of the three classes is used to then assign a gene to that probability's class. The top five genes with the highest probabilities for the *most likely* and *probable* classes had their gene functionality described. The *least likely* gene class had only the top three genes described due to 172 genes having the same probability (0.9397) to be potentially ranked fourth.

The genes with the highest probabilities for being classed as *most likely* and *probable* (*MLIP* and *LEF1* respectively) had their ML prediction visualised via SHAP (Figure 3.8). Their difference in influencing features for model decision-making highlight the inverse influence of HIPred for each of the two classes. The higher the HIPred scores strengthening probable gene classification, whilst the lower HIPred scores influence most likely gene prediction. This different use of the feature per class by the model is also shown by the distribution comparison across all predicted classes for the feature (Figure 3.9).

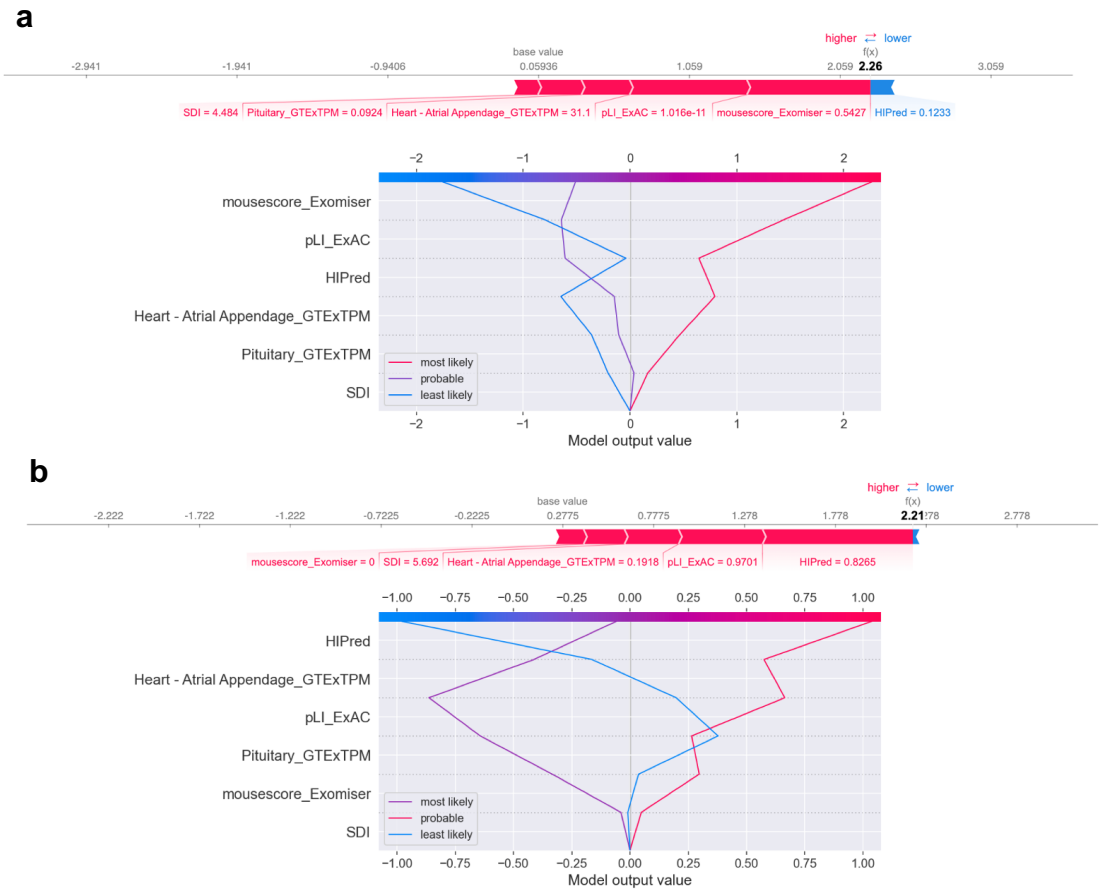


Figure 3.8. Shapley interpretation of predictions for *MLIP* and *LEF1*. Both (**a**) and (**b**) provide SHAP force plots and decision plots of CatBoost decision-making for prediction of individual genes *MLIP* (**a**) and *LEF1* (**b**). The horizontal force plots show the directionality of influence each feature had on model decision-making for each gene's predicted class, and the decision plots show the feature influences for all classes (*most likely*, *probable* and *least likely* model predictions).

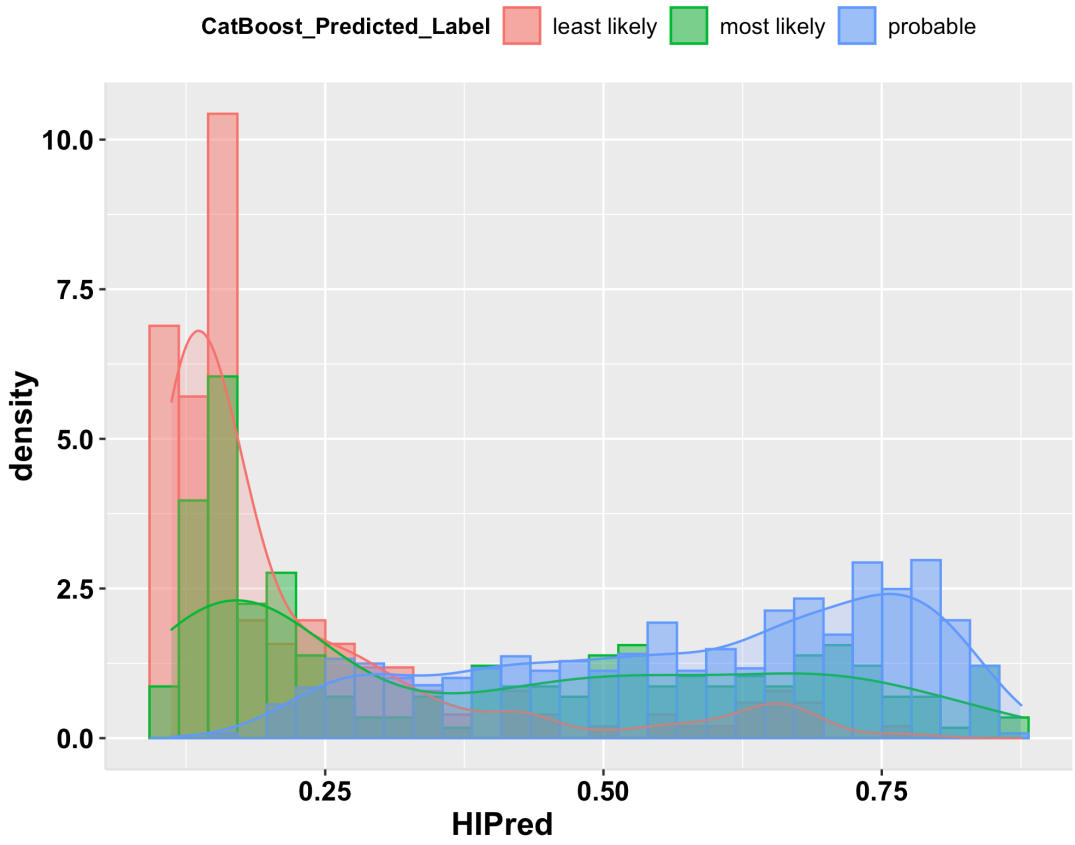


Figure 3.9. HIPred distribution comparison across predicted classes. The distributional difference of HIPred between the predicted gene groups (223 genes predicted *most likely* in green, 1069 genes predicted *probable* in blue, and 512 genes predicted as *least likely* in red).

The *most likely* predicted genes (n=223) acted as a gene group for enrichment analysis comparisons. KEGG analysis of that predicted class, alongside sentinel genes from the GWAS, and the training genes classed as *most likely* and *probable*, were compared. Pathway analysis showed less significant enrichment for known BP pathways across all three predicted gene groups (Figure 3.10). The *most likely* gene group had its most significant enrichment for cardiovascular and renal pathways matching the BP training

genes, however, the BP training genes were more significantly enriched across all pathways

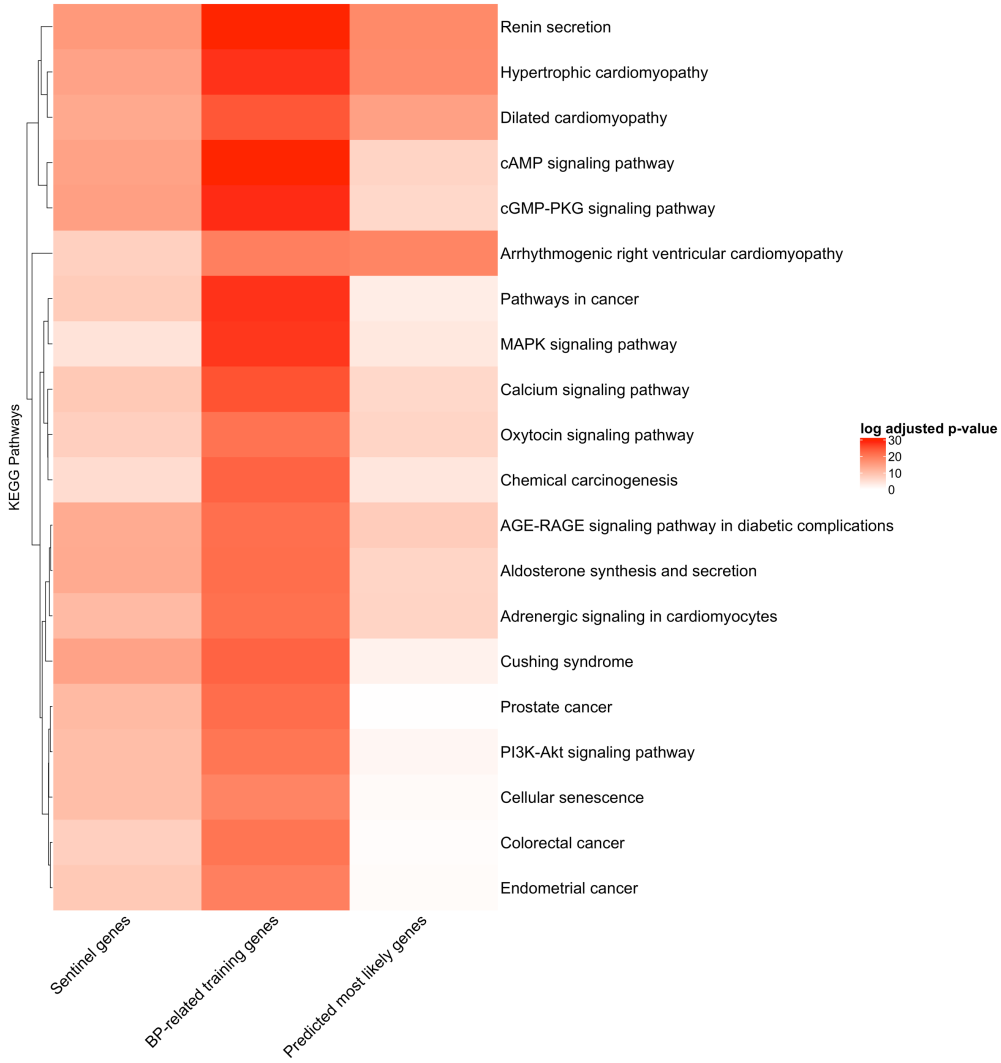


Figure 3.10. Pathway analysis of classified most likely blood pressure genes. Heatmap of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. The heatmap shows more significant values are indicated by darker shades of red. The heatmap compares three gene lists, composed of sentinel genes (identified by the Evangelou et al. (2018) genome-wide association study), BP-related training genes

(*most likely* and *probable* labelled training genes), and genes predicted as *most likely.* The top 20 most significant pathways for the BP training genes group were visualised.

3.4 Discussion

The training data curated here gave two approaches to multiclass classification, 3-labelled and 4-labelled datasets, providing ML that learnt from a stratified scale of example genes. The 3-label approach proved to give a better performance for all models, indicating that the 4th added possible gene group increased the difficulty of prediction. This lower performance could be due to the possible gene group adding more noise to the selected features than recognisable and distinguishing patterns. To investigate this, I visualised the 4-labelled pairwise feature distributions on exploratory data analysis (chapter 2 Figure 2.7b) showing several features with minimal distinguishing differences between each group. In contrast, the 3-label training data showed almost all of the selected feature's distributions had different distributions for each curated gene group (only excluding the pituitary gene expression feature) (chapter 2 Figure 2.7a). The *possible* gene group may have also created difficulty in ML prediction by the genes themselves and their curation using IPA not being robust enough to use standalone as a class identifier. IPA was chosen due to the resource curating experimental analysis of whether each gene has had BP study, giving the possible gene group functional evidence to create its category. However, its poor performance as a gene grouping criterion, it not being selected as a feature in the 3-label classification, and the added fact that IPA's data is collected from behind a paywall overall suggests that it is likely not a worthwhile data point to collect for ML.

The better performance of 3-label classification also suggests binary classification may improve ML performance, as two classes with more significant differences (i.e., *most likely* versus *least likely* genes) may give a more confident model. However, due to the small sample size of *most likely* BP genes (n=51), this was not investigated as there is minimal ability to expand upon this group without rigorous experimental validation and provide a ML model with more diverse knowledge.

Overall, the ML performance on the 4-labelled training data indicates that the possible gene group hindered ML and needed either to be removed, further curated, or integrated with the probable BP gene group. However, integration with the probable gene group would have also amplified the class imbalance. Stratified training data curation beyond binary classification is an ongoing area of research as multi-labelling between positive and negative examples is often difficult to curate with a strong evidence base that justifies having a gradient of classes. Whilst most gene prioritisation studies focus on binary classification[50] or positive-unlabelled learning[33], this issue is an ongoing area of ML research in general. For example, packages are being developed such as Snorkel[127] that focus on programmatically labelling and managing groupings within training data, providing a statistical evidence base for groupings. However, thus far this tool has focused on natural language processing and medical imaging ML problems[127] and further development would be needed to re-direct such tools to tabular biological data.

On taking the 3-label training data into further analysis, SMOTE oversampling and class weighting were compared to address the underlying class imbalance. SMOTE creates synthetic example data points, doing so by duplicating data points that are closest to decision boundaries for their labels (avoiding duplicating data points that are further from the decision boundary and therefore easily defined, and so avoiding overfitting on well-understood patterns). However, SMOTE has an unavoidable overfitting risk if the data is non-linear and therefore all data points are not well defined[128], causing patterns to be amplified in the data that the model may overfit to. The SMOTE approach benchmarked here has the highest performance for any ML iteration tested, suggesting a potential risk of overfitting that is also overriding metrics used to catch overfitting such as F1, precision and recall.

Meanwhile, class weighting is a cost-sensitive approach that involves assigning greater penalties for each model when it misclassifies the minority class, forcing the model to make greater adjustments to try and more accurately predict that gene group. Whilst this does not directly alter the training data, class weighting still also provides an overfitting risk if the model is overestimating the value of minority class distributions that are not representative of the true population. On comparing these two approaches to address the class imbalance, class weighting shows more conservative benchmarking results (e.g., 71% median balanced accuracy of the top-performing model CB, Table 3.2), which led to it being the method chosen for further downstream analysis when supported by probability calibration. However, ultimately the underlying class imbalance is still a limitation within the training data, and both oversampling and class reweighting by design are creating new local minima in the

cost function of ML models that may not be optimal. Testing the reliability of class-balancing would be best performed by having an external validation training set of more *most likely* BP-regulator genes, however for most disease-gene prediction problems these genes are laborious and difficult to identify. Due to the lack of such known BP-regulating genes this validation test set could not be curated and in the future, these genes will be crucial for validating any chosen prioritisation model and its findings.

Using the 3-label training dataset and class weighting approach, the model benchmarking of the best approach showed CB to be the top-performing model alongside most other models having comparably high performances (between 65% to 71% median balanced accuracy) except for LR (51% balanced accuracy). The lower performance of LR suggests that the heterogeneity of the training data needs a more complex model that can test several hypotheses to understand the gene groupings, as shown by the higher-performing ensemble models. This result is further emphasized by the F1 scores for all ML iterations tested, where LR consistently had a score of 0 for the *most likely* gene group (Figures 3.2 and 3.3) and had the lowest F1 scores for each class on class-weighted data (Figure 3.3).

CB being the top-performing model aligns with studies benchmarking models on multiple dataset types, finding CB (alongside XGB) is a top performer across supervised learning problems[129]. CB is also unique in its algorithmic principles that are designed to combat overfitting, for example by adding an extra regularisation parameter (Bayesian matrix regularisation), suggesting its 71% balanced accuracy has

more reliability than other models. From exploring its feature interpretation, HIPred and mouse Exomiser scores were the most important features (Appendix B Table 2), indicating the models understanding of the BP phenotype and gene functionality. However, the output predictions by CB indicate a degree of confusion in interpreting the model's classifications. For example, the *most likely* predicted genes have the potential for further research. *MLIP* (predicted with a 0.926 probability of being a most likely BP gene) has been shown to induce cardiac hyperactivation in knockout mouse models of the gene[117]. However, from the top five most likely genes to the best of our knowledge only *NOTCH3* has had functional research concluding it may play a role in BP[120]. Meanwhile, the probable gene group showed genes that are also strong contenders, such as *LEF1* which has been shown to modulate the expression of angiotensin II[130], bringing into question how CB is distinguishing these genes from being classified as probable instead of *most likely*. For *MLIP*, SHAP shows the mouse Exomiser and pLI scores are the most important features, meanwhile for *LEF1* HIPred is the most important feature followed by the atrial heart gene expression (Figure 3.8). Notably, *MLIP* has a lower HIPred score than *LEF1* (0.1233 versus 0.8265 respectively), which may be affecting how the two differ in classification, as SHAP shows the higher HIPred score drives up *MLIP*'s probability of being a *probable* BP gene. This difference in HIPred aligns with the summary SHAP values showing HIPred as the least important feature for most likely gene predictions and the most important for probable gene predictions. The distributional difference of HIPred between the two predicted gene groups (223 genes predicted *most likely* versus 1069 genes predicted *probable*) shows the *probable* predicted genes have higher HIPred scores (Figure 3.9) and that *most likely* genes are more similar to *least likely* predicted

genes with lower HIPred scores. These results conflict with the domain biology expectation that more likely disease-causing genes would be more often haploinsufficient and have greater functional impacts on loss-of-function, suggesting a misinterpretation by the model as it tries to discern categorically between *most likely* and *probable* BP genes, when their differences may be more opaque than multiclass classification can capture.

In further contrast, for 172/1,804 predicted genes that had similar missingness patterns (Appendix B Table 4) that led to all 172 genes having the same exact probability and *least likely* classification, making it impossible to discern between them. This result highlights the difficulty in the *least likely* gene classification, as such genes are, by their unexplored nature, less likely to be as well annotated as genes more likely to impact disease.

The gene enrichment analysis also presents difficulty in analysing the gene groups to identify genes pathways worth further investigation. The *most likely* predicted genes had less significant gene enrichment in comparison to the BP training genes, and pathways that were the most significantly enriched were established cardiovascular and BP pathways (e.g., renin secretion, dilated and hypertrophic cardiomyopathies, etc.) (Figure 3.10). Alongside the circularity of pattern recognition, this issue is potentially impacted by the classification ML approach itself. The multiclass classification does not treat the labels as ordinal, possibly leading to a misclassification between the gene classes as the differences between the groups is not distinct enough to have them act as categorical groupings.

Overall, the class imbalance and difficulty of distinct gene labelling in the training data greatly impact the multiclass classification approach. The model's predictions suggest each grouping of prioritised genes has various levels of evidence linking the genes to BP in predominantly cardiovascular pathways. These conclusions suggest the multiclass classification developed here requires further comparison with other methods to ensure an optimised approach is being applied to prioritise BP genes.

4 Regression for Prioritising Blood Pressure Genes

4.1 Introduction

In this chapter, I applied fourteen ML algorithms and benchmarked them on the BP-GWAS[47] data curated previously, however, reframing the methodology from a classification problem to a regression analysis. I explore the genetic landscape of the top-performing model's highly prioritised genes, investigating the prioritised genes at their loci, and compare the ML method's concordance with other ML prioritisation methods, providing a stronger evidence-base for how GWAS results fit into the broad biology of blood pressure.

One aspect that is the most time-consuming in supervised machine learning is training data curation. Having representative and plentiful examples is pivotal for a model's training, however, these qualities can be difficult to ensure when it comes to identifying disease-causing and non-disease-causing genes. With methods, such as OpenTargets Genetics' Locus 2 Gene score taking several filtering steps to identify gold standard positive and gold standard negative genes[115]. For example, their training gene curation involved collecting 400 gold-standard positive loci by identifying: loci overlapping with drug target-disease pairs, loci with strong orthogonal evidence, loci with functional follow-up, and loci inferred from observational functional data[84]. These distinctions would then classify loci between high, medium, and low gold standard quality. They also identified gold-standard negatives, doing so by finding genes that were not within 500kb of any positively labelled locus and that had low to

no PPIs with the positive labelled genes. Whilst gold-standard positive genes can be difficult to identify due to lack of causal evidence, it could be argued negative gold standards are even harder to find as their criteria in most cases hold a risk of false negatives and biased labelling criteria – a point which is made in studies opting to use positive-unlabelled learning[19]. This is particularly true for the use of PPI data where a lot of interactions have not been experimentally proven, or for when an interaction with a causal gene does not necessarily mean causality for the connected gene in the PPI. Negative labelling is also made difficult by the lack of research focusing on non-causal/negative gene identification. However, a ML model with no distinction of negative examples risks a more limited understanding of gene diversity and potentially increases the risk of false positive predictions.

Ultimately the pros and cons of negative labelling for non-disease-causing genes suggest stringent criteria are needed to use them. However, it also presents an opportunity for an alternative approach using gene scoring and applying a regression analysis machine learning approach – which has had little to no exploration in post-GWAS prioritisation, although it has been used in other bioinformatics applications such as predicting protein expression levels[131] and in time-series analysis of gene expression[132]. In ML, regression is used when the measurement of the target variable is continuous, and classification is applied when the target variable is categorical and usually has no natural order. Regression analysis applied to the BP training data curated here would avoid the prioritisation of genes into fixed categories and allow a greater degree of freedom for prioritisation on a continuous scale that accounts for all three gene groupings having an order from most to least likely BP genes. Also, giving

all prioritisations in one output ranking as opposed to the three separate probabilities given in multiclass classification. Furthermore, a regression analysis approach combats the difficultly of non-disease-causing gene curation in the training data, as these genes can be assigned a score reflecting the uncertainty of their grouping (e.g., having a score of 0.1 as opposed to 0 on a scale between 1-0 of most to least likely BP genes).

In comparison the previously applied multiclass classification aimed to prioritise genes based on predefined classes, dividing any new data space into three discrete groupings based on probabilities. This approach is less interpretable than regression as it outputs probabilities based on model weights and biases, whilst regression can be expressed as an equation that uses coefficients, which in turn can be directly interpreted to understand how the input features change the output model prioritisation. Also, in the direct comparison between multiclass classification and regression, regression can consider ordinal data while the classification applied in chapter 3 will only consider categorical groups. From a ML perspective this allows for the loss function calculated on each training data iteration by a model to consider ordinal error rates[133], as opposed to only correct or incorrect classifications, giving a model a better chance of getting closer to the true values of ordinal groups. These benefits of a regression analysis justify its exploration in gene prioritisation. However, it should be noted that, as regression is not predicting predefined classes and is instead aiming for prioritisations as close to the original/true prioritisation as possible, it cannot be directly aware of class imbalance and so is still susceptible to overfitting despite its advantages[134].

In this chapter. I explore the optimisation of a regression analysis ML framework, informed by the results of chapters 2 and 3, to finalise a methodology that can prioritise most likely BP genes post-GWAS for further investigation.

4.2 Methods

4.2.1 Data Collection and Pre-processing

From exploratory data analysis to output model scoring of the trained top-performing model, all steps of the machine learning pipeline were conducted using Python (v3.8.5). The BP GWAS data that underwent pre-processing and feature selection was fully described in chapter 2. From the 7 million variants that were annotated to genes, these genes had 114 features collected that were assessed for missingness and correlation, undergoing feature cleaning and selection described in chapter 2. Features were removed if found to be missing for all genes by >25%. Features with a pearson's correlation coefficient >0.9 and not meeting in i.i.d assumptions between train and test data were also removed (i.i.d testing found in https://github.com/hlnicholls/PhD-Thesis/tree/main/Chapter4/Machine%20learning/).

Different correlation thresholds (0.85 and 0.99) were also tested for feature cleaning (test runs included in https://github.com/hlnicholls/PhD-Thesis/tree/main/Chapter4). The initial feature removal resulted in 20 features entering feature selection. The 20

features were imputed using random forest imputation (using the missingpy package, v0.2.0) and underwent feature selection using the BorutaShap package (v1.0.13).

4.2.2 Training Data

Genes used in the training data, that were labelled with one of 3 gene groupings between most likely, probable and least likely to affecting BP in previous chapters 2-4, were scored with values between 0 to 1 for regression analysis. Firstly, the genes known to interact with BP drugs, curated by an expert in the cardiovascular field (herein referred to as *BP-regulator* genes, that were labelled as most likely BP genes in multiclass classification) were scored at 1 (Appendix A Table 4). Genes were assigned a score of 0.75 if they were considered probable to affect BP (genes labelled as probable in multiclass classification, herein referred to as *text-mining genes* Appendix A Table 6). Finally, the genes labelled as least likely BP genes, as defined in chapter 2, were given a score of 0.1.

These three scorings provided 293 training genes (51 *BP-regulator genes* scored at 1, 149 *text-mining genes* scored at 0.75, and 93 least likely BP genes scored at 0.1) (Appendix C Table 1). Scores 1, 0.75 and 0.1 were designated to reflect the degree of certainty provided by the grouping criteria, with multiple scoring intervals tested on ML performance (such as 1, 0.6 and 0.1; 1, 0.5 and 0; 1, 0.75 and 0.1 – Appendix C Table 2, with all benchmarking test runs also included in https://github.com/hlnicholls/PhD-Thesis/tree/main/Chapter4). Each scoring scale for the three gene groups had fourteen ML models benchmarked - further described in the

next section (4.2.3) - with the best ML performance selecting the final scoring to use in further analysis.

4.2.3 Machine Learning Model Benchmarking Methods

Fourteen models were benchmarked: random forest, gradient boosting, extreme gradient boosting, catboost, lightgbm, decision tree, extratrees, k-nearest neighbours, support vector regressor, linear regression using elastic net and LASSO, a voting model and a stacking model both consisting of all other models, and a bagging model using the top-performing model (extreme gradient boosting). All models except extreme gradient boosting, lightgbm, and catboost were applied using scikit-learn (v0.23.2), and extreme gradient boosting was applied from the xgboost package (v1.2.0), lightgbm using the lightgbm package (v3.3.2), and catboost using the catboost package (v1.0.6). Each model underwent hyper-parameter tuning using a Bayesian optimisation to tune hyper-parameters and nested 5-fold cross-validation repeated three times - giving 15 model performances to take average and median assessments from. The 15 folds of training and test data underwent i.i.d (independent and identically distributed) assumptions testing using the kolmogorov-Smirnov test, finding only GDI scores to have significant differences – with this feature not being selected for final model benchmarking. Model performance was evaluated with r^2, predicted r^2, mean squared error, mean absolute error, and explained variance to select the top-performing model for further analysis. The top-performing model was then chosen to prioritise the remaining *BP-genes* that were not in the training data

(n=1,804), scoring the genes on a continuous scale. The output prioritised genes could then undergo downstream analysis (described in 4.2.4).

The top-performing model, XGB, also underwent interpretation using the SHAP package[26] (v0.36.0), providing feature importance values both globally for overall model performance and individually for each gene. Plots of the feature importance (for both overall predicting and individual predictions) were created alongside feature-feature interactions.

4.2.4 Gene Prioritisation Analysis

After ML prioritisation, an algorithm was developed to select the top prioritised gene per locus (genes within a 500kb+/- region of sentinel SNPs – SNPs with a GWAS p-value $<5x10^{-8}$). This algorithm enables tiebreaking gene selection for when XGB scores are close (e.g., < 0.01 difference between scores) for multiple genes in a locus and offers a failsafe step to select genes that may have false negative prioritisation by XGB but still have strong evidence of a BP relationship. This strategy combines the XGB scores with supporting PPI information (PPI data not used in the model) to select the most likely BP gene(s) at a locus. The algorithm consisted of seven steps:

1) If there is a training gene scored at 1 (a gene labelled as most likely to impact BP) in the locus that gene is retained.

2) The top-scored gene per locus is selected if the score is greater than +1 standard deviation (SD) of the ML model score distribution for all genes at that locus.

3) If no genes are more than +1 SD and only one gene has a score greater than the average score of that locus, then that gene is selected.

4) If multiple genes have a higher-than-average score at their locus, then all genes with scores larger than the average are selected to be compared with PPI filtering.

5) The gene with the largest number of PPIs directly with known BP genes is selected.

6) If more than 1 gene has equal direct PPIs, then the gene with the largest number of secondary PPIs of BP genes (interactors of the gene that interact with interactors of the known BP genes) is selected.

7) If the genes have both equal direct and secondary PPI counts, then the multiple genes are all selected for that locus.

All genes prioritised (including training genes scored at 1 and 0.75) entered this gene selection algorithm. Application of this selection algorithm gave 768 loci with 794 genes selected to enter enrichment analysis (19 loci having more than one gene selected per locus). These 794 genes selected per loci are herein referred to as the "*selected-genes*".

I investigated the genes scored > 0.8 by XGB - herein referred to as the "*highly-scored genes"* - and the *selected-genes* in downstream analysis by investigating their distributional differences for several collected annotations using the Mann-Whitney U

test in R and plotting their gene expression across all tissues in GTEx (v8) using ComplexHeatMap (v2.6.2). Clustered gene identified on plotting GTEx analysis were further explored in STRINGdb (v11.5). The R package GeneOverlap (v1.30.0) was used to perform hypergeometric tests on gene hits in IMPC mouse model phenotypes, testing the overlap of gene hits for the *highly-scored genes* and the *selected-genes* against the total number of genes in each phenotype in comparison to a total 1,875 genes annotated in IMPC. I also explored gene enrichment using the R package enrichrplot (v1.14.1).

I collected Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways from the 2021 KEGG database within Enrichr[116] and compared pathway enrichment in four groups: the *selected-genes,* the *highly-scored genes,* the genes containing sentinel SNPs[47] GWAS, and the BP genes used in training (*BP-regulator genes* and *text-mining* genes scored at 1 and 0.75 respectively). KEGG gene enrichment analyses were visualised using ComplexHeatMap (v2.6.2). On investigating the interacting genes within the most significantly enriched pathways, I plotted pathway interactions and overlaid the druggability of genes from the Drug Gene Interaction database (DGIdb), showing interactive genes in BP pathways that are also potential drug targets.

Gene-drug interactions were taken from DGIdb and drugs with BP side effects were identified using Side Effect Resource (SIDER)[135] and the British National Formulary (BNF). Genes with encoded protein-protein interactions with proteins involved in drug mechanisms were recorded using STRINdb. Total DGIdb recorded drug interactions

were plotted using Circlize (v0.4.15) in R. SIDER was used to extract all the drugs with BP side effects by searching for 'hypotension' and 'hypertension' terms in the data resource. 555 of drugs with hypertension side effects were downloaded and 660 drugs for hypotension were downloaded. These drugs were queried in DGIdb to identify their gene-drug interactions, identifying genes also prioritised by XGB. DGIdb collates crowdsourced databases, each with varying methods of curating gene-drug interactions[136]. The level of validation differs between gene-drug interactions; however, all interactions were reported to be able to identify those of high interest.

4.2.5 Machine Learning Prioritisation Methods Comparison

The ML method developed was compared with other methods applying ML for gene prioritisation (OpenTargets Genetics L2G[137], Mantis-ml[33], ToppGene [138], and GPrior[19]). OpenTargets Genetics L2G scores were already predicted for the Evangelou et al. (2018) GWAS and available to download from the OpenTargets web-interface[137] - with L2G scores being provided for the three BP traits (SBP, DBP and PP) individually. GPrior, Mantis-ml and ToppGene required an input of positive genes, being given the 51 genes – most likely labelled BP training genes (scored at 1 for regression analysis) as positive examples[19, 33, 138]. They then required different parameters run. ToppGene only required the gene list to be prioritised as input, all other parameters were set to their default training parameters[138]. Mantis-ml required the phenotype term of interest input alongside any exclusion terms[33] – for this we input 'blood pressure' and 'hypertension' whilst excluding 'pulmonary hypertension'. GPrior is the only method that allows the user to input their own features[19], and so for

this method we input the features also used by our XGB model corresponding with our gene list to be prioritised.

XGB prioritisation was also compared with the commonly used gene-based test MAGMA (multi-marker analysis of genomic annotation)[139]. Gene-based tests aggregate associations across a gene to calculate their statistical significance, with aggregation enabling improved statistical power and few tests allowing for a lower significance threshold when correcting for multiple testing[140]. MAGMA was ran via the FUMA web-interface (https://fuma.ctglab.nl/)[141], requiring the GWAS summary statistics from Evangelou et al. (2018) as input with the UK Biobank selected as the reference panel.

4.3 Results

4.3.1 Data Pre-processing

I developed a ML framework for prioritising BP-associated genes post-GWAS (Figure 4.1), in which the model aims to interpret biological knowledge of genes in regions using a range of data types such as genetic (e.g., gene expression across tissues in GTEx), epigenetic (e.g., methylation and DNase sites), and phenotypic (e.g., Exomiser

scores that are calculated using clinical phenotype terms for BP). I applied this framework to the UK Biobank BP-GWAS performed by Evangelou et al. (2018), in which over 7 million SNPs were analysed in over 750k individuals. Genes were curated and annotated to act as ML training data (n=293), with each gene receiving a score for regression analysis. Eight features (HIPred, pLI, Exomiser mouse scores, IPA BP annotation, SDI, and gene expression from the liver, pituitary, and EBV-transformed lymphocytes) were selected and used as model input (Figure 4.2), with benchmarking fourteen models on repeated nested cross-validation. From the selected features only one feature (pituitary gene expression) was in a correlating pair $r^2>0.9$, with the other correlating feature in the pair being removed (Appendix A Table 8). After model benchmarking, the top-performing model (XGB) was fitted to the training data and the 1,804 *BP-genes* - that were taken from the GWAS[47] and were not included in the training data - were then prioritised by that model for further analysis.

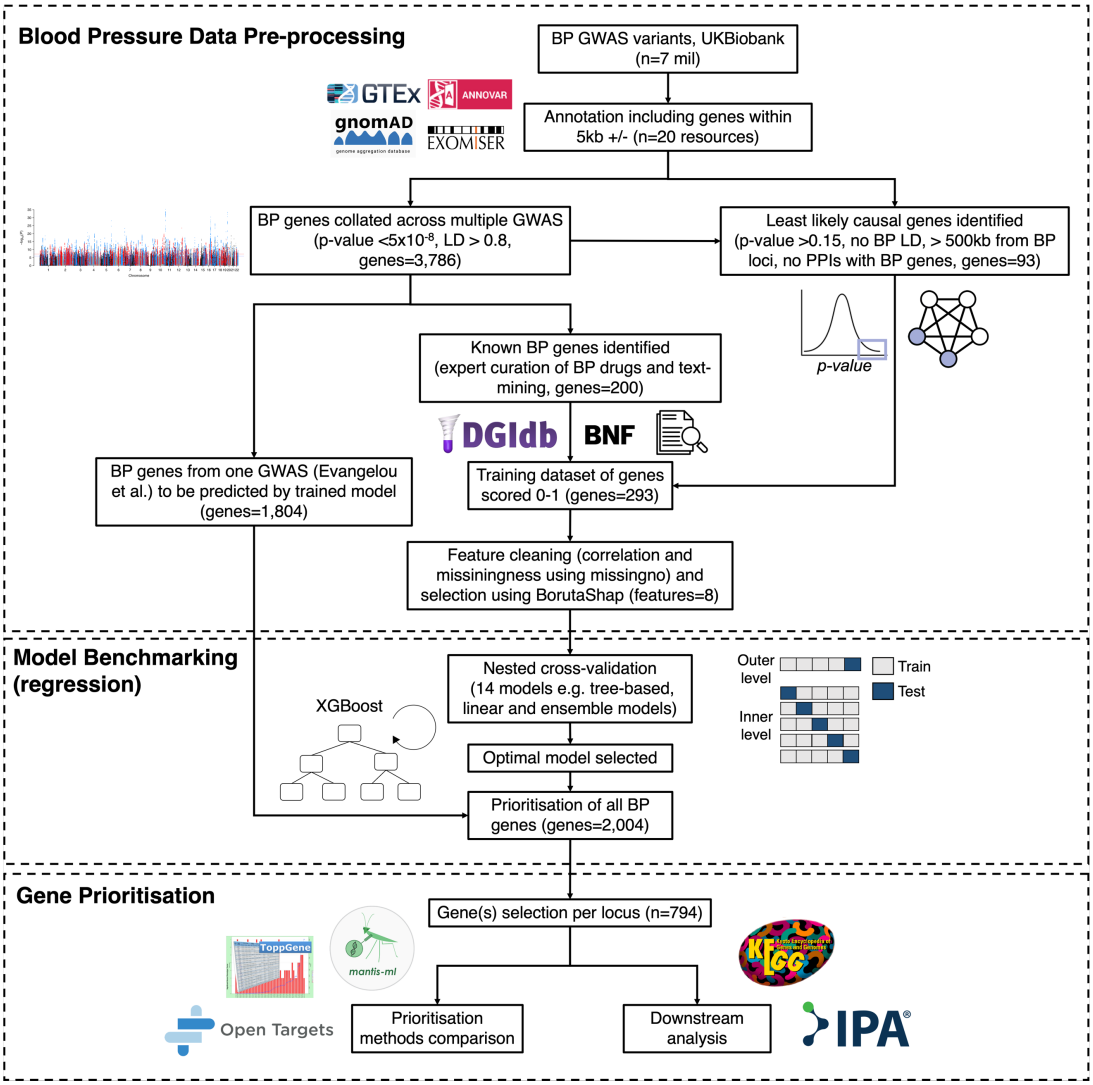


Figure 4.1. Overview of the Gene Prioritisation Framework. Blood pressure genome-wide association study (BP-GWAS) variants tested or included in the analysis by Evangelou et al. (2018) were annotated to genes and evaluated by benchmarked machine learning. Data pre-processing involved annotating variants to genes from the whole GWAS and collecting gene-level annotations from several databases. The genes were then filtered to identify select subsets of *BP-genes* (with linkage disequilibrium, LD, $r^2 > 0.8$ and a p-value $<5x10^{-8}$) and *non-BP genes* that are genes least likely to affect BP (selected by meeting criteria of: not in LD, p-value >0.15, not within 500kb

+/- loci and no direct or secondary protein-protein interactions with BP genes, with interactions defined by experimentally measured interactions at > 0.15 confidence in STRINGdb, n=93). A subset of the *BP-genes* was categorised in two groups to be used in training data depending on if they met selection criteria: most likely labelled BP genes (*BP-regulator genes* with clinical evidenced, n=51), and probable labelled BP genes (*text-mining* genes with published evidence of BP interactions, n=149). These select groups of *BP-genes* (n=200) and *non-BP genes* (n=93) create the training dataset and are each assigned scores for regression analysis. After model benchmarking the unscored *BP-genes* that did not meet the training dataset criteria are those to be predicted by the top-performing model. The top-performing trained model (extreme gradient boosting) was then used for gene prioritisation, with the genes and their corresponding scores being assessed within their loci to select the best gene(s) per locus. The prioritised genes underwent downstream analyses and were compared with other prioritisation methods.

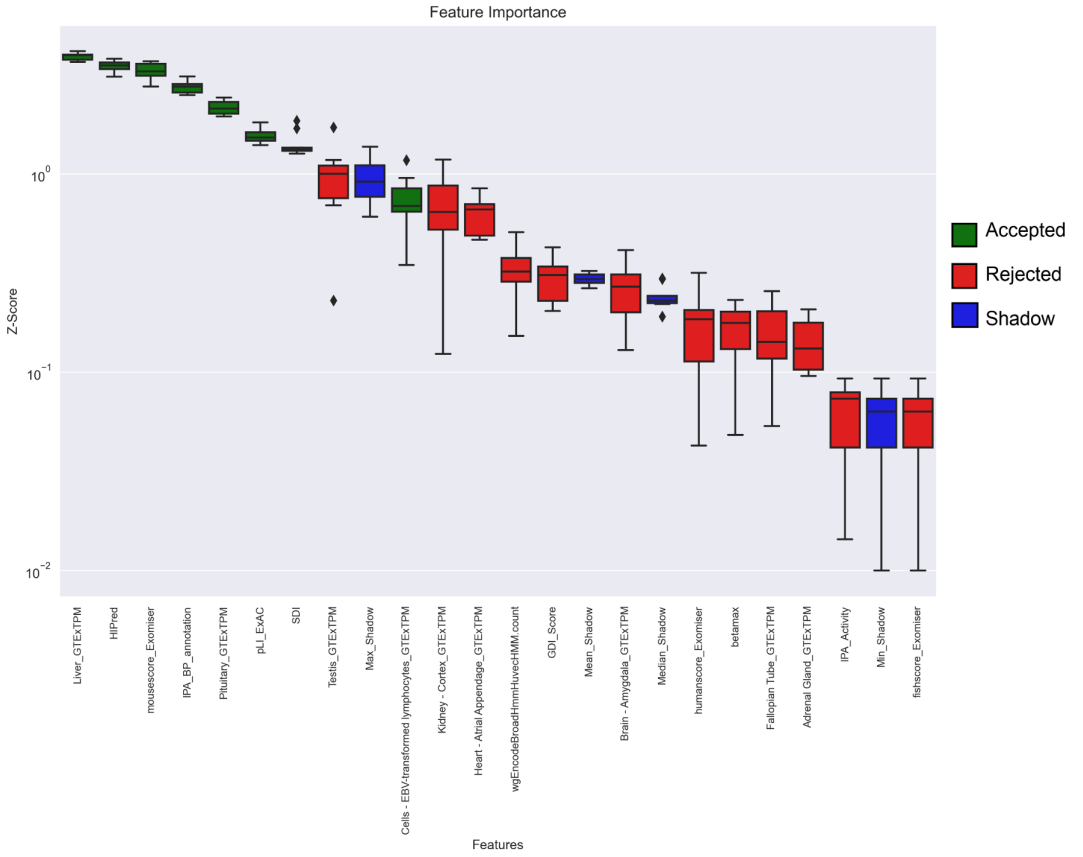


Figure 4.2. Overall feature importance for all features by BorutaShap. 102 features pass data cleaning (removing heavily correlating or missing features) to then enter BorutaShap feature selection which applies the boruta algorithm measured by SHAP feature importance. The box plot shows all 20 cleaned features (<25% missingness and <0.9 r^2) and their measured importance by BorutaShap over 200 iterations of the BorutaShap algorithm (using z-scores), ordered from left to right in descending feature importance. Green boxes indicate selected features, red boxes indicate rejected features, and blue boxes indicate shadow features.

4.3.2 Model Benchmarking

From the fourteen models benchmarked, XGB was selected as the top-performing model to use for further analysis. All models were evaluated using repeated 5-fold nested cross-validation to calculate median performances. Assessment of performance was measured by metrics r^2, predicted r^2, mean squared error, root mean square error, explained variance, and median absolute error (Table 4.1). XGB was selected as the top-performing model to use for further analysis. The XGB model had the highest median r^2 (0.744) (Figure 4.3) and predicted r^2 (0.897). The predicted r^2 importantly measures how well the model will generalise to new data, which led to XGB being the selected model for further analysis. XGB was closely followed by other gradient boosting models such as CB, GBM, and LGBM (with 0.721, 0.718, and 0.71 median r^2 respectively).

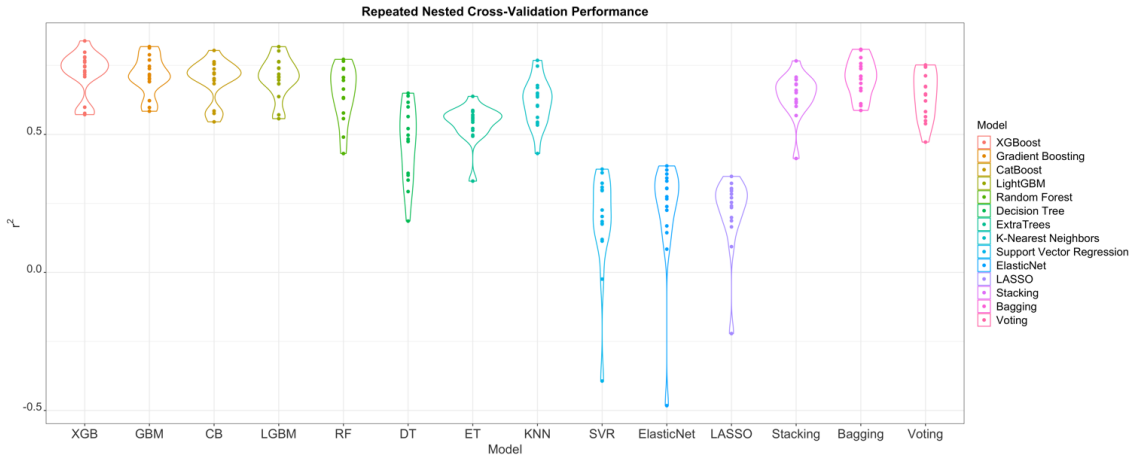


Figure 4.3. Model benchmarking performance on repeated nested cross-validation. Fourteen models were benchmarked: extreme gradient boosting (XGB), gradient boosting (GBM), catboost (CB), LightGBM (LGBM), random forest (RF), decision tree (DT), Extratrees (ET), K-nearest neighbours (KNN), support vector regressor (SVR), two linear models using regularisation of elastic net and LASSO, respectively, and three meta-ensemble methods – stacking, bagging, and voting

models. The model performance was assessed on 5-fold nested cross-validation repeated three times.

Model	Median r^2	predicted r^2	Mean Square Error	RMSE	Explained Variance	Mean Absolute Error
XGB	0.744	0.897	0.031	0.176	0.745	0.125
GB	0.712	0.968	0.035	0.188	0.714	0.131
CB	0.721	0.886	0.034	0.184	0.738	0.138
LGBM	0.71	0.85	0.033	0.183	0.719	0.138
RF	0.664	0.752	0.04	0.2	0.667	0.145
DT	0.483	0.582	0.057	0.241	0.498	0.146
ET	0.558	0.589	0.054	0.231	0.564	0.185
KNN	0.636	0.663	0.044	0.21	0.638	0.133
SVR	0.203	0.233	0.088	0.297	0.294	0.248
LASSO	0.254	0.19	0.087	0.296	0.258	0.26
ElasticNet	0.274	0.2	0.083	0.288	0.303	0.254
Stacking	0.659	0.83	0.042	0.205	0.666	0.127
Bagging	0.71	0.849	0.035	0.188	0.722	0.137
Voting	0.646	0.784	0.04	0.2	0.664	0.155

Table 4.1. Model benchmarking performance. Median performance comparison on nested 5-fold cross-validation across several metrics - r^2, predicted r^2, mean squared error, root mean square error (RMSE), explained variance, and median absolute error. Only predicted r^2 measurements were not median calculations but calculated from each model's performance after hyper-parameter tuning. The fourteen models benchmarked were: extreme gradient boosting (XGB), gradient boosting (GBM),

catboost (CB), LightGBM (LGBM), random forest (RF), decision tree (DT), Extratrees (ET), K-nearest neighbours (KNN), support vector regressor (SVR), two linear models using regularisation of elastic net and LASSO, respectively, and three meta-ensemble methods – stacking, bagging, and voting models.

SHAP values showed Liver GTEx expression, haploinsufficiency scores (HIPred), and mouse Exomiser scores were the most important features for the XGB model (Figure 4.4a). All the 51 most likely BP training genes (scored at 1) were successfully scored highly by the model (with a minimum score of 0.68 a median score of 0.88 and all genes scored above SHAP's expected baseline score of 0.59) (Figure 4.4b). On investigating how feature-feature interactions influenced the model several features were shown to interact with one another. For example, the interaction between haploinsufficiency scores (HIPred) and probability of being loss-of-function intolerant (pLI ExAc) had the strongest influencing interaction for XGB, followed by liver expression interacting with mouse Exomiser and HIPred scores (Figure 4.5).

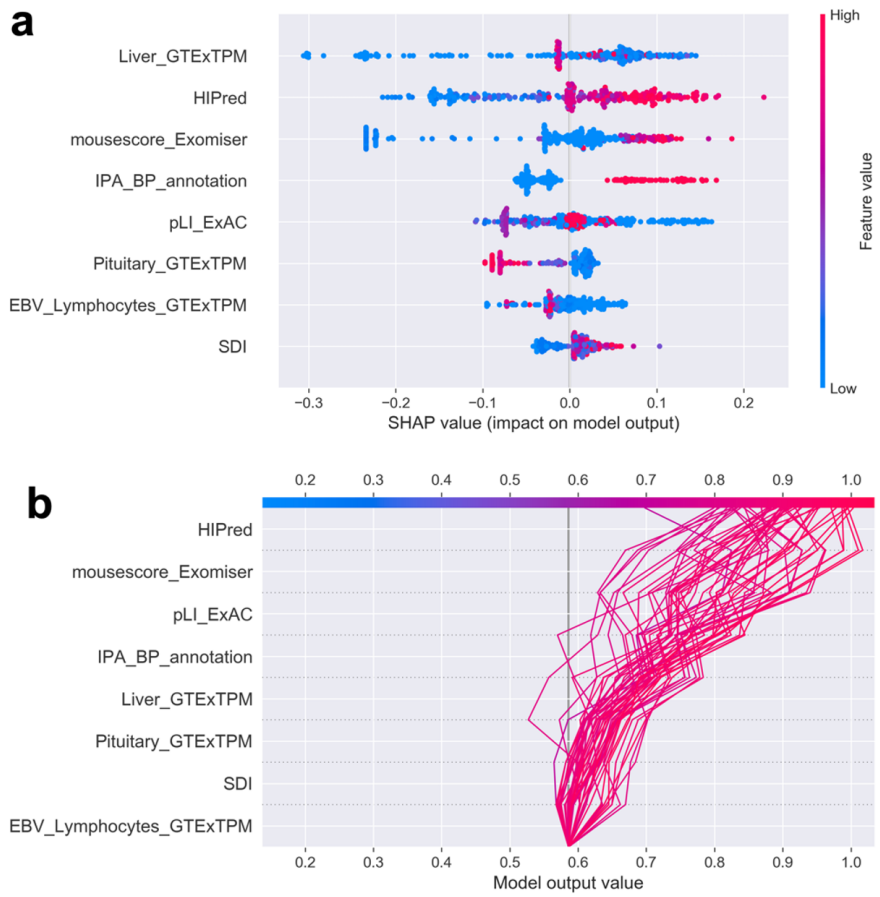


Figure 4.4. Shapley additive explanation of model decision-making. a SHapley Additive exPlanation (SHAP) summary plot of the top-performing model (extreme gradient boosting) predictions of all training genes (n=293) and how they were each influenced by each feature. The SHAP value on x-axis indicates the direction of model influence from that feature for each gene (e.g., a higher SHAP value indicates a higher output model prioritisation). The colour-coding of points (genes) indicates whether their feature value was high (red) or low (blue), and the ordering of features on the y-axis is by descending feature importance overall. **b** SHAP summary plot of the 51 most likely labelled BP genes (scored at 1) predictions, visualising the model's use of features for predicting each of the gene's predicted scores (on the x-axis) – with these

also being plotted against a black vertical line which is the average model score for all training data (0.59).

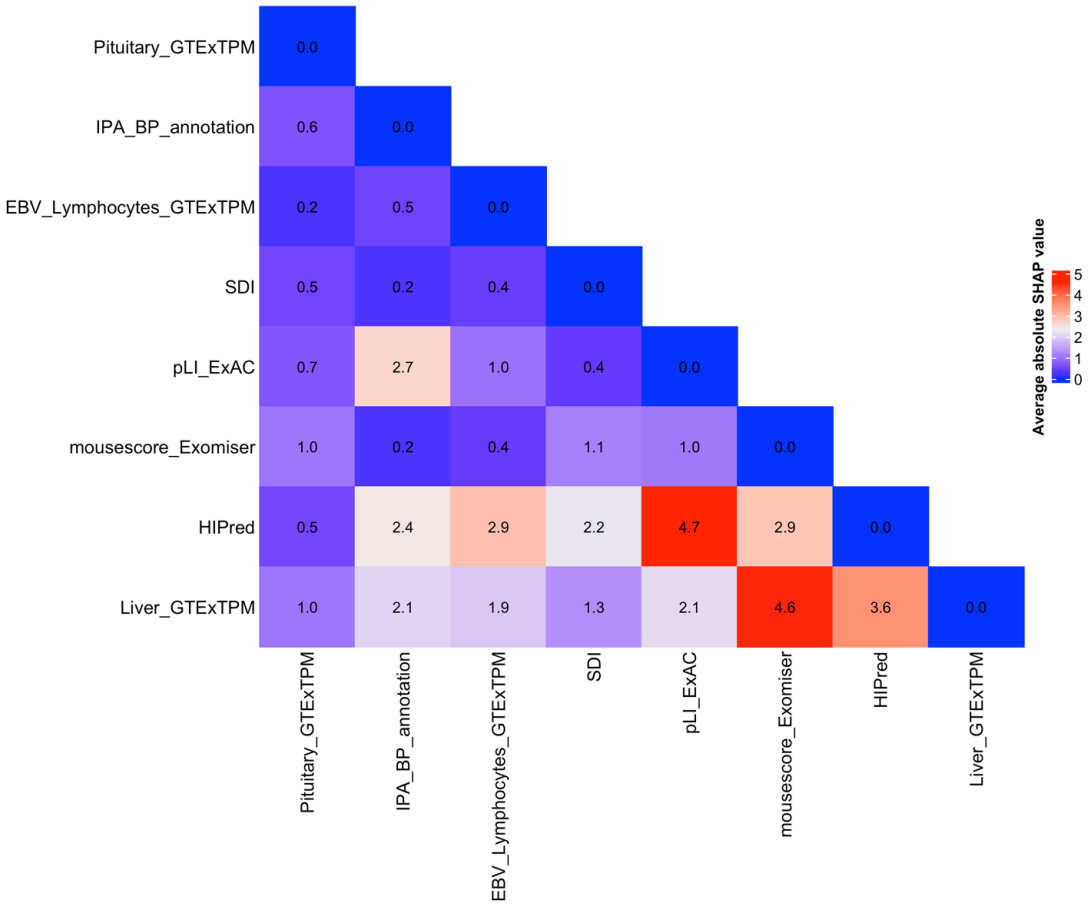


Figure 4.5. Shapley additive explanation of feature interactions. Absolute SHAP value of feature-feature interactions, measuring the impact feature interactions had on model decision-making overall, with a red colour gradient indicating a larger influence on the model and a blue colour gradient indicating less to no model influence.

4.3.3 Gene Prioritisation and Downstream Analyses

Once benchmarked and fitted to the training data, the top-performing model, XGB, was used to prioritise all *BP-genes* (n=2,004 with $r^2 > 0.8$ LD and p-value $< 5x10^{-8}$) (Appendix C Table 3). I investigated all the prioritised genes in two groups: the *highly-*

scored genes and the *selected-genes* (Appendix C Table 4). I investigated these gene groups by assessing intolerance metrics which were not used by the XGB model due to missingness > 25% or not passing feature selection. The Mann-Whitney U test was used to give an indication as to whether the highly prioritised genes by XGB had a significantly different distribution for these annotations in comparison to other gene groups. The *highly-scored genes* had significantly different values on Mann-Whitney U tests in comparison to genes with an XGB score < 0.8 for gene essentiality (measured by Avana mean), mutational damage (measured by the GDI, gene damage index), genic intolerance (measured by RVIS, residual variation intolerance score) and level of ubiquitous expression across cell types (measured by PanglaoDB) (Appendix C Table 5, Figure 4.6). The most significant difference was the Avana mean, a gene essentiality measure, with a Mann-Whitney U test adjusted p-value of 7.8×10^{-77} when comparing the annotation for the *highly-scored genes* against all other XGB scored genes (Appendix C Table 5). The *highly-scored genes* had negative Avana mean values indicating that more essential genes were highly prioritised (Figure 4.6).

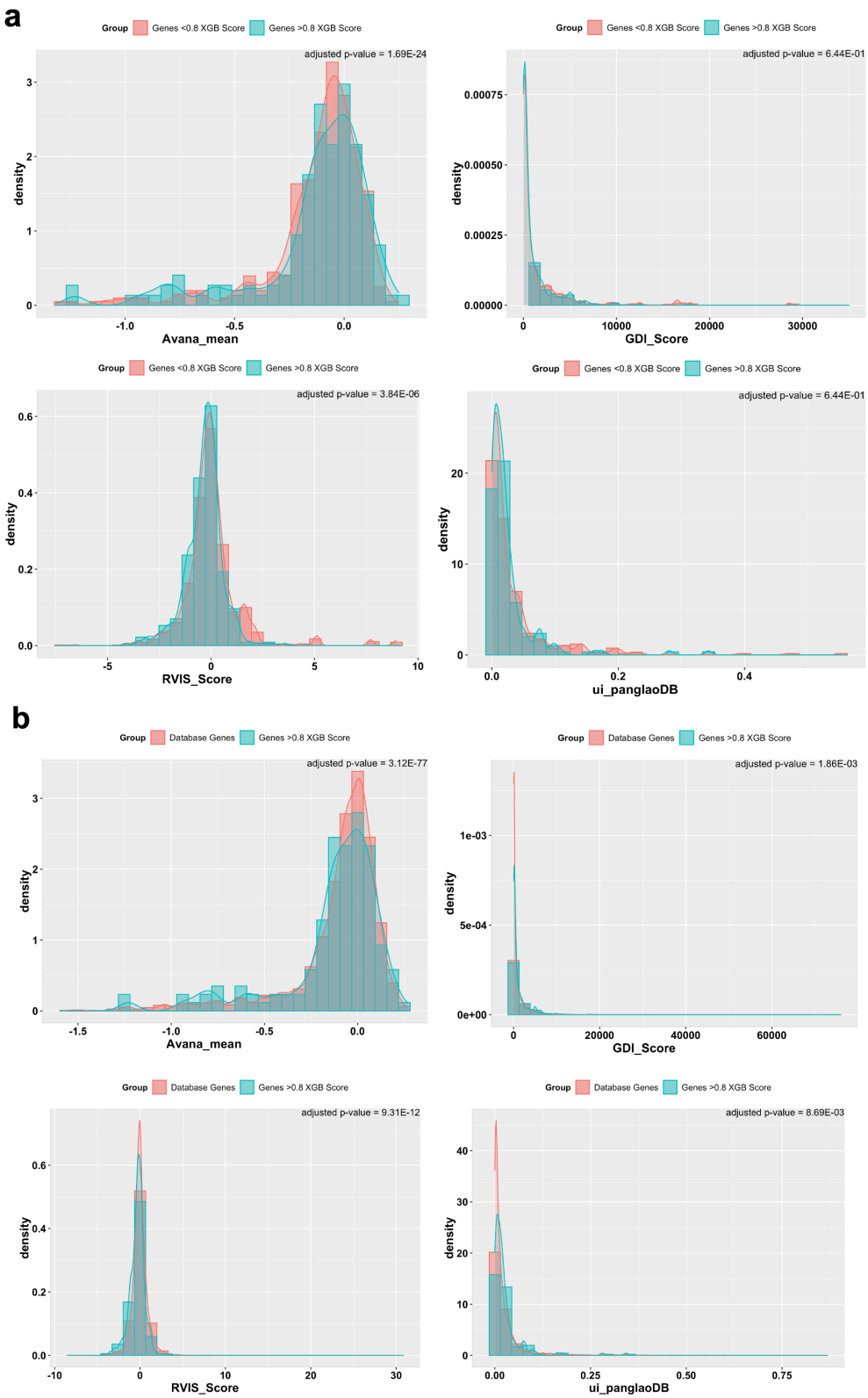


Figure 4.6. Distributions of annotations for genes prioritised > 0.8 versus genes scored < 0.8 (a) and genes > 0.8 versus total database annotations (b). Annotations

not used in machine learning were plotted comparing genesets across several measures: mutation damage (GDI), gene essentiality (Avana mean), genic intolerance (RVIS), and ubiquitous cell-type expression (panglaoDB). Genes scored > 0.8 had their annotations compared against that of genes < 0.8 (**a**), and that of the total genes in database for each annotation (**b**). The Mann-Whitney U test identified significance differences in distributions.

When comparing the *selected-genes,* 329/434 of the *highly-scored genes* by the model passed the selection strategy to be the selected genes per their locus. Exploring Mann Whitney U tests for the selected-genes showed significant differences in the distribution of several measures (Appendix C Table 5, Figure 4.7). The *selected-genes* had the most significant difference on comparing their RVIS scores with that of all other genes scored by XGB (adjusted p-value = 3.41×10^{-7}) (Appendix C Table 5, Figure 4.7). The *selected-genes* had lower RVIS scores than all other scored genes (Figure 4.7), indicating that genes with more intolerance to variation were highly prioritised by XGB.

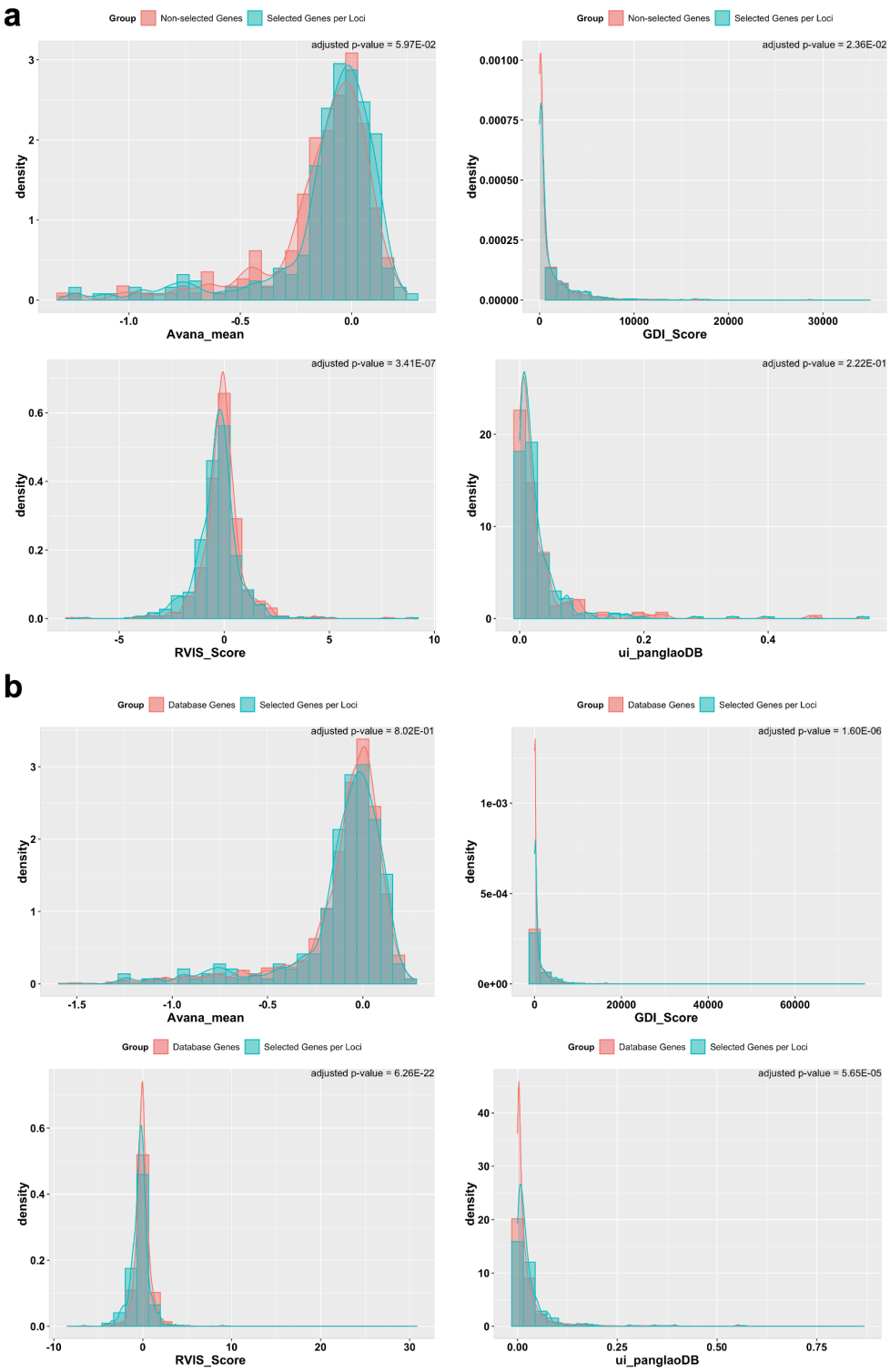


Figure 4.7. Density distributions of annotations for selected genes per locus versus all other scored genes (a) and selected genes per locus versus total database

annotations (b). Annotations not used in machine learning were plotted comparing genesets across several measures: mutation damage (GDI), gene essentiality (avana mean), genic intolerance (RVIS), and ubiquitous cell-type expression (panglaoDB). Genes scored > 0.8 had their annotations compared against that of genes < 0.8 (**a**), and that of the total genes in the database for each annotation (**b**). The Mann-Whitney U test identified significance differences in distributions.

Furthermore, I used the International Mouse Phenotyping Consortium (IMPC) database[142] to explore the gene groups annotated to mouse knockout phenotypes. 216/434 *highly-scored genes* have knockout mouse models with 56 of the genes having phenotypes relating to BP physiology (e.g., cardiovascular and kidney abnormalities, Appendix C Table 6). Enrichment testing for *highly-scored genes* and their overlap with all genes annotated to IMPC phenotypes showed 210 statistically significant phenotypes in total (adjusted p-value < 0.01), with "preweaning lethality, complete penetrance" being the most significantly enriched (adjusted p-value $=1.27 \times 10^{-11}$) followed by "increased heart weight" (adjusted p-value = 7.45×10^{-9}) and "increased circulating cholesterol level" (adjusted p-value = 5.7×10^{-8}) (Figure 4.8, Appendix C Table 7). For the *selected-genes*, 375/794 genes have knockout models, with 96 of those having phenotypes relating to cardiovascular or kidney abnormalities. Enrichment testing for *selected-genes* overlap with IMPC phenotypes showed 278 statistically significant phenotypes in total, with "preweaning lethality, complete penetrance" also being the most significantly enriched phenotype (adjusted p-value = 4.31×10^{-20}) followed by "hyperactivity" (adjusted p-value = 1.15×10^{-12}), and

"abnormal kidney morphology" (adjusted p-value = 1.27×10^{-12}) (Figure 4.8, Appendix C Table 7).

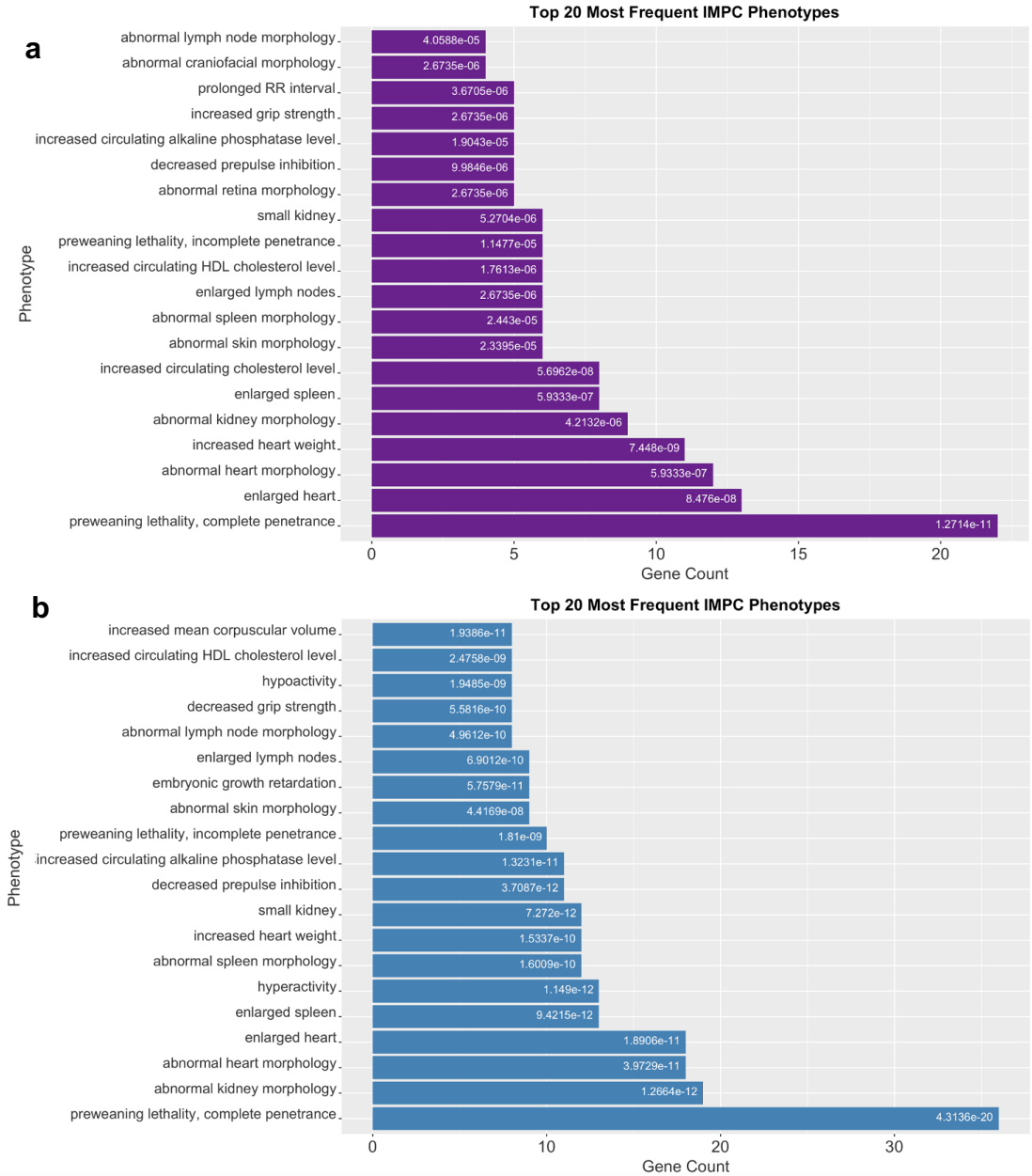


Figure 4.8. Bar plot of the most frequent mouse knockout phenotypes for highly scored genes (a) and for selected genes per locus (b). Each bar indicates the number of genes present for each knockout mouse phenotype for that gene group out of the total number of genes annotated to each phenotype in the International Mouse Phenotyping Consortium (IMPC) database. Each phenotype had hypergeometric

testing for adjusted p-value significance of gene overlap. **a)** Counts of highly scored genes (scored > 0.8) by XGB and their phenotype hits in the IMPC are coloured in purple and **b)** counts of selected genes' phenotype hits were coloured in blue.

Overall *COL15A1* was the top scored predicted gene (0.978) with other highly prioritised predicted genes by XGB including: *SMOC2I* (0.958), *MLIP* (0.956), *SGCD* (0.945), *CHRM2* (0.94), *ELK3* (0.929), *SNTB2* (0.929), *ARHGEF26* (0.921), *PTPN3* (0.92), *NOTCH3* (0.919) (Table 4.2). Several of these genes have had experimental and *in vivo* model research investigating their roles in BP[143] and related cardiovascular conditions such as cardiomyopathy[144], pulmonary arterial hypertension[145], and atherosclerosis[146]. Furthermore, 62 of the predicted genes have interactions, as recorded by DGIdb, with drugs that have potential BP side effects (recorded by SIDER) (Appendix C Table 12). For example, *CHRM2* is listed as interacting with 6 drugs with either potential hypertensive or hypotensive side effects (Appendix C Table 8). From the *selected-genes* that have encoded proteins that interact with any kind of drug, regardless of BP side effects, there are 2,478 total interactions (Figure 4.9, Appendix C Table 9). 743 genes have interactions with at least one drug mechanism, while 548 interact with drugs with a BP indication. 246 *ranked-genes* do not interact with any drug with a BP indication, while 199, 176 and 173 *ranked-genes* respectively interact with 1-2, 3-4 or >5 drugs with BP mechanisms. The pathways that were represented in these four groups of genes were evaluated compared to all *selected-genes* (Fig. 6c). From each of the BP drug mechanisms, 4.5-26.1% of their total interactions recorded in STRINGdb are with *selected-genes* (Table 4.3).

Gene	XGB Score	Potential Druggability	Most Significant Pathway	Locus Gene(s)	Median GWAS p-value	OpenTargets Score
COL15A1	0.978	Druggable genome	Protein digestion and absorption	*COL15A1*	1.13E-16	0.89
SMOC2	0.958	Druggable genome	NA	*SMOC2*	4.45E-07	0.82
MLIP	0.956	NA	NA	*MLIP*	9.17E-11	0.79
SGCD	0.945	NA	Viral myocarditis	*SGCD*	1.15E-08	0.53
CHRM2	*0.94*	Druggable genome	Cholinergic synapse	*CHRM2*	1.37E-08	0.86
ELK3	0.929	Transcription factor	NA	*CDK17, ELK3*	2.72E-09	0.3
SNTB2	0.929	NA	NA	*CHTF8, CIRH1A, SNTB2, TERF2*	4.15E-10	0.28
ARHGEF 26	0.921	NA	Bacterial invasion of epithelial cells	*ARHGEF26, ARHGEF26-AS1, RPL21P42*	1.63E-12	0.88
PTPN3	0.92	Druggable genome phosphatase	NA	*PTPN3*	7.15E-12	0.42
NOTCH3	0.919	Clinically actionable	Notch signaling pathway	*NOTCH3*	6.23E-17	0.74

Table 4.2. Description of the top ten prioritised genes. The top ten scored genes by XGBoost (XGB) and descriptions of: their druggability as annotated by the Drug-

Gene Interaction database, their most significant KEGG pathway, their other locus gene(s), their median GWAS p-value, and their prioritisation scored predicted by OpenTargets.

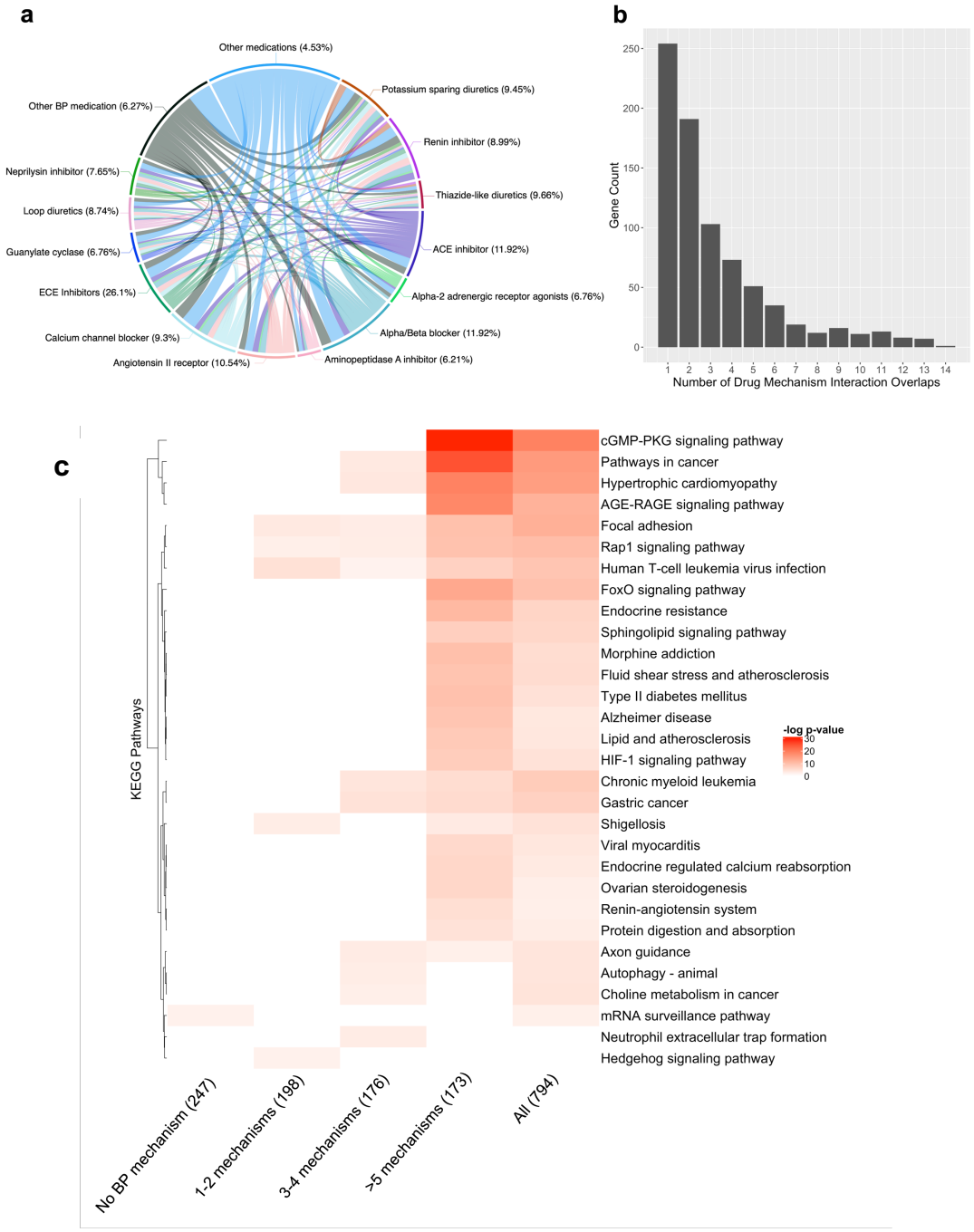


Figure 4.9. Drug mechanism overlaps between selected genes per loci. a Chord diagram of fifteen categories of drug mechanisms annotated as interacting with the

selected genes per loci. Each sector denotes a drug mechanism with the numbering per section representing the number of genes interacting with that drug mechanism that also overlap with another mechanism. Each mechanism is also annotated with the percentage of interactions between selected-genes out of all total interactions per mechanism. **b** barplot of the number of drug mechanism overlaps counted across all genes, e.g., showing only one gene had interactions with 14/15 drug mechanisms. 203 genes from the 794 selected genes at their loci do not have an overlap with multiple mechanisms (interacting with only one drug mechanism) and 51 genes had no interactions with any drug mechanisms. **c** heatmap of enriched pathways for *selected-genes* grouped by number of drug interactions by mechanism (genes with no blood pressure drug mechanism interaction, genes interacting with 1-2 mechanisms, genes interacting with 3-4 mechanisms, genes interacting with more than 5 mechanisms and all *selected-genes*).

Mechanism	Number loci interactions	Total Drug Mechanism interactions	Total (%)
Other medications	733	16175	4.53
Other BP medications	313	4989	6.27
Alpha/Beta Blockers	194	1627	11.92
ACE Inhibitors	136	1544	8.81
Renin inhibitors	135	1502	8.99
Potassium sparing diuretics	130	1376	9.45
Endothelin/ECE Inhibitors	130	498	26.10
Angiotensin II receptor	105	996	10.54
Guanylate cyclase	88	1302	6.76
Alpha-2 adrenergic receptor agonists	79	1197	6.6
Neprilysin inhibitors	75	980	7.65
Loop Diuretics	59	675	8.74
Aminopeptidase A	54	870	6.21
No mechanism linked	51	NA	NA
Thiazide-like diuretics	49	507	9.66
Total	2529		

Table 4.3. Total number of loci with encoded protein drug interactions across drug mechanisms. For each drug mechanism collected, the number of selected genes per loci with interacting encoded proteins were counted (giving the number of loci

interactions per each drug mechanism). This was compared against the total number of protein-protein interactions for each drug mechanism, and the total percentage of loci interacting within those total protein-protein interactions was calculated per mechanism.

4.3.4 Gene Expression

I next explored gene expression across all 54 tissues available in GTEx[147] for both the *highly-scored genes* and *selected-genes* (Figure 4.10 and 4.11). For the *highly-scored genes*, k-means clustering found one cluster of genes (*DUSP1, S100A4, RRAS, CD151, LTBP4, and CAV1*) that had high expression in arterial tissues (aorta, coronary and tibial arteries), adipose tissue, and the bladder (Figure 4.10). The *selected-genes* identified a group of genes (*HLA-B, MYH11, ACTA2, IGFBP7,* and *TPM2*) with a similar pattern of high gene expression across the same arterial tissues alongside colon and reproductive tissues (Figure 4.11). On analysis in STRINGdb, the GTEx cluster for the *selected-genes* (*HLA-B, MYH11, ACTA2, IGFBP7,* and *TPM2*) showed *MYH11*, *ACTA2* and *TPM2* have PPIs with one another, and they also act in smooth muscle contraction and were annotated to arterial diseases - aligning with their clustered gene expression in cardiovascular tissues by GTEx.

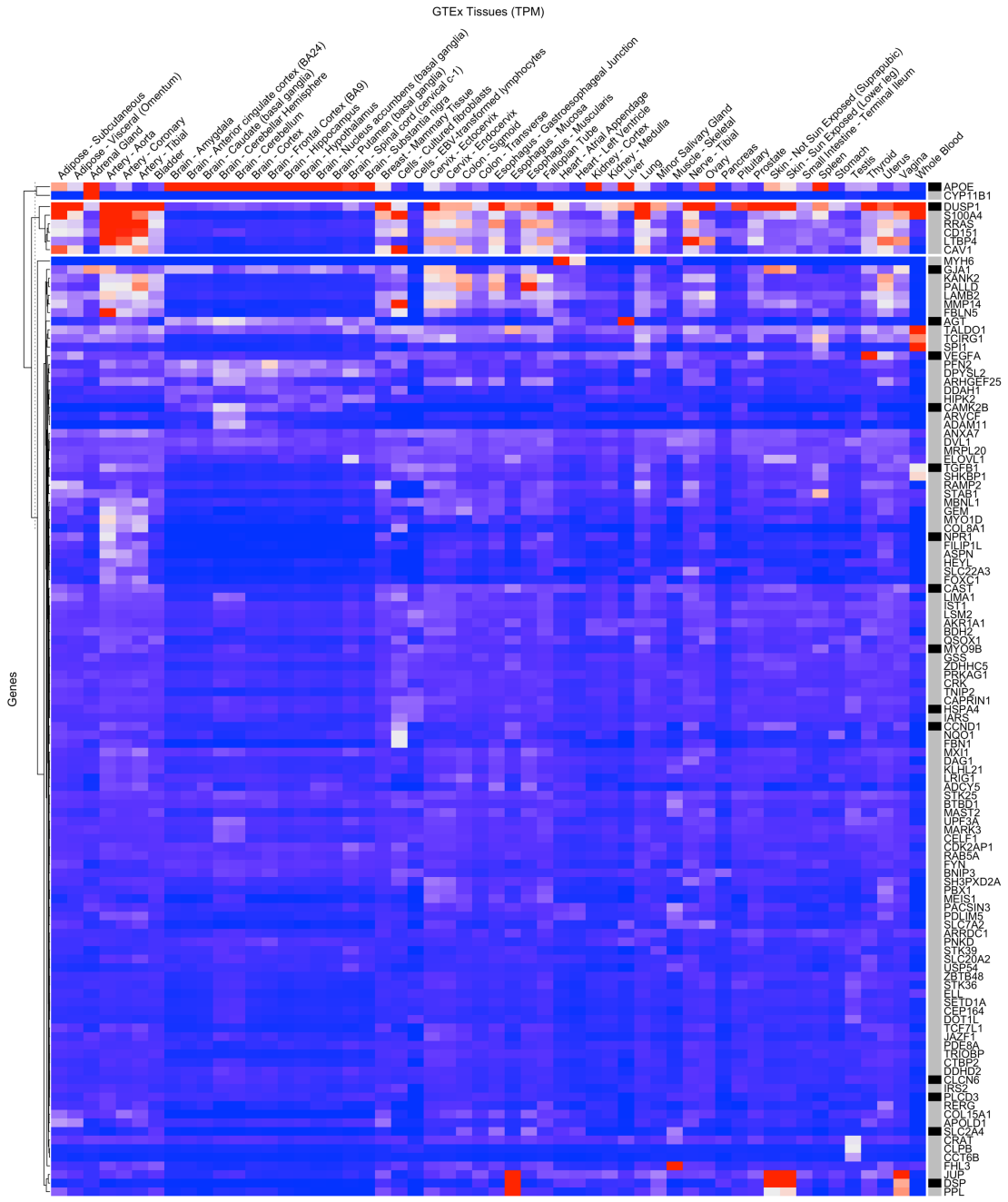


Figure 4.10. Heatmap of gene expression for the most highly expressed genes scored > 0.8 for all 54 tissues in GTEx. Gene expression is measured in median transcripts per million (TPM) with the 51 known interacting BP used in XGB training also marked in black and all other genes marked by grey.

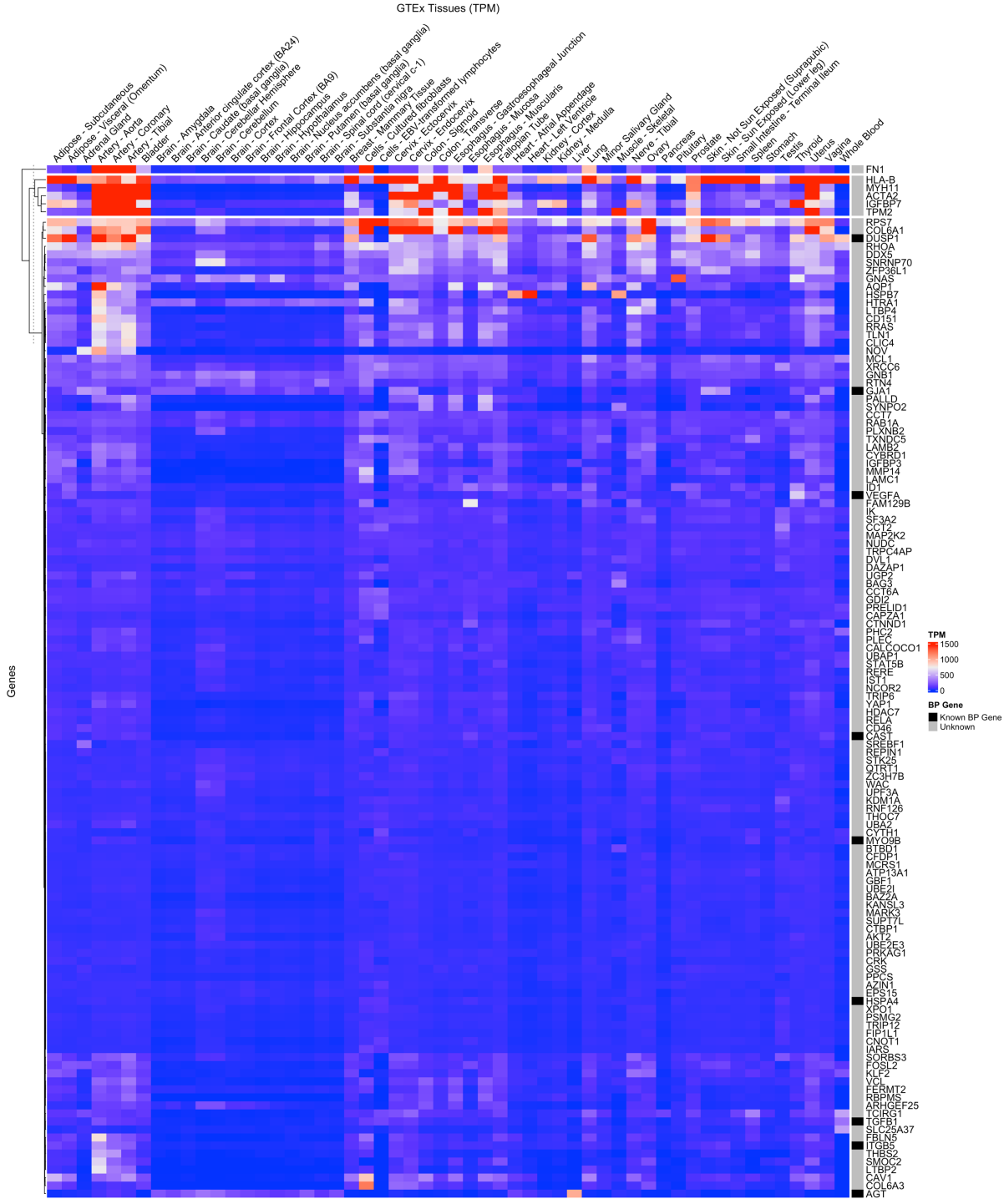


Figure 4.11. Heatmap of gene expression for the most highly expressed genes selected per loci for 54 tissues in GTEx. Gene expression is measured in median transcripts per million (TPM) with the 51 BP-regulator genes used in XGB training also marked in black and all other genes marked by grey.

4.3.5 Gene Enrichment Analysis

Pathway analysis showed the *highly-scored genes* with the most significant pathways being cardiovascular-related (e.g., hypertrophic and dilated cardiomyopathies followed by cGMP-PKG signalling) (Figure 4.12a, Appendix C Table 10). Pathway interactions and the overlaid druggability of genes showed interactive genes in BP pathways that are also potential drug targets (Figure 4.12b). For example, *SLC8A1* (scored 0.86 by XGB) interacts in the pathways of cardiomyopathies and cGMP-PKG signalling and is also a druggable target (Figure 4.12b).

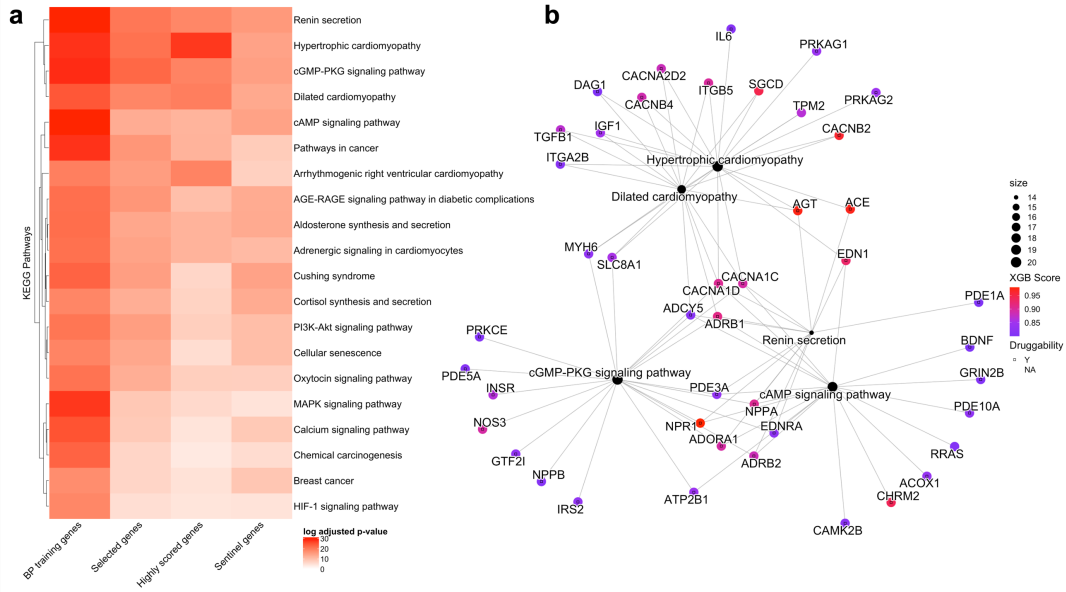


Figure 4.12. **Gene enrichment analysis of prioritised genes. a** Heatmap of the top 20 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and **b** Gene interaction network of the top five most significantly enriched KEGG pathways for the *highly-scored genes.* The heatmap **a** shows more significant values are indicated by darker shades of red and both compare four gene groups, composed of: *highly-scored genes* (genes with a > 0.8 XGB score), BP training genes (most likely and probable labelled BP training genes), *selected-genes* (genes elected at their locus), and

sentinel genes (identified by the Evangelou et al. (2018) genome-wide association study). The gene pathway interactions in **b** have gene nodes colour-coded with higher prioritised genes by extreme gradient boosting in dark red and lower scored genes scored in light red. Pathway node size indicates enrichment log p-value for each pathway node. Square symbols represent whether the gene had druggability recorded in the Drug Gene Interaction Database.

4.3.6 Machine Learning Methods Comparison

All prioritisation methods compared showed positive correlations for prioritising BP genes used in the training data (the 51 *BP-regulator* and 149 *text-mining* BP training genes) and for all predicted genes (the 1,804 genes prioritised by the trained model) (Table 4.4, Figure 4.13). The highest positive correlations were for the predicted genes (when against GPrior and ToppGene with 0.63 and 0.62 correlations respectively) followed by 0.19 correlation for the most likely BP gene group with both OpenTargets and Mantis-ml. Whilst all comparisons had a positive correlation, only the correlation with predicted genes showed statistical significance, for example, with OpenTargets having a p-value of 2.8×10^{-10} for its 0.18 correlation with XGB predictions.

	OpenTargets	GPrior	Mantis-ml	ToppGene
BP-regulator genes (scored 1.0 on training)	0.19	NA	0.19	NA
Text-mining genes (scored 0.75 on training)	0.14	0.12	0.12	0.16
Predicted genes	0.18	0.63	0.3	0.62

Table 4.4. Comparison of machine learning gene prioritisation methods. Table comparing the prioritisation of training genes that were scored as *BP-regulator genes* (scored at 1.0, n=51) or *text-mining genes* (scored at 0.75, n=149) and predicted genes (n=1,804) by several methods in comparison to extreme gradient boosting, measured by their correlation (R) for their predicted gene scores.


a
R = 0.19, p = 0.18
R = 0.14, p = 0.088
R = 0.18, p = 2.8e-10
OpenTargets_L2G_Score
XGB_Score
0.00
0.25
0.50
0.75
1.00
Training_label
BP-regulator gene
Text-mining BP gene
Predicted gene


b
R = 0.19, p = 0.18
R = 0.12, p = 0.13
R = 0.3, p < 2.2e-16
Mantis_probability
XGB_Score
0.00
0.25
0.50
0.75
1.00
Training_label
BP-regulator gene
Text-mining BP gene
Predicted gene

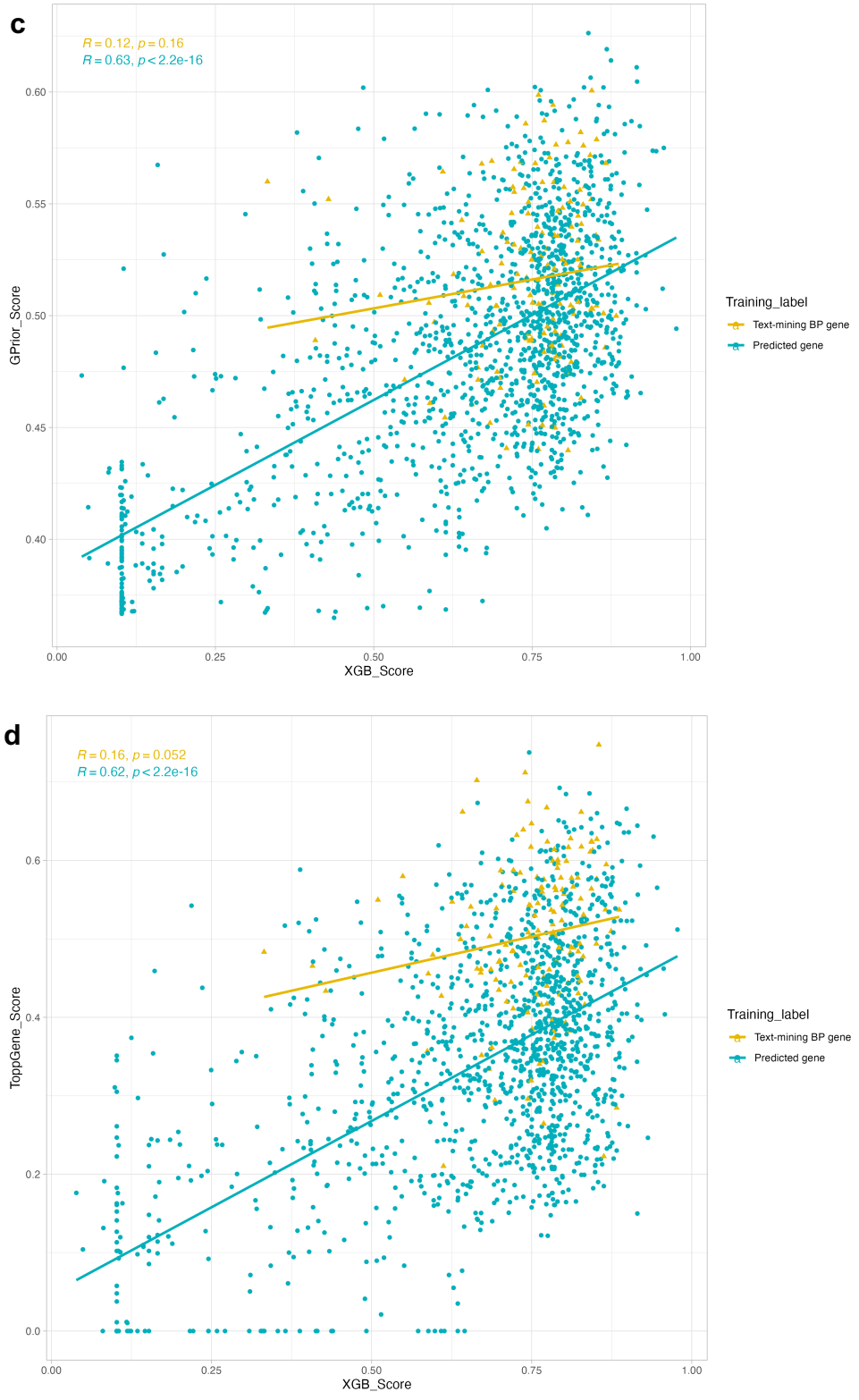


Figure 4.13. Prioritisation method comparison on predicting blood pressure genes. The extreme gradient boosting (XGB) method was plotted against OpenTargets

Genetics locus 2 gene (L2G) score (**a**), Mantis-ml (**b**), GPrior (**c**), and ToppGene (**d**) prioritisation methods comparing gene prediction. The comparison of scores between XGB and each method shows how the methods score the BP training genes (the 51 BP-regulator genes and 149 text-mining BP genes scored at 1 and 0.75 respectively) and all other BP genes predicted by the trained model (predicted genes, n=1,804). For each plot, the score the XGB model gave to each gene is plotted along the x-axis in comparison to the other method's prediction for each gene plotted on the y-axis. Correlation (R) between the two methods for each prediction is calculated alongside the p-value significance of the R. The 51 BP-regulator genes are coloured in red, the text-mining BP genes are coloured in yellow, and the 1,804 predicted genes are coloured in blue.

4.4 Discussion

The regression approach applied here is, to the best of our knowledge, the first of its kind for post-GWAS prioritisation. Regression removes the need for models to categorise genes, as is the case in multiclass classification. The curation of *non-BP-genes* (labelled as least likely genes in previous chapters and scored at 0.1) also develops model knowledge, with predictions made by regression allowing for the implication of uncertainty in the least likely gene predictions. I.e., a score of 0 reflects definitively negative cases whilst a score of 0.1 indicates a degree of uncertainty that can be assessed by the model. However, it should be noted that the underlying class imbalance that was prevalent in the ML performances on multiclassification is still underlying in the same training data converted to regression analysis. Furthermore,

the setting of gene scores for regression (1, 0.75 and 0.1) provides a beneficial ordinal understanding of genes for a model, but the exact scoring was based on ML tests of multiple scoring intervals and may provide biased predictions if a model is overfitting. These adaptations of the framework are difficult to optimise without a further external validating dataset that could confirm the benefits of regression over multiclass classification and confirm or rule out any overfitting. The benchmarked models explored also compared a range of commonly used methods to reach an optimised approach. The top-performing model being XGB, and its performance being similar to other gradient boosting models (CB, LGBM, and GBM) aligns with previous ML prioritisation research post-GWAS[19, 33] and research that found XGB and CB are top-performing models on tabular datasets for supervised learning (when compared against 15 other models, including those benchmarked here)[129].

On exploring the performance of XGB, SHAP interpretation showed *BP-regulator genes* were predicted based on HIPred followed by mouse Exomiser scores and pLI measures as the most important features (Figure 4.3b). The use of these genetic intolerance and phenotypic features shows the model highly prioritises genes that are more haploinsufficient, have more observed mouse phenotypes relating to BP via Exomiser, and have a higher probability of being loss-of-function intolerant genes. This model interpretation is also validated by our Mann-Whitney U test analysis finding other annotations not used in ML showed that the highly prioritised genes were more likely to have intolerant variations and have significantly enriched cardiovascular and renal mouse phenotypes (Appendix C Tables 6 & 7) – suggesting the model develops a BP-tailored understanding for more informed decision-making.

Furthermore, the feature-feature interactions between gene expression features in the selected GTEx tissues (e.g., liver and lymphocytes Figure 4.5) have large SHAP values alongside that of the gene intolerance and phenotypic features. These interactions highlight how the model identifies relevant biological relationships and suggests novel directions for downstream analysis to investigate potential BP pathways. For example, both the liver and lymphocytes have published roles in BP both separately[148, 149] and interacting together[150]. However, this research requires further investigation, and the gene prioritisation here presents opportunities for specific genetic BP hypotheses focused on these tissues.

Downstream analysis validated the ML approach, highlighting supporting evidence for the *highest-scored genes* and their potential roles in BP. Notably, *CHRM2, ELK3* and *NOTCH3* have had research focusing on their roles in BP regulation, with downregulation of *CHRM2* worsening hypertension in rats[151], *ELK3* being a repressor of nitric oxide synthase[152], and *NOTCH3* knockout mice having shown vascular abnormalities that alter BP[120]. All of the top ten genes have had either cardiovascular or renal research, e.g., *SMOC2* has been shown to develop kidney fibrosis (and therapeutic silencing of *SMOC2* has been shown to maintain kidney function)[153]. The top prioritised gene, *COL15A1,* encodes the alpha chain of type XV collagen protein and has had research relating to cardiomyopathy[144] and atherosclerosis[146], with it also being druggable in the genome (and having an interacting drug ocriplasmin in DGIdb that is used to break down fibronectin in vitreomacular adhesion).

In comparison, from all 1,804 predicted genes, 62 of them interact with drugs that have either potential hypertensive or hypotensive side effects as recorded by SIDER, and 52/62 genes were also validated by having hypertensive/hypotensive side effects recorded by the British National Formulary (BNF) (Appendix C Table 8). Attention deficit hyperactivity disorder and central nervous stimulant drug methylphenidate had the most gene interactions (with *CORO7, FARP2, SENP3, FXR2, ARHGAP12,* and *ELP5*) followed by the BP-lowering angiotensin-receptor antagonist drug candesartan (*DOT1L, PLEKHJ1*, and *SULTC3*). Nine of the 62 genes (*DOT1L, PTPRD, ATXN2, CBX1, TBX2, RRP1B, ATAD5, PLEKHJ1* and *SULT1C3*) were shown by DGIdb to interact with drugs indicated for hypertension treatment by the BNF. For example, *PTPRD* (scored 0.853 by XGB) interacts with calcium channel blocker verapamil - used to treat hypertension and angina (Appendix C Table 8). On investigating this interaction, Gong et al. (2015) found a *PTPRD* polymorphism (rs4742610) associated with resistant hypertension in those taking verapamil, suggesting this gene may be beneficial in understanding clinical BP response. Overall, each of these examples validates the ML framework and puts forward new avenues for further translational research.

From the top 10 prioritised genes, DGIdb showed *CHRM2* has an agonistic interaction with antidepressive drugs (olanzapine, doxepin and desipramine) - with SIDER and BNF both identifying olanzapine and doxepin as having hypotensive side effects (Appendix C Table 8). Padmanabhan et al. (2021) have also written about *CHRM2* amongst other BP loci that have interactions with antidepressive drugs that have BP side effects, suggesting such drugs could have a multi-purpose use for treating those

with depression and hypertension[155]. *CHRM2* is a muscarinic acetylcholine receptor and its potential link to BP (both via its indicated drug interactions and by research in rats that has shown its downregulation in combination with other genes worsens hypertension[156]) suggests that focusing on such nervous signalling roles possibly underpinning BP may highlight targets with therapeutic potential and novel biological insights that are not in the cardiovascular and renal sites of action for BP.

Another notable gene-drug interaction in DGIdb was *INSR* (scored 0.863 by XGB) interacting with the migraine drug topiramate which has a hypotensive side effect (recorded in both SIDER and BNF) and has had clinical research finding it lowered BP in obese patients with essential hypertension[157]. The interaction between *INSR* and topiramate has also been focused on in a pharmacogenetic study, finding genetic variants in *INSR* impacted the effectiveness of topiramate to treat weight loss and that insulin-related genes regulated topiramate response[158]. These results imply an impactful relationship between *INSR* and topiramate, and that this interaction may relate to the drug's potential effects on BP.

The total interactions between each *selected-gene's* encoded proteins and drug mechanisms were annotated, showing the potential novel drug targets within BP GWAS loci awaiting discovery (Figure 4.5, Appendix C Table 9). 246 of *selected-genes* genes interact with 4 to 14 drug mechanisms each, suggesting those genes possibly have more ubiquitous effects, whilst most genes (n=548) had 1-3 interactions that involve both BP-related and non-BP-related drug mechanisms. Three genes had interactions between BP drug mechanisms only (*BRD1, MFSD6,* and *PLEKHG1*),

suggesting these could have more targeted drug effects for further investigation, however, these genes are not annotated with any interacting drugs, as recorded in DGIdb. Examining the *selected-gene* interactions at each BP drug mechanism also presented subsets of genes with translational potential. For example, 9.3% of the total gene-drug interactions for calcium channel blockers recorded in STRINGdb interact with *selected-genes* (Appendix C Table 9), which make up 194 of the 768 loci investigated here, validating their prioritisation, and highlighting these genes for further targeted study.

Furthermore, the gene expression analysis found *MYH11* (scored 0.915 by XGB), *ACTA2* (scored 0.9 by XGB), and *TPM2* (scored 0.865 by XGB) as highly expressed in cardiovascular tissues and have interconnected PPIs. In comparison, studies have found all three of these genes are coexpressed in vascular smooth muscle cells to weaken the aortic wall[159] and to be markers of pulmonary hypertension[160]. I also analysed these clustered genes in DGIdb, finding *MYH11* and *ACTA2* are both clinically actionable drug targets, suggesting these highly prioritised genes and their interaction network also offer targeted directions of investigation with putative drug targeting for BP.

In total, 794 genes were selected per loci as most likely BP genes with 747 of them having a model score >0.5 and 515 of them also being genes associated from sentinel SNPs by Evangelou et al. (2018) (Appendix C Table 4) – indicating that the prioritisation results align with the genes' significance found in their GWAS summary statistics. The most significantly enriched pathways for the highest prioritised genes

are also known BP pathways (e.g. renin secretion[161] and cGMP-PKG signaling[162] – Figure 4.12, Appendix C Table 10), validating the model's prediction of most likely BP genes but also showing the circular pattern recognition underlying ML prediction, as seen in chapter 3. However, novel genes were shown to be highly prioritised in these established pathways, such as *SLC8A1* (Figure 4.12b), which is annotated as a druggable target in DGIdb, and encodes a Na^{+}/Ca^{2+} exchanger that has had multiple studies linking its function to hypertension[163, 164].

When comparing the XGB prioritisation to other ML methods, the approach was validated by positive correlations for all gene predictions. The OpenTargets L2G score overall, followed by Mantis-ml, had the highest positive correlations for all training genes across BP traits (Table 4.4). This may be due to both methods using XGB (although differing in using classification or regression) and both being optimised to the data of individual GWAS, supporting bespoke data collection. Uniquely, the XGB method developed in this chapter incorporates phenotypic data specific for BP and positive training examples that were curated specifically from the BP GWAS depending on whether they had an interacting BP drug. Meanwhile, OpenTargets L2G uses co-localisation and fine-mapping data amongst their features, suggesting that highly prioritised genes by both methods have a strong evidence base across different resources indicating their link to BP. For example, there are 286 genes both methods score >0.7 and 76 genes both methods score >0.8 (Appendix C Table 11), creating tailored gene lists for functional follow-up.

Furthermore, the XGB's prioritisation was also compared against gene prioritisation using MAGMA[139] (Appendix C Table 12). This comparison showed that the MAGMA p-values for each BP trait had no strong positive or negative correlation with XGB prioritisation or the median GWAS p-values (Appendix C Table 12), with no absolute correlation value greater than 0.1. However, on investigating the log-transformed p-values, and thereby comparing the prioritisation methods on the same scale, each BP trait had positive correlations between 0.33 – 0.44. These correlations highlight agreement between the methods and suggest that their overlapping highly prioritised genes warrant further investigation. For example, 196 of the 1,804 prioritised BP genes had significant MAGMA and GWAS p-values (less than $5x10^{-8}$ as also used MAGMA's default p-value threshold) for all three BP traits. Furthermore, 103 of those 196 genes are also *selected-genes* and 55/196 are *highly-scored genes*.

As XGB does not use any GWAS summary statistics as input features the lack of overlapping information used by MAGMA and XGB offers the ability to use both methods to have a more selective gene list - utilising concordance across the prioritisation methods to identify genes for further investigation. The overall lack of correlation between MAGMA and XGB prioritisation or GWAS p-values emphasizes the differences in the underlying methodology. For example, MAGMA uses linear regression to estimate the effect size of each genetic variant on the phenotype and then aggregates these effect sizes across all variants within each gene, outputting gene-level statistics[139]. Meanwhile, XGB uses optimised decision trees for non-linear pattern recognition from various datatypes measured across molecular scales. Additionally, XGB was developed to prioritise the 1,804 associated genes (from 47,249 SNPs in LD

$r^2 > 0.8$ and with p-values less than $5x10^{-8}$) while MAGMA is run on the whole GWAS summary statistics (approximately 7 million variants and their 33,847 annotated genes). To create a more direct comparison with equal gene subsets, MAGMA was run on only the 47,249 SNPs in LD $r^2 > 0.8$ and with p-values less than $5x10^{-8}$. This test showed that MAGMA had an extremely high correlation with the median GWAS p-values directly (0.998), alongside a 0.37 correlation when comparing log-transformed p-values, and a 0.025 correlation with the XGB prioritisation. However, MAGMA is designed to be run on whole GWAS summary statistics, making this comparison less reliable, but it does represent that the data size and subset differences between MAGMA and XGB may impact their interpretation of gene significance, making them difficult to directly compare. Moreover, when benchmarking the methods, linear regression methods were the lowest performers - with 0.27 median r^2 and 0.19 predicted r^2 for linear regression using elasticnet and 0.25 median r^2 and 0.19 predicted r^2 for linear regression using LASSO. This performance highlights the non-linear patterns within the data curated for ML and emphasizes that different gene prioritisation methods may have different strengths and weaknesses depending on the specific data.

Overall, the ML prioritisation provides a different interpretation of potential BP genes in comparison to MAGMA, presenting an opportunity to use these methods in parallel to validate highly prioritised genes for functional follow-up. To explore the potential of this ML prioritisation further, a re-application of the framework was also developed for the post-GWAS analysis of blood lipid traits (detailed in full in Appendix D).

While this chapter successfully optimises the ML framework to achieve promising gene prioritisation, limitations are still underlying in the methodology. For example, similar to chapter 3, XGB prioritises genes that are enriched for already known BP pathways, as proven by gene enrichment analysis (Figure 4.5). While this validates the model's decision-making it also indicates less opportunity for identifying novel insights downstream with enough evidence to justify functional research. As more functional data becomes available there may be more insight into important genes and mechanisms that can then act as inputs to increase the current training data in size and quality. Also, the gene per locus selection method was developed to combat situations where a ML model may score genes closely at a locus, without large score margins between genes that clearly identify one most likely causal gene at a locus. With the use of BP gene PPI data for the gene(s) selection, this filters out less likely top genes per loci via a known BP biology filter. However, this selection method is also limiting, with PPI information relating to known BP genes abetting circular selection for genes with more PPI evidence. Additionally, given that there are few genes with interacting BP drugs, and with the model using these genes in its training data, it was not possible to create an external validation set for further model testing. This testing will be required as new *BP-regulator genes* are discovered in order to assess any inaccurate modelling assumptions[165]. Curation of the *non-BP-genes* that are least likely training genes with any certainty is also a challenge across diseases, despite stringent criteria for selecting our *non-BP-genes*, there is a risk of false negative examples being used in training. Also, for our downstream analysis of prioritised genes, it should be noted that the mouse Exomiser Scores are derived partly using IMPC data, indicating circularity in the mouse phenotype analysis.

Overall, this research puts forward a novel method to navigate and interpret most likely causal signals within the thousands of BP-associated genes being identified by GWAS. It identifies new genes of interest from the GWAS whilst also validating previous downstream analysis. The method provides an opportunity as a complementary tool to support fine-mapping and functional investigation, ultimately leading to increased evidence available for BP GWAS and an enhanced understanding of causality within loci. Furthermore, the performance of this ML framework suggests that it could also have success when re-applied to other phenotypes, especially other traits that have undergone large GWAS research and so present ample opportunity for high-quality training gene curation. Re-applying the methodology developed here to a trait with similar quality training data would also validate the performance seen in this chapter and suggest the framework has potential as an automatable application across phenotypes.

5 Discussion

5.1 Key Findings

In comparison to other traits, genetic insight into blood pressure has reached an enviable level of depth and complexity. This has been enabled by some of the largest GWAS studies performed to date, highlighting over 901 loci in over 1 million individuals[2]. Yet this wealth of information presents a huge bottleneck that genetic analysis is not well placed to resolve. The machine learning approach optimised in this

work not only addresses the current bottleneck of genetic BP locus to gene prioritisation but capitalises on it and presents opportunities for re-application to other phenotypes with amassing GWAS data. The size and quality of BP genetic association studies serve as an opportune source for the high standard of training data that is needed in robust machine learning. Building on this potential for advancing post-BP GWAS analysis curated training data tailored to BP was used to benchmark ML prioritisation methods, finalising a tuned regression analysis approach. The model successfully prioritised candidate genes with a high likelihood of causality in 768 BP loci, emphasizing key processes and pathways in BP biology and bringing new insight into the genetic basis of antihypertensive drug action. The ML framework was then re-applied to prioritise genes from blood lipid trait GWAS data, and separately a variant-level prioritisation framework was developed, showing the promise of ML to discover novel insights across post-GWAS analysis with varying aims.

5.1.1 Exploratory Data Analysis Summary

In this thesis I employed in-depth exploratory data analysis that considered genetic characteristics (i.e., gene length and genomic distance of training genes), going beyond the common explored factors such as feature distributions, correlation, and missingness. Whalen et al. (2022) reviewed how ML is applied to genomic problems and the common pitfalls, focusing on the importance of checking genomic distance due to potential annotations being correlated for neighbouring genes. They also note such pitfalls are common and need thorough inspection[166], which I have aimed to employ in this thesis. The exploratory genomic data analysis implemented here provides a template for genomic data pre-processing before ML application post-

GWAS, as seen with the gene length and genomic distance assessment developed in chapter 2 and re-applied in Appendix D for BP and blood lipid trait GWAS respectively. Furthermore, the results in chapters 2 and Appendix D indicate that in both cases the training genes have no outstanding risk of bias based on genomic position and gene length (due to the features impacted by gene length being removed and gene length itself not being included as a feature). However, as genetic data expands to better encapsulate gene relationships (in not only gene distance and size but also possible interactions with one another) all such genetic relationships need to be understood in training data to ascertain how well the model can generalise its learning to the rest of the genome that it may be applied to.

Chapter 2 also showed the curation of features across molecular scales and datatypes with both continuous and categorical variables, making for a complex multi-omic dataset. The predominant features collected were from GTEx with all 53 tissues having an individual feature for their TPM expression alongside gene functionality and phenotypic measures. This curation followed similar feature collections by other genetic prioritisation studies. For example, GPrior also uses median GTEx TPM expression for all 53 available tissues and variant-level pathogenic scores in their applications to their case studies (IBD, educational attainment, coronary artery disease, and schizophrenia)[19]. In their prioritisation of IBD genes, for example, from the GTEx tissues, the most important features were from the colon, oesophagus, salivary gland, skin, kidney, and whole blood – with the researchers highlighting the importance of these features and their specificity to IBD[19]. In comparison, the BP gene prioritisation in chapter 4 identified gene expression from the pituitary gland, which

is a site of hormone signalling for BP, as well as liver and EBV-transformed lymphocytes. These tissues highlight how the model identified relevant biological relationships but also suggests directions to investigate potential BP pathways. As seen in the example between the liver and lymphocytes, which both separately have published roles in BP[167, 168], and have a recent study identifying a relationship between the two to regulate BP[150]. Specifically, the liver has been shown to relate to BP via liver-derived insulin-like growth factor 1[167], lymphocytes relate to BP by neural signalling[168], and one study has found both tissues interact as the "*constitutive sulfhydration of liver kinase B1 by cystathionine γ lyase-derived H2S activates its target kinase, AMP-activated protein kinase, and promotes Treg differentiation and proliferation, which attenuates the vascular and renal immune-inflammation, thereby preventing hypertension.*" [150] The ML selection of these tissues' features in parallel with this research supports such studies focusing on the importance of the tissues that produce signalling molecules to regulate BP - with the XGB algorithm also prioritising pituitary gene expression and not the immediate sites of action for BP (i.e., cardiovascular and renal tissues).

The valuing of gene expression data by feature selection and XGB (as interpreted via SHAP) in this thesis also validates the use of gene expression across all GTEx tissues as ML features. Using all 53 tissues as input, as opposed to singling out only tissues believed relevant to the phenotype, presents an opportunity to combat and avoid reinforcement of existing knowledge that ML is potentially biased towards. A model can also identify genes known to act in established BP pathways but newly link them to tissues with unknown contributions to BP, thereby developing novel directions of

research. However, as discussed in Appendix D, the influence of sex-biased gene expression, which has been shown in 38% of genes for at least one tissue in GTEx[169], may bias the measure of relevance a gene has in a tissue – showing that follow-up research should be approached with caution to validate the gene's true impact.

The feature collection was curated in a bespoke manner to specifically study BP, providing unique features that may not be available for similar studies of other phenotypes. The work also benefitted from a close collaboration with the Exomiser development team, who provided, Exomiser scores for increased BP. This, alongside IPA annotation for genes reported in experimental BP studies, provides disease-specific information for machine learning. The use of such phenotypic features is also recommended by other methods (Mantis-ml[33] and GPrior[19]), and publicly available resources such as Exomiser provide an opportunity to annotate disease-specific information with increased efficiency on re-application of an ML framework. Uniquely, Exomiser scores integrate disease information from human, mouse and fish model databases. In comparison, other methods such as Mantis-ml have similar data queried from several databases (MGI, OMIM and disease-specific databases such as the Chronic Kidney Database)[33], aggregating several disease-specific measures from varying sources. Meanwhile, Exomiser offers a similar diversity in collating data from both human and animal model databases (HPO, IMPC and MGD) in one method requiring only one query for annotation. Furthermore, the benefit of Exomiser's disease-specific annotation was shown as important in both BP and blood lipid ML frameworks developed in this thesis, validating its benefit and use as a ML annotation

for GWAS, in contrast to the intended use of Exomiser for disease variant prioritisation in individual patients.

On the other hand, IPA BP annotation was also an important feature for BP prioritisation, but IPA lipid annotation was not selected for the blood lipid trait application. The IPA data is curated as a binary feature, identifying if a gene had a disease-specific annotation in the database, which may make it less informative than continuous Exomiser scores. Furthermore, as the annotation is provided from a commercial database it may be of more limited wider utility to the research community. It seems likely that other public domain sources of disease annotation, such as those provided in OpenTargets[115] would be suitable replacements for proprietary data.

This thesis also explored the curation of other novel datatypes. For example, I collected cell-type data from PangloaDB but this data was not complete enough to pass feature cleaning, and will be interesting to include in future work as the experimental data expands. Other novel databases, such as the Human Cell Atlas and Expression Atlas, are also building cell-type data, suggesting that over time cell-type information will become available. This data may be particularly important feature input as it could provide ML models with potentially more granular information that could lead to more specific biological insights. For instance, from the BP ML framework, the top prioritised gene (*COL15A1)* was shown in PangloaDB to have its highest cell clustering in human endothelial cells and smooth muscle cells in the testis – which has also been shown to interact with the renin-angiotensin system to affect

antihypertensive drugs[170]. Such cell-type data can increase the specificity of supporting evidence and so it should be revisited to both incorporate as a feature in ML and overlay with previously prioritised genes to better understand their biology.

In addition to cell-type data, animal model, epigenetic, and variant-level data (e.g., pathogenic scores and LoF measures) were also too missing to enter ML, indicating that as experimental data becomes more readily available, features should be updated to reflect more complete information that could advance ML for gene prioritisation. This point is especially pertinent as with improved annotations ML can be readily applied to prioritise genes in the non-coding genome, which is understudied in post-GWAS machine learning prioritisation[24].

5.1.2 Evaluation of Supervised Learning Approaches

Alongside the curation of features, the training genes identified in chapter 2 showed how BP genetics is ideally suited to a ML approach, with thousands of associations from which to subset training data for supervised learning. The 7 million variants from the BP GWAS allowed for the testing of various filters for dividing gene groups. Specifically, from the 33,847 genes annotated to the 7 million variants, genes with BP drug interactions, genes with text-mining significance in relation to BP literature, genes with BP annotations in IPA, and genes meeting several selection criteria (outlined in the chapter 2 methods section 2.2.3) to be deemed least likely BP genes were partitioned into four groups. These four groups then served as classifications in ML, used in both multiclass classification and regression analysis. The multiclass approach in chapter 3 showed that three of the gene groups as opposed to all four gave

a better ML performance. This difference may be due to the IPA annotation not truly being informative enough (which is supported by it not being a selected feature on the 3-label multiclass approach). It may also be due to simple chance, as a 3-label approach increases the odds of a model predicting correctly (1/3 probability of a correction prediction at random) in comparison to a 4-label approach (1/4 probability of a correct prediction at random). However, whether these observations are true would need to be further tested, with using balanced classes as the training data size and class imbalance differences between the 3-label and 4-label data will also heavily influence a model's accuracy.

In comparison, binary classification could also have been investigated. However, each group in a multiclass approach provides more distinction than binary classification as models can recognise similar or differential patterns between definitively known causal genes and probable BP genes that have a less established BP relationship. Unlike in binary classification, which is a common ML prioritisation approach[19, 33], where the opaquer intermediate gene groupings will be forced into either definitively positive or negative classes, and their probability for their positive prediction is used as a prioritisation score, overall providing less stratification and therefore a less informative gene ranking.

One interesting aspect explored here is the diverse range of supervised-learning methods. Multiclass classification in chapter 3, showed a poor ML performance overall on balanced accuracy for training data with both 3 labels and 4 labels (the highest being 68% balanced accuracy from the 3-label dataset) - suggesting the

training data would need further curation or class balancing approaches to improve the performance. Oversampling and class weighting with probability calibration were also able to improve performance (81% balanced accuracy for oversampling versus 71% for class weighting). However, due to the risk of overfitting known in class balancing approaches[128] the more conservative improvement in model performance with class weighting was selected for further analysis. Furthermore, probability calibration was used to adjust output predictions, providing a confidence-level interpretation of the predicted probabilities by overlaying a regression on to the output multiclass probabilities. However, on exploring the prioritised genes, in multiclassification, 172/512 of the genes classified as least likely had the same probability and almost entirely missing features. The classification of these genes may be due in part to the quality of least likely training gene curation, with insufficient information about non-causal BP genes to prioritise them in a way which reflects the possibility that they may act in unexplored BP mechanisms. However, on re-framing the ML into a regression analysis (using the same features and training data) these 172 genes had various predicted scores. For example, the multiclassification predicted all olfactory genes, which have had research identifying their potential link with BP[171], in the data had a least likely label of the same probability. Meanwhile for the regression analysis, selected olfactory genes were prioritised with scores > 0.5 (e.g., *OR2I1P* scored at 0.67 on regression but classed as least likely on multiclassification) showing the benefit of applying ML prioritisation on a continuous scale rather than classes. However, there is a risk that these prioritisations and their selection at their locus are spurious. In comparison, the regression method prioritised 68 out of the 172 genes with the same low prioritisation score (0.102 XGB score) – suggesting the

underlying issue is still present and that this will need improved least likely training gene curation in further work.

As mentioned in chapter 4, the regression approach applied here is, to the best of our knowledge, the first of its kind for post-GWAS prioritisation. In the context of optimising the ML framework, changing from classification to regression uniquely provides more flexibility in model-decision making despite the underlying class imbalance being unchanged. The regression approach also provided further ML comparison, which became necessary due to the overfitting performance on classification and the ML decision-making conflicting with domain biology (as seen in chapter 3, where the most likely and probable BP gene classifications indicated conflicting use of HIPred to discern between the groups). While classification and regression metrics measure model error in ways that are not directly comparable, regression analysis provided a slightly increased feature selection (including IPA BP annotation, liver gene expression, EBV-transformed lymphocytes gene expression and dropping heart atrial appendage gene expression in comparison to the selected classification features). Notably, the selection of IPA BP annotation increases the phenotypic understanding of the regression model, whilst the selected tissues that are not sites of action for BP potentially increase the ability for a model to recognise novel biological BP functions – overall suggesting the regression analysis may have more data to develop more nuanced pattern recognition.

5.1.3 Model Benchmarking for Gene Prioritisation

The models tested here matched the performances of previous studies. The top-performing model being XGB in chapters 4 and 5, and its performance being very close to other gradient boosting models (CB, LGBM, and GBM) aligns with previous ML prioritisation research post-GWAS[19, 33] and research that found XGB and CB are top-performing models on tabular datasets for supervised learning (when compared against 15 other models, including those benchmarked here)[129]. Meta-ensemble models also showed similarly high performances to the gradient boosting models, with voting and stacking meta-ensemble models consisting of each of the benchmarked models and the bagging model using the top-performing model. However, whilst the meta-ensemble models perform well and validate the performance of their base learners, they were also more time-consuming on hyperparameter tuning for equivalent or slightly lower performances in comparison to the base models in chapters 3-5. Overall, suggesting they are beneficial to benchmark for validation but should only be selected for further analysis if they have a notably stronger performance that would make the computational efficiency trade-off worthwhile.

Vitsios et al. (2020) found from their benchmarked models (RF, XGB, GBM, ET, DNN, and SVM) all had performances between 0.831-0.85 AUC, with XGB being the top-performing and SVM the lowest. However, whilst their SVM and DNN performed very similarly to the tree-based approaches, the benchmarking in this thesis found both performed worse than other models. For example, on multiclass classification SVM and NN had 0.7-0.71 balanced accuracy compared to 0.79-0.84 for the gradient boosting models, and on regression SVM performed poorly (0.2 median r^2 versus

0.744 median r^2 for XGB). SVM's lower performance (alongside the lower performances also seen in linear regression models in chapters 3-5) is possibly due to the complexity of the data, having non-linear features of both categorical and continuous data, creating difficulty as the underlying decision-making of the models is driven by linear functions. In comparison, NNs are known to perform badly on tabular data in comparison to other commonly used models (XGB, CB, GBM, KNN etc.)[129], and on application, in chapter 3 it had lower performance combined with the additional computation time (>12 hours required to tuned a benchmark on classification due to the larger number of neural net parameters that are possible to tune). Furthermore, tuning the hyperparameters of NNs is complex due to deep learning architecture having hundreds if not thousands of possible structures to test and tune to create a neural network[172]. Each of these points led to the NN model not being benchmarked on regression analysis.

Overall the model benchmarking here, alongside that also explored by Vitsios et al. (2020), agree with the no free lunch theorem. There is no one-size-fits-all model for ML applied to gene prioritisation and benchmarking a range of models is a crucial step in developing robust prioritisation. However, models such as XGB and CB which are known to perform well on tabular data and have had validation as top performers, highlight themselves as leading models of choice for gene prioritisation. Furthermore, their speed and functions to address overfitting (additional regularisation parameters beyond that in GBM) also best position them for biological problems where it is important to understand as much as possible within model decision-making.

5.1.4 Gene Prioritisation Key Findings

5.1.4.1 Defining Causal Genes for High Blood Pressure

The optimised ML prioritisation of genes in BP GWAS loci offers a ranked list of putative causal genes for functional investigation. The prioritised genes represent many functional categories, including genes that may represent novel drug targets, or which interact with known BP drug targets, new genes acting in established BP pathways, and genes with overlapping high prioritisation across all methods. Each of these categories demonstrate the benefit of using a comprehensive ML framework and presents good evidence to streamline hypotheses development to improve understanding of BP biology and identify new drug targets.

The prioritised genes from the total 1,804 predicted genes by the optimised XGB model found several genes worthy of functional follow-up, whether by taking directly from the highest prioritised genes (such as the top-prioritised *COL15A1*[144, 146]), gene-drug interactors (e.g., *PTPRD*[154], *CHRM2*[156] or *INSR*[158]), or gene enrichment analysis (*SLC8A1*[163, 164]). While each of these putative candidate genes may warrant further investigation, their true impact on BP (and the accuracy of their ML prioritisation) can only be confirmed by experimental follow-up. However, many prioritised genes already have some evidence of therapeutic potential, at the highest level, 9 genes are recorded to interact directly with BP drugs, while 62 genes interact with drugs that have BP side effects. Although it should be noted that these findings were annotated from drug databases (BNF, SIDER and DGIdb), all require validation, some database annotations may be spurious, and ultimately, laboratory follow-up and clinician input,

may be required to confirm their potential. Clinical validation, especially for the 9 BP-drug interacting genes, would also indicate that these genes could serve as additional training data, increasing the most likely BP training gene size from 51 to 60, possibly increasing ML performance as well as our understanding of BP biology.

The analysis of *selected-gene* PPIs with BP-drug mechanisms found hundreds of prioritised genes with potential roles in BP mechanisms (Figure 4.5, Appendix C Table 9) awaiting discovery. Furthermore, these genes at most represented 26.1% of PPIs for a given BP-drug mechanism, suggesting that the genes posited as having therapeutic potential here are the tip of the iceberg. From the *selected-genes,* 733 also interact with a drug mechanism that is not a BP drug, suggesting there is potential in these mechanisms to have a re-purposed effect on BP regulation. Additionally, *BRD1, MFSD6* and *PLEKHG1* were highlighted as having multiple BP drug mechanism overlaps only and there were also 7 other genes (*TRPC4AP, INO80, CCDC68, ODF2L, TNRC6A, OR4C13, LAMB2)* that interacted with singular BP drug mechanism PPIs (and no other drug mechanisms). Only *LAMB2* interacts with a drug annotated in DGIdb, interacting with ocraplasmin (similar to the top prioritised gene *COL15A1* as discussed in chapter 4). Ocraplasmin is used to treat vitreomacular adhesion by breaking down extracellular matrix components. It has also been shown to be beneficial to diabetic macular oedema[173]. Further study of this treatment mechanism in diabetes may increase the understanding of the genes interacting with ocraplasmin, and particularly their role in fluid balancing for oedema more generally – with this being of interest as it is also a related mechanism that impacts BP. However, while this suggestion is speculative, the fact that *COL15A1* and *LAMB2* are highly

prioritised and interact with this drug suggests there is an underlying mechanism of interest for further BP research.

It is also intriguing to note, that pathway analysis of *selected-genes,* divided into four groups with increasing known drug mechanism interaction (Figure 4.9), shows that the strongest pathway enrichment seen in all *selected-genes,* is substantially represented by under 25% of *selected-genes* which interact with 5 or more drug mechanisms.

These genes may offer immediate opportunities for genetic stratification of patients who are more likely to respond to certain classes of drugs, using recently described pharmacogenomic polygenic risk scores[174]. Perhaps more importantly, the 621/794 (78%) *selected-genes* with <5 drug mechanism interactions, potentially define a substantial range of novel or less-exploited biological mechanisms underpinning BP, suggesting that antihypertensive drug discovery has so far focused on a quite limited mechanistic range of BP biology. These genes include 124 drugged (only 7 with BP indication) and 174 druggable genes, all of which could potentially represent highly novel target mechanisms for blood pressure. Considering some potential repositioning opportunities among these 621 genes, melatonin receptor 1B (*MTNR1B*) has a selective agonist ramelteon, approved for insomnia, which has been shown to attenuate age-associated hypertension and weight gain in spontaneously hypertensive (SHR) rats[175]. In some cases, molecules are identified which act on a number of targets among the *selected-genes.* Hesperadin is a naturally occuring citrus flavanone, with aurora kinase inhibitor activity, it is known to act on over 24 kinase targets, including

7 *selected-genes* (*EIF2AK4*, *MERTK, FER, FES, BLK, CSK, FYN*)[176]. A number of studies highlight the cardioprotective properties of hesperidin and its aglycone, hesperitin[177], including reduction of systolic BP, endothelial dysfunction and oxidative stress in SHR rats[178]. Mechanistic understanding of hesperidin action is limited, but multiple hypotheses are proposed[177], identification of a network of kinase targets of hesperidin with BP association, may offer a valuable insight into a potentially novel drug mechanism.

Novel pathway enrichment observed among *selected-genes* (Figure 4.9c) with <5 drug mechanism interactions, include *Autophagy* (*MAP2K2, RRAS, ATG7, MRAS, VMP1*) which has been extensively linked with BP via mechanisms of mitochondrial and endothelial dysfunction[179, 180]. Two pathways show enrichment potentially with a common angiotensin II mediated mechanism. *Neutrophil extracellular traps* (*MAP2K2, RELA, HDAC9, ATG7, HDAC7, H2BC12, CDK6, HLA-B*) are released by angiotensin II and mediate essential hypertension by a fibrotic mechanism leading to endothelial dysfunction[181]. Conversely, *hedgehog signalling* (*BCL2, CSNK1G3, BTRC*) has been shown to correct angiotensin II-induced hypertension and endothelial dysfunction in aorta through induction of NO production and reduction of oxidative stress[182].

Total interactions between each *selected-gene's* encoded proteins and drug mechanisms were annotated, showing the potential for novel drug targets within BP GWAS loci (Fig. 4.9). 173 of *selected-genes* genes interact with more than five drug mechanisms, possibly having more ubiquitous effects, whilst the majority of genes

(n=374) had 1-5 interactions all involving BP-related and non-BP-related drug mechanisms. Overall, the heavy interconnectivity between the *selected-genes* interacting with multiple BP drug mechanisms, alongside their higher enrichment across pathways especially those that are BP-related such as renin secretion and cGMP-PKG signalling (Figure 4.9), shows that genes that overlap in multiple drug mechanisms are likely to overlap in multiple pathways. This result, combined with the *selected-genes*' prioritisation, puts forward a stratified view of the genetics underpinning BP drug mechanisms for further investigation.

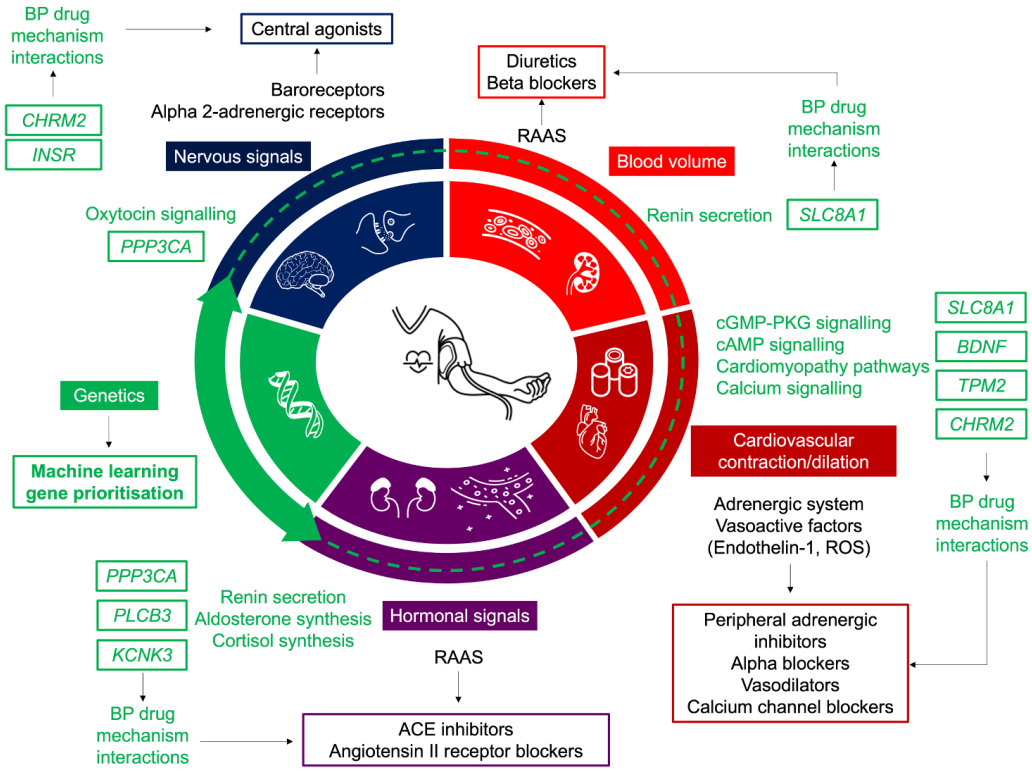
The drug mechanisms also have a range of representations among the prioritised genes by GWAS. For example, the largest coverage by the *selected-genes* was 26.1% of the total endothelin inhibitor mechanism PPIs, meanwhile, the smallest coverage was 6.6% of the alpha-2 adrenergic receptor agonist interaction mechanisms (Table 4.3), suggesting further work is needed to illuminate the genes relevant to these drug mechanisms. However, from the mechanism interactions represented by the *selected-genes*, and their heavy interconnectivity between the BP drug mechanisms (Figure 4.9a), suggests they may all be acting on similar pathways that affect BP. The *selected-genes* being active in overlapping BP-relevant drug pathways indicates the reliability of their ML prioritisation, but also shows a need for further analysis to put forward novel drug mechanisms for BP.

The prioritised genes discussed in chapter 4 as having potential translational insights are also highly prioritised by OpenTargets L2G score[137] (for example the L2G score

for *CHRM2* is 0.879, for *INSR* it is 0.83, and for *PTPRD* it is 0.878), with these genes being a part of the 76 genes prioritised highly by both methods. The L2G score prioritises genes with similar ML principles (for example XGB is used, although binary classification is then applied as opposed to regression) but uses a different set of features (fine-mapping, disease-disease colocalisation analysis and disease-molecular trait colocalisation analysis across 92 tissues and cell types, phenome-wide association study analysis, and enriched trait evidence), making its pattern recognition based on entirely different influences compared to the regression analysis developed in chapter 4. The high prioritisation by both methods further emphasizes that these genes are key findings that need further investigation and shows how combined approaches of prioritisation can be used to support findings. Moreover, the different ML methodologies make their concordance even more validating for their prioritised genes. Therefore, whilst the ML framework developed here can suggest potential candidate BP genes for further investigation, the overlaying of additional prioritisation methods can strengthen the biological evidence - which then will enable less laborious experimental research with an even more refined gene list.

From the 794 s*elected-genes* several key findings arose that expand research directions and can better illuminate the biology of high BP (Figure 5.1). Notably, *selected-genes* genes that interact with nervous system targeting drugs have highlighted themselves (*CHRM2* and *INSR*) as their interacting drugs approved for other conditions (depression, migraines and weight loss) show promise as re-purposing targets for hypertension[155, 157]. The understanding of cardiovascular and hormone signalling genes with likely BP roles has also grown, with *selected-genes* such as *SLC8A1*

(scored 0.86 by XGB) being shown here as interacting across BP-drug mechanisms and in the enriched pathways of renin secretion, cardiomyopathies, and cGMP-PKG



signalling. Importantly, many *selected-genes* were found to have impacts on multiple pathways and mechanisms across BP biology. For example, alongside *SLC8A1*, *PPP3CA* (scored 0.78 by XGB) acts on renin secretion, cGMP-PKG signalling, and oxytocin signalling, all of which were significantly enriched pathways (Figure 4.12). Such results emphasize the knotted tangle of overlapping genetic contributions for BP, with more examples likely to be discovered in the future. However, the gene prioritisation and downstream analysis presented here sets the foundation for targeted functional investigations, which may more effectively unravel the genetic influences underpinning the landscape of BP biology.

Figure 5.1. Overview of blood pressure biology and the implications of gene prioritisation. Regulation of blood pressure (BP) involves the interactions of several

organ systems, which are predominantly cardiovascular, renal and neurological. Furthering the understanding of the genetic component of BP by machine learning prioritisation in this thesis has led to an emphasis on the importance of nervous and hormonal signalling pathways for BP, it has highlighted genes with roles in established cardiovascular disease pathways that warrant further investigation, and it has shown gene-drug interactions with highly prioritised genes having interactions with BP drug mechanism as well as other drugs (such as antidepressant and migraine drugs which act on nervous signalling pathways) that have potential for drug re-purposing.

5.1.4.2 Extending the ML Framework to Blood Lipid Traits

Extension of this ML framework to blood lipid traits identified novel insights to better understand the regulation of lipid metabolism. However, to meet the requirements of this different GWAS trait, the framework had to be adapted (with differing least likely gene curation) and the training data groups (gold and silver set genes) were previously developed in another study not focused on ML[183]. These differences also led to a framework applied to all five blood lipid traits at once, as opposed to individually, and overall gave a comparable ML performance to that of the BP prioritisation. For example, XGB was the selected top-performing model and had a median r^2 of 0.707 and predicted r^2 of 0.826 for blood lipid trait prioritisation, whilst for BP prioritisation XGB was also the top-performing model with 0.744 median r^2 and 0.897 predicted r^2. These similar ML performances, despite bespoke tweaks to the framework, suggest that the ML framework is robust enough to perform well on re-application - even when the framework is altered to meet the needs of a specific phenotype and its GWAS data.

However, considering the blood lipid prioritisation framework was most notably altered to have less least likely gene filtering, this also indicates that the ML performance could still be improved with opportunities for higher-quality training data.

On exploring the model decision-making, the SHAP feature importance showed consistency in highly ranked features compared to the BP prioritisation (e.g., HIPred, Exomiser scores, and pLI), however, it uniquely valued sex-specific GTEx tissues and also several GTEx tissues passed feature selection that were then shown to have minimal XGB model influence (Appendix D Figure 5.5). This feature importance ranking indicates that the model may be identifying the biological relevance of sex-specific tissues that are known to have roles in lipid metabolism[184-186], but on the other hand, the model may also be being swayed by sex-specific bias, which has been noted in GTEx data[169]. Overall, this result highlights a limitation of the ML framework and its black-box nature, making further research reliant on functional validation to understand any sex-specific genetic relationships being prioritised by ML.

When focusing on the training data, the gold and silver standard genes were not curated using drug or text-mining data like the BP framework, and this is potentially reflected in the highly prioritised genes. For example, on analysing the highest prioritised *selected-genes*, none have interacting drugs in DGIdb for any condition and only two have been points of focus in lipid metabolism research (*KHK*[187] and *CREB3L3*[188]), with the rest being briefly mentioned in other lipid or cardiovascular-related research (e.g. due to their transcriptomic expression[189]). Whilst these results

do not eliminate the genes from potentially being influential lipid-regulating genes, they imply the impact of the differing training gene criteria and feature selection compared to the BP ML. For instance, despite having little drug or literature evidence, the actual proteins encoded by each of the top ten genes have metabolic functions that relate to lipid metabolism (e.g., fructose metabolism or bile synthesis, Appendix D Table 5.2) suggesting that with further research these genes may reveal translational insights. This implication is also validated when considering the genes scored >0.8 and the *selected-genes* have significantly higher SDI drug probabilities in comparison to the lower prioritised genes and all genes in the SDI database (Appendix D Table 5.7).

Furthermore, the most significantly enriched pathways for the *selected-genes* were known lipid metabolism pathways (e.g., cholesterol metabolism and PPAR signalling – Appendix D Figure 4.5, Appendix D Table 10). This result, similar to that of the BP prioritisation, shows that further analysis of the genes inside these pathways may be informative. For example, *CD36* was shown as interacting within the top five enriched pathways (Appendix D Figure 5.10). The gene also interacts with the angiogenesis inhibitor and cancer drug ABT-510[190] and it has known roles in atherosclerosis and lipid metabolism[191]. This gene highlights the circularity in pattern recognition being provided by the ML framework, however, it also suggests that by further researching genes such as *CD36*, we may improve drug target specification in known disease-causing pathways.

One of the key findings in Appendix D was the comparison of ML prioritisation with other methods, particularly those from recent studies that also focused on ranking genes from the same GWAS. In a similar way to the BP prioritisation in chapter 4, the blood lipid trait prioritisation had positive correlations with all other prioritisation methods it was compared with (Appendix D Table 5.3). It also showed concordance with confidence assignments given to genes by Kanoni et al. (2021), suggesting reliability in the model's decision-making. However, the model also had conflicting prioritisations of genes with medium-low and low confidence assignments by Kanoni et al. (2021). Ultimately, conflicting prioritisations of individual genes provide opportunities for future targeted research. For example, a broader comparison of prioritisation methods or multi-layered evidence could be collected to confirm a gene's prioritisation (as shown by Ramdas et al. (2021)), or conflictingly prioritised genes could be removed to develop an even more select gene list, ensuring the most concordantly prioritised genes are efficiently ranked for functional research. Furthermore, conflicting gene rankings also suggest that gene groupings could be altered to improve prioritisations. For instance, the medium-low and low-confidence gene groups by Kanoni et al. (2021) could be combined and have a stronger evidence base for their grouping. Meanwhile, for the XGB model the silver set training group provides a large class imbalance, suggesting this gene group could be further filtered to identify genes with stronger blood lipid relationships than mouse models alone, and this may then reduce the majority of genes that XGB prioritised ~0.7 and improve model performance.

5.2 Limitations

Although powerful, machine learning approaches applied to data analysis do have limitations which should not be overlooked and taken at face value could hinder efficient locus prioritisation. For example, any ML framework is susceptible to overfitting, and this is also further impacted by a lack of complete transparency in model interpretation via SHAP. While in chapter 4 overfitting is somewhat circumvented by switching from multiclass classification to regression analysis, the class imbalance is still present in the unchanged training data. The regression analysis has metrics for assessing overfitting (e.g., predicted r^2 or adjusted r^2) however in all model performances no model had a perfect performance in these measures, and even then, these metrics are not able to catch all overfitting. This prevents an unavoidable limitation of this work that can only be further investigated with external validation datasets, enabling exploration of how well a model generalises to unseen data.

I also explored model decision-making using SHAP, providing an interpretation of how features influenced prediction, which could be aligned with domain biology to check a model's understanding of disease-causing genes. However, SHAP can only illuminate so much of the black-box decision-making being undertaken, providing insight only into how features influence model predictions and not into how the features influence the true target variable. This means that there are still unanswered questions in relation to which of the highly prioritised genes are truly causal and why, and that answering such questions can likely only be found in functional research.

Perhaps one of the greatest challenges is model training that enables accurate predictions on non-training data – sometimes referred to as "knowable unknowns". The work in this thesis is limited by this challenge. The optimised XGB model prioritising genes in chapters 4 and 5 output highly ranked genes that were enriched for already known BP and blood lipid pathways, as proven by gene enrichment analysis. While this validates the model's decision-making it also indicates less opportunity for identifying novel insights downstream with enough evidence to justify functional research. Furthermore, the gene per locus selection method was developed to combat situations where a ML model may score genes closely at a locus, without large score margins between genes that clearly identify one most likely causal gene at a locus. With the use of BP and blood lipid gene PPI data for the gene(s) selection per loci there are also limitations. The PPIs collected were lists of genes with direct and secondary PPIs with the BP-regulating and gold standard blood lipid genes, which were then used to select genes with the highest number of these PPIs per locus, if ML prioritisation did not distinctly prioritise any one gene at a locus. These PPIs filter out less likely top genes per loci via a known biology filter. However, this selection method is also limiting, with PPI information relating to known disease-related genes abetting circular selection for genes with more PPI evidence.

Additionally, for the BP GWAS application, given that there are few genes with interacting BP drugs, and with the model using these genes in its training data, it was not possible to create an external validation set for further model testing (with this also being an issue for the small gold standard set in the blood lipid application). This testing will be required as new known disease-causing genes are discovered to assess

any inaccurate modelling assumptions[165]. Curation of the *non-BP-genes* that are least likely causal training genes with any certainty is also a challenge across diseases, despite stringent criteria for selecting the *non-BP-genes*, there is a risk of false negative examples being used in training. These issues in training gene curation also created an imbalanced minority class that, whilst regression analysis and nested cross-validation aim to minimise, is still underlying in the training data and poses an overfitting risk. This issue is also present in the blood lipid application, with few least likely genes that provide minimal training opportunities for a model in comparison to the majority probable gene grouping. Furthermore, both least likely gene curations for BP and blood lipids produced genes with the smallest gene lengths in their training data, which in turn will provide less opportunity for these genes to have annotations across databases.

Also on training, more features (e.g., cell-type data or pathogenic variant scores) were not tested on ML due to heavily missing features – however, it is with more diverse features at different molecular scales that ML may become more empowered to recognise novel patterns and provide new insights. In theory, the approach of imputing all missing values with zeros could be applied. However, this would in turn limit the model benchmarking, requiring models that are robust to such noisy features, and this also still does not guarantee the more robust models (such as XGB) are not being affected by overfitting. Furthermore, genomic data such as gene length was collected to assess any risk of genetic bias, however further testing of biological bias could still be performed (e.g., by identifying interacting genes) and will be needed in future work to ensure robustly selected genes are entering the training data.

Another key limitation of this work is the focus on protein-coding gene prioritisation. This was due to the training genes being protein-coding, which is also because protein-coding genes are better studied and therefore more likely to have annotations that equate to more complete features in ML. As the non-coding genome is researched and data is collected this will hopefully expand the possible training data and ability for ML to prioritise genes across the whole genome. Overall, as research expands (such as larger GWAS across more diverse populations, and the development of more comprehensive annotations) new inputs for BP gene prioritisation will develop that can advance ML prioritisation.

5.3 Future Work

For multiclass classification, improvement in the training data quality would be the most beneficial step forward to the framework built in this chapter. Hyper-parameter tuning of the models could also be more comprehensive. For example, Bayesian hyper-parameter tuning was performed here, but a more thorough grid search of a large range of hyper-parameters could be investigated to ensure the global optimum is being selected. However, this comes at a cost of computational efficiency.

The regression analysis could be further developed by advancing the model benchmarking phase – again by testing more comprehensive hyper-parameter tuning, but by also comparing with deep learning. While the time investment required in a deep learning test would further hinder the computational efficiency, confirming the

deep learning regression analysis and whether it performs similarly to its benchmarking in multiclass classification would work to validate the model comparison seen in chapter 3 and affirm whether it should be included in future benchmarking tests.

Beyond ML improvements, a key focus point for future work of the top-performing model will also be in furthering model decision-making interpretation. For example, SHAP was used throughout this thesis as it has comprehensive visualisation parameters as well as both local and global ML interpretation. However, other explainability tools exist (such as LIME) that would confirm or rebuke local model interpretations, identifying which gene prioritisations may need further analysis. Furthermore, a developing area of ML is bias auditing, and ML fairness toolkits are being built by companies such as IBM and Google, which develop metrics for assessing model bias[193]. However, these are susceptible to "fairness gerrymandering" where the metrics are generated as they are more likely to produce more appealing results[193]. Whilst such packages are still being developed, dedicated ML for bias auditing that is tailored to genomic data could be developed in future work, such as that shown by Eid et al. (2021) for drug and PPI ML prediction problems.

For the future work of blood lipid trait prioritisation, there are several potential directions to improve the ML framework built in Appendix D. For example, training gene curation could be further fortified with expert clinical validation of the gold and silver set genes, overlaying the BP training gene criteria with that already defined by Kanoni et al. (2021), which would likely improve ML performance and address the

large class imbalance. Exomiser scores could be calculated for more stratified HPO terms. 'Abnormal lipid circulating concentration' was used in Appendix D as it is the highest overarching lipid term in HPO, however, sub-branching terms for high cholesterol, triglyceride levels, HDL and LDL levels could be used to divide the ML framework into iterations per each blood lipid trait. The development of individual Exomiser scores tailored to each trait would warrant their individual prioritisation framework, as otherwise, the features provided would not differ per the training data for each trait. Five individual re-applications per each blood lipid trait would increase the amount of computational effort needed. However, the number of genes prioritised per trait would be smaller, and if the features can be more bespoke to that blood lipid trait it may provide a more informed trait-specific ML prioritisation, alongside making each individual re-application more efficient. Also, the number of least likely blood lipid training genes would increase as the least likely gene filtering criteria (LD, p-value, and PPIs) would only need to be filtered against genes associated with one blood lipid trait as opposed to all five. Furthermore, alongside altering the training data, the comparative prioritisation by Kanoni et al. (2021) also highlights the potential for future work to combine prioritisation methods, For example, a meta-analysis merging the prioritisations from ML and other methods using different input features (e.g., fine-mapping, TWAS, PoPs, OpenTargets L2G, Mantis-ml etc.) may produce a more robust gene ranking.

The machine learning explored in this thesis also has potential as a variant prioritisation framework. However, such an approach needs high quality variant annotations to address data missingness and positive-labelling of disease-specific

pathogenic variants. If such roadblcoks are addressed there is great potential for ML including deep learning, for which whislt it has shown poor performance on tabular data, it has been shown to prioritise variants impacting immune disease accurately when given sequencing data[39]. This suggests deep learning could uniquely use natural language processing and input textual data (e.g., reference and alternative allele information or even variant function descriptions) to potentially have a more competitive ML performance for variants than seen at the gene-level.

5.4 Future Implications

5.4.1 Machine Learning Methodology Post-GWAS

The results of this work enable researchers of complex traits to analyse genes without selecting them based on 'cherry-picking' bias but by selecting prioritised genes with a stronger evidence base for their potential impact. However, as multi-omic data improves the prioritised genes and the ML methods underlying them should also be updated. Particularly, as ML is a broad research field with methodologies that have not yet been tested in genomic applications. For example, unsupervised ML has not been focused on in this thesis due to a lack of interpretability, however, unsupervised methods do exist that give a view of models under the hood. For example, one-class learning (using a training set with only one gene group) is an unsupervised method that plots a decision boundary between one group versus everything else and is interpretable via packages such as SHAP. It is usually used for problems such as anomaly detection, but it could be tested for identifying the least likely disease-causing

genes as ‘outliers’ from a list of known disease-causing genes. Furthermore, unsupervised learning could be applied after supervised learning – for example, by using output SHAP values to cluster genes and gain an abstraction of gene-gene relationships with ML. More complex ML could also be developed, such as reinforcement learning, which has been used to detect SNP-SNP interactions or deep learning[194]. Deep learning has had success with sequence data as input, however, it has yet to have equivalent success with gene-level tabular data. Methods are being developed to address deep learning’s often poorer performance on tabular data[129], suggesting it should not be ruled out from model benchmarking in the future.

Another key aspect of ML for prioritisation that needs future focus is interpretability. Whilst in this thesis I focus on SHAP, other explainability tools exist that could also be compared - and due to different underlying mathematical principles, they may interpret model influences differently. This has been a focus of a recent package interpretML developed by Microsoft, which allows the user to create explainability dashboards showing model interpretations from several methods (SHAP, LIME, and sensitivity analysis). Furthermore, they also developed an explainable boosting machine which aims to be a ‘glass box’ model and have more transparent decision-making than black box models – by the approach learning one feature at a time as an additive model. However, currently, this method is very time-consuming to train or tune the parameters and so would need improved computational efficiency to be effective on large GWAS data.

5.4.2 Accessibility and Combinational Approaches

Research should continue to develop models aiming to prioritise genes across diseases, with methods that can be reused by other researchers, and with consideration for the size of present GWAS data, varying datatypes, and feature importance. Doing so could then lead to more accessible and reusable models - for example with source code or web interfaces that are useable by a wider range of GWAS researchers – and create more globally implemented ML applications for GWAS prioritisation, thus accelerating researchers towards the post-GWAS endgame of understanding disease. OpenTargets is one of the leading methods on this front, addressing this need for an accessible tool with a web interface containing interactive gene rankings per loci for many GWAS'[137].

Future work may benefit most from exploring combinational methods, enabling augmented results with a stronger statistical evidence base. This works as each method can lessen the limitations of the other or work together as a voting system to support or disagree with each other's predictions, providing more reliable evidence and allowing for functional researchers to make more informed decisions. A few combinational approaches have already been developed, such as that used by Kanoni et al. (2021). Although in most combined approaches, when ML is applied it is used as a pre-processing step, whether developing priors in fine-mapping or in developing MR tests[195, 196]. Using ML particularly for providing functional information in fine-mapping may shed more insight on biological trends underpinning select variants and indicates ML can solidify its place in the post-GWAS pipeline whether as a standalone tool, such as the frameworks developed in this thesis, or in hybrid methods. Beyond this, ML is also on a steep curve of research attention across many domains, suggesting

that it unlike other methods may advance further with intersectional developments and implies the full potential of ML in post-GWAS analysis is still to come.

5.5 Conclusion

This thesis shows how ML is gradually proving itself to be a valuable tool for post-GWAS analysis. The methodology and training data developed here show an optimised performance for prioritising loci. The output prioritisation results have the potential for functional and translational validation. For complex diseases, such as hypertension, the ML method's ability to generate hypotheses has streamlined functional work that gives biological insights - enabling the unravelling of how associated loci may affect cardiovascular traits. However, before ML models, such as the frameworks put forward in this thesis, can consolidate their role in the post-GWAS analyses, research needs to address several aspects ranging from genomic training data curation to reproducibility, and accessibility. There also needs to be a greater comparison between ML and other prioritisation methods to understand ML's place in the post-GWAS pipeline and enable GWAS to truly provide the projected biological insights and translational capability that it has been so long promised.

6 Abbreviations

Abbreviation	Description
BNF	British National Formulary
BP	Blood pressure

CB	CatBoost
CV	Cross-validation
DBP	Diastolic blood pressure
DEPICT	Data-driven Expression Prioritized Integration for Complex Traits
DGIdb	Drug Gene Interaction Database
DT	Decision tree
EDA	Exploratory data analysis
eQTL	Expression quantitative trait loci
GB	Gradient boosting
GDI	Gene damage index
GO	Gene ontology
GWAS	Genome-wide association study
HIPred	Haploinsufficiency prediction score
i.i.d	independent and identically distributed
IPA	Ingenuity Pathway Analysis
IMPC	International mouse phenotyping consortium
KNN	k-nearest neighbours
LD	Linkage disequilibrium
LGBM	Light gradient boosting (lightgbm)
LIME	Local interpretable model-agnostic explanations
LoF	Loss-of-function
LR	Logistic regression
ML	Machine learning
NN	Neural network
pLI	Probability loss-of-function
PoPs	Polygenic priority score
PP	Pulse pressure
PPI	Protein-protein interaction
RF	Random forest
SBP	Systolic blood pressure
SDI	Subcellular Diversity Index
SHAP	SHapley Additive exPlanations
SNP	Single nucleotide polymorphism
SVM	Support vector machine
TPM	Transcripts per million
XGB	eXtreme Gradient Boosting

7 Bibliography

1. Roth GA, Johnson C, Abajobir A, Abd-Allah F, Abera SF, Abyu G, et al. Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. Journal of the American College of Cardiology. 2017;70(1):1.
2. Surendran P, Feofanova EV, Lahrouchi N, Ntalla I, Karthikeyan S, Cook J, et al. Discovery of rare variants associated with blood pressure regulation through meta-analysis of 1.3 million individuals. Nature Genetics. 2020;52(12):1314-32.
3. Evangelou E, Warren HR, Mosen-Ansorena D, Mifsud B, Pazoki R, Gao H, et al. Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. Nature Genetics. 2018;50(10):1412-25.
4. Giri A, Hellwege JN, Keaton JM, Park J, Qiu C, Warren HR, et al. Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. Nat Genet. 2019;51(1):51-62.
5. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nature Reviews Genetics. 2019;20(8):467-84.
6. Hurle MR, Nelson MR, Agarwal P, Cardon LR. Impact of genetically supported target selection on R&D productivity. 2016.
7. Hormozdiari F, Kichaev G, Yang WY, Pasaniuc B, Eskin E. Identification of causal genes for complex traits. Bioinformatics. 2015;31(12):i206-13.
8. Weissbrod O, Hormozdiari F, Benner C, Cui R, Ulirsch J, Gazal S, et al. Functionally informed fine-mapping and polygenic localization of complex trait heritability. Nature Genetics. 2020;52(12):1355-63.
9. Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E. Identifying causal variants at loci with multiple signals of association. Genetics. 2014;198(2):497-508.
10. Pickrell Joseph K. Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits. The American Journal of Human Genetics. 2014;94(4):559-73.
11. Kichaev G, Yang W-Y, Lindstrom S, Hormozdiari F, Eskin E, Price AL, et al. Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. PLOS Genetics. 2014;10(10):e1004722.
12. Olczak KJ, Taylor-Bateman V, Nicholls HL, Traylor M, Cabrera CP, Munroe PB. Hypertension genetics past, present and future applications. Journal of Internal Medicine. 2021;290(6):1130-52.
13. Wu M, Zeng W, Liu W, Lv H, Chen T, Jiang R. Leveraging multiple gene networks to prioritize GWAS candidate genes via network representation learning. Methods. 2018;145:41-50.
14. Zhao Y, Blencowe M, Shi X, Shu L, Levian C, Ahn IS, et al. Integrative Genomics Analysis Unravels Tissue-Specific Pathways, Networks, and Key Regulators of Blood Pressure Regulation. Frontiers in Cardiovascular Medicine. 2019;6:21.
15. Leiserson MDM, Eldridge JV, Ramachandran S, Raphael BJ. Network analysis of GWAS data. Curr Opin Genet Dev. 2013;23(6):602-10.

16. Verbanck M, Chen C-Y, Neale B, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. Nature genetics. 2018;50(5):693-8.
17. Chesmore K, Bartlett J, Williams SM. The ubiquity of pleiotropy in human disease. Human Genetics. 2018;137(1):39-44.
18. Deo RC, Musso G, Tasan M, Tang P, Poon A, Yuan C, et al. Prioritizing causal disease genes using unbiased genomic features. Genome Biol. 2014;15(12):534.
19. Kolosov N, Daly MJ, Artomov M. Prioritization of disease genes from GWAS using ensemble-based positive-unlabeled learning. European Journal of Human Genetics. 2021.
20. Mishra Manoj K, Liang Eugene Y, Geurts Aron M, Auer Paul WL, Liu P, Rao S, et al. Comparative and Functional Genomic Resource for Mechanistic Studies of Human Blood Pressure–Associated Single Nucleotide Polymorphisms. Hypertension. 2020;75(3):859-68.
21. Oguz C, Sen SK, Davis AR, Fu Y-P, O'Donnell CJ, Gibbons GH. Machine learning identifies SNPs predictive of advanced coronary artery calcium in ClinSeq® and Framingham Heart Study cohorts. bioRxiv. 2017:102350.
22. Huang K, Nogueira R. EpiRL: A Reinforcement Learning Agent to Facilitate Epistasis Detection. In: Shaban-Nejad A, Michalowski M, editors. Precision Health and Medicine: A Digital Revolution in Healthcare. Cham: Springer International Publishing; 2020. p. 187-91.
23. Wang H, Liu X, Tao Y, Ye W, Jin Q, Cohen WW, et al. Automatic Human-like Mining and Constructing Reliable Genetic Association Database with Deep Reinforcement Learning. Biocomputing 2019: WORLD SCIENTIFIC; 2018. p. 112-23.
24. Nicholls HL, John CR, Watson DS, Munroe PB, Barnes MR, Cabrera CP. Reaching the End-Game for GWAS: Machine Learning Approaches for the Prioritization of Complex Disease Loci. Frontiers in Genetics. 2020;11(350).
25. Eid F-E, Elmarakeby HA, Chan YA, Fornelos N, ElHefnawi M, Van Allen EM, et al. Systematic auditing is essential to debiasing machine learning in biology. Communications Biology. 2021;4(1):183.
26. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems; Long Beach, California, USA: Curran Associates Inc.; 2017. p. 4768–77.
27. Dickinson Q, Meyer JG. Positional SHAP (PoSHAP) for Interpretation of machine learning models trained from biological sequences. PLOS Computational Biology. 2022;18(1):e1009736.
28. Seyyedrazzagi E, Navimipour NJ. Disease genes prioritizing mechanisms: a comprehensive and systematic literature review. Network Modeling and Analysis in Health Informatics and Bioinformatics. 2017;6(1).
29. Raj MR, Sreeja A. Analysis of Computational Gene Prioritization Approaches. 8th International Conference on Advances in Computing & Communications (Icacc-2018). 2018;143:395-410.
30. Mieth B, Kloft M, Rodriguez JA, Sonnenburg S, Vobruba R, Morcillo-Suarez C, et al. Combining Multiple Hypothesis Testing with Machine Learning Increases the Statistical Power of Genome-wide Association Studies. Sci Rep. 2016;6(1):36671.

31. Leem S, Jeong HH, Lee J, Wee K, Sohn KA. Fast detection of high-order epistatic interactions in genome-wide association studies using information theoretic measure. Comput Biol Chem. 2014;50:19-28.
32. Pare G, Mao S, Deng WQ. A machine-learning heuristic to improve gene score prediction of polygenic traits. Sci Rep. 2017;7(1):12665.
33. Vitsios D, Petrovski S. Mantis-ml: Disease-Agnostic Gene Prioritization from High-Throughput Genomic Screens by Stochastic Semi-supervised Learning. The American Journal of Human Genetics. 2020;106(5):659-78.
34. Lewinger JP, Conti DV, Baurley JW, Triche TJ, Thomas DC. Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. Genet Epidemiol. 2007;31(8):871-82.
35. Fridley BL, Iversen E, Tsai YY, Jenkins GD, Goode EL, Sellers TA. A latent model for prioritization of SNPs for functional studies. PLoS One. 2011;6(6):e20764.
36. Wang Y, Goh W, Wong L, Montana G, Alzheimer's Disease Neuroimaging I. Random forests on Hadoop for genome-wide association studies of multivariate neuroimaging phenotypes. BMC Bioinformatics. 2013;14 Suppl 16(Suppl 16):S6.
37. Oh JH, Kerns S, Ostrer H, Powell SN, Rosenstein B, Deasy JO. Computational methods using genome-wide association studies to predict radiotherapy complications and to identify correlative molecular processes. Sci Rep. 2017;7(1):43381.
38. Khan A, Liu Q, Wang K. iMEGES: integrated mental-disorder GEnome score by deep neural network for prioritizing the susceptibility genes for mental disorders in personal genomes. BMC Bioinformatics. 2018;19(17):501.
39. Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nat Genet. 2018;50(8):1171-9.
40. Sanseau P, Agarwal P, Barnes MR, Pastinen T, Richards JB, Cardon LR, et al. Use of genome-wide association studies for drug repositioning. Nat Biotechnol. 2012;30(4):317-20.
41. Wang K, Wan M, Wang R-S, Weng Z. Opportunities for Web-based Drug Repositioning: Searching for Potential Antihypertensive Agents with Hypotension Adverse Events. J Med Internet Res. 2016;18(4):e76-e.
42. Jarari N, Rao N, Peela JR, Ellafi KA, Shakila S, Said AR, et al. A review on prescribing patterns of antihypertensive drugs. Clin Hypertens. 2016;22:7-.
43. Olczak KJ, Taylor-Bateman V, Nicholls HL, Traylor M, Cabrera CP, Munroe PB. Hypertension genetics past, present and future applications. Journal of Internal Medicine. 2021;n/a(n/a).
44. Ramdas S, Judd J, Graham SE, Kanoni S, Wang Y, Surakka I, et al. A multi-layer functional genomic analysis to understand noncoding genetic variation in lipids. bioRxiv. 2021:2021.12.07.470215.
45. Kanoni S, Graham SE, Wang Y, Surakka I, Ramdas S, Zhu X, et al. Implicating genes, pleiotropy and sexual dimorphism at blood lipid loci through multi-ancestry meta-analysis. medRxiv. 2021:2021.12.15.21267852.
46. Khan A, Liu Q, Wang K. iMEGES: integrated mental-disorder GEnome score by deep neural network for prioritizing the susceptibility genes for mental disorders in personal genomes. BMC Bioinformatics. 2018;19(Suppl 17):501.

47. Evangelou E, Warren HR, Mosen-Ansorena D, Mifsud B, Pazoki R, Gao H, et al. Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. Nat Genet. 2018;50(10):1412-25.
48. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. (1362-4962 (Electronic)).
49. Gettler K, Giri M, Kenigsberg E, Martin J, Chuang LS, Hsu NY, et al. Prioritizing Crohn's disease genes by integrating association signals with gene expression implicates monocyte subsets. Genes Immun. 2019;20(7):577-88.
50. Isakov O, Dotan I, Ben-Shachar S. Machine Learning-Based Gene Prioritization Identifies Novel Candidate Risk Genes for Inflammatory Bowel Disease. Inflamm Bowel Dis. 2017;23(9):1516-23.
51. Dai Y, Pei G, Zhao Z, Jia P. A Convergent Study of Genetic Variants Associated With Crohn's Disease: Evidence From GWAS, Gene Expression, Methylation, eQTL and TWAS. Front Genet. 2019;10:318.
52. Branco PR, de Araujo GS, Barrera J, Suarez-Kurtz G, de Souza SJ. Uncovering association networks through an eQTL analysis involving human miRNAs and lincRNAs. Sci Rep. 2018;8(1):15050.
53. Consortium GT. The Genotype-Tissue Expression (GTEx) project. Nature genetics. 2013;45(6):580-5.
54. Rancati G, Moffat J, Typas A, Pavelka N. Emerging and evolving concepts in gene essentiality. Nature Reviews Genetics. 2018;19(1):34-49.
55. Jia K, Zhou Y, Cui Q. Quantifying Gene Essentiality Based on the Context of Cellular Components. Frontiers in genetics. 2020;10:1342-.
56. Minikel EV, Karczewski KJ, Martin HC, Cummings BB, Whiffin N, Rhodes D, et al. Evaluating drug targets through human loss-of-function genetic variation. Nature. 2020;581(7809):459-64.
57. Fadista J, Oskolkov N, Hansson O, Groop L. LoFtool: a gene intolerance score based on loss-of-function variants in 60 706 individuals. Bioinformatics. 2017;33(4):471-4.
58. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285-91.
59. Whiffin N, Karczewski KJ, Zhang X, Chothani S, Smith MJ, Evans DG, et al. Characterising the loss-of-function impact of 5' untranslated region variants in 15,708 individuals. Nature Communications. 2020;11(1):2523.
60. Finer S, Martin HC, Khan A, Hunt KA, MacLaughlin B, Ahmed Z, et al. Cohort Profile: East London Genes & Health (ELGH), a community-based population genomics and health study in British Bangladeshi and British Pakistani people. International Journal of Epidemiology. 2019;49(1):20-1i.
61. Cacheiro P, Muñoz-Fuentes V, Murray SA, Dickinson ME, Bucan M, Nutter LMJ, et al. Human and mouse essentiality screens as a resource for disease gene discovery. Nature Communications. 2020;11(1):655.
62. Steinberg J, Honti F, Meader S, Webber C. Haploinsufficiency predictions without study bias. Nucleic Acids Research. 2015;43(15):e101-e.
63. Shihab HA, Rogers MF, Campbell C, Gaunt TR. HIPred: an integrative approach to predicting haploinsufficient genes. Bioinformatics. 2017;33(12):1751-7.

64. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Research. 2017;45(D1):D833-D9.
65. Yang C, Li J, Wu Q, Yang X, Huang AY, Zhang J, et al. AutismKB 2.0: a knowledgebase for the genetic evidence of autism spectrum disorder. Database. 2018;2018.
66. V A, Nayar PG, Murugesan R, Mary B, P D, Ahmed SSSJ. CardioGenBase: A Literature Based Multi-Omics Database for Major Cardiovascular Diseases. PLoS One. 2015;10(12):e0143188-e.
67. Smedley D, Jacobsen JOB, Jäger M, Köhler S, Holtgrewe M, Schubach M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. Nature Protocols. 2015;10(12):2004-15.
68. Ritchie GRS FP. Functional Annotation of Rare Genetic Variants NCBI2015 [Available from: https://www.ncbi.nlm.nih.gov/books/NBK539450/.
69. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285-91.
70. Ward LD, Kellis M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. Nucleic Acids Research. 2016;44(D1):D877-D81.
71. Mossotto E, Ashton JJ, O'Gorman L, Pengelly RJ, Beattie RM, MacArthur BD, et al. GenePy - a score for estimating gene pathogenicity in individuals using next-generation sequencing data. BMC Bioinformatics. 2019;20(1):254.
72. Maciukiewicz M, Marshe VS, Hauschild AC, Foster JA, Rotzinger S, Kennedy JL, et al. GWAS-based machine learning approach to predict duloxetine response in major depressive disorder. J Psychiatr Res. 2018;99:62-8.
73. Mentch GHaL. Please Stop Permuting Features: An Explanation and Alternatives. arXiv. 2019.
74. Keany E. BorutaShap : A wrapper feature selection method which combines the Boruta feature selection algorithm with Shapley values. 2020.
75. Miller JE, Veturi Y, Ritchie MD. Innovative strategies for annotating the "relationSNP" between variants and molecular phenotypes. BioData Mining. 2019;12(1):10.
76. Mills MC, Rahal C. The GWAS Diversity Monitor tracks diversity by disease in real time. Nature Genetics. 2020;52(3):242-3.
77. Oliva M, Muñoz-Aguirre M, Kim-Hellmuth S, Wucher V, Gewirtz ADH, Cotter DJ, et al. The impact of sex on gene expression across human tissues. Science. 2020;369(6509):eaba3066.
78. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3(null):1157–82.
79. Darst BF, Malecki KC, Engelman CD. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. BMC Genetics. 2018;19(1):65.
80. Kotu V, Deshpande B. Chapter 13 - Anomaly Detection. In: Kotu V, Deshpande B, editors. Data Science (Second Edition): Morgan Kaufmann; 2019. p. 447-65.

81. Nembrini S, Konig IR, Wright MN. The revival of the Gini importance? Bioinformatics. 2018;34(21):3711-8.
82. Lopes I, Altab G, Raina P, de Magalhães JP. Gene Size Matters: An Analysis of Gene Length in the Human Genome. Frontiers in Genetics. 2021;12:30.
83. Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, et al. Common Sequence Polymorphisms Shaping Genetic Diversity in <em>Arabidopsis thaliana</em&gt. Science. 2007;317(5836):338.
84. Mountjoy E, Schmidt EM, Carmona M, Peat G, Miranda A, Fumis L, et al. Open Targets Genetics: An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. bioRxiv. 2020:2020.09.16.299271.
85. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic acids research. 2019;47(D1):D607-D13.
86. Schaefer MH, Serrano L, Andrade-Navarro MA. Correcting for the study bias associated with protein–protein interaction measurements reveals differences between protein degree distributions from different cancer types. Frontiers in Genetics. 2015;6(260).
87. Hoffmann TJ, Ehret GB, Nandakumar P, Ranatunga D, Schaefer C, Kwok P-Y, et al. Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. Nature genetics. 2017;49(1):54-64.
88. Sung YJ, de las Fuentes L, Winkler TW, Chasman DI, Bentley AR, Kraja AT, et al. A multi-ancestry genome-wide study incorporating gene–smoking interactions identifies multiple new loci for pulse pressure and mean arterial pressure. Human Molecular Genetics. 2019;28(15):2615-33.
89. de las Fuentes L, Sung YJ, Noordam R, Winkler T, Feitosa MF, Schwander K, et al. Gene-educational attainment interactions in a multi-ancestry genome-wide meta-analysis identify novel blood pressure loci. Molecular Psychiatry. 2021;26(6):2111-25.
90. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research. 2010;38(16):e164-e.
91. Fontaine J-F, Priller F, Barbosa-Silva A, Andrade-Navarro MA. Génie: literature-based gene prioritization at multi genomic scale. Nucleic Acids Research. 2011;39(suppl_2):W455-W61.
92. Mostafavi H, Spence JP, Naqvi S, Pritchard JK. Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery. bioRxiv. 2022:2022.05.07.491045.
93. Franzén O, Gan L-M, Björkegren JLM. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database. 2019;2019:baz046.
94. Papatheodorou I, Moreno P, Manning J, Fuentes AM-P, George N, Fexova S, et al. Expression Atlas update: from tissues to single cells. Nucleic Acids Research. 2020;48(D1):D77-D83.

95. Osumi-Sutherland D, Xu C, Keays M, Levine AP, Kharchenko PV, Regev A, et al. Cell type ontologies of the Human Cell Atlas. Nature Cell Biology. 2021;23(11):1129-35.
96. Petrazzini BO, Naya H, Lopez-Bello F, Vazquez G, Spangenberg L. Evaluation of different approaches for missing data imputation on features associated to genomic data. BioData Min. 2021;14(1):44.
97. Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. Genetic effects on gene expression across human tissues. Nature. 2017;550(7675):204-13.
98. Darst BF, Malecki KC, Engelman CD. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. BMC Genet. 2018;19(Suppl 1):65.
99. Demir-Kavuk O, Kamada M, Akutsu T, Knapp EW. Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features. BMC Bioinformatics. 2011;12(1):412.
100. Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society Series B (Statistical Methodology). 2005;67(2):301-20.
101. Ogutu JO, Schulz-Streeck T, Piepho HP. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. BMC Proc. 2012;6 Suppl 2(S2):S10.
102. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot. 2013;7:21.
103. Dietterich TG. Ensemble methods in machine learning. Multiple Classifier Systems. 2000;1857:1-15.
104. Breiman L. Random forests. Machine Learning. 2001;45(1):5-32.
105. Kafaie S, Chen Y, Hu T. A network approach to prioritizing susceptibility genes for genome-wide association studies. Genet Epidemiol. 2019;43(5):477-91.
106. Vitsios D, Petrovski S. Stochastic semi-supervised learning to prioritise genes from high-throughput genomic screens. bioRxiv. 2019.
107. Roshan U, Chikkagoudar S, Wei Z, Wang K, Hakonarson H. Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest. Nucleic Acids Res. 2011;39(9):e62.
108. Smola AJ, Scholkopf B. A tutorial on support vector regression. Statistics and Computing. 2004;14(3):199-222.
109. Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E. Deep learning applications and challenges in big data analytics. Journal of Big Data. 2015;2(1):1.
110. Aung N, Vargas JD, Yang C, Cabrera CP, Warren HR, Fung K, et al. Genome-Wide Analysis of Left Ventricular Image-Derived Phenotypes Identifies Fourteen Loci Associated With Cardiac Morphogenesis and Heart Failure Development. Circulation. 2019;140(16):1318-30.
111. Hampe N, Wolterink JM, van Velzen SGM, Leiner T, Išgum I. Machine Learning for Assessment of Coronary Artery Disease in Cardiac CT: A Survey. Frontiers in Cardiovascular Medicine. 2019;6.
112. Zeng W, Wu M, Jiang R. Prediction of enhancer-promoter interactions via natural language processing. BMC Genomics. 2018;19(Suppl 2):84.

113. Bao F, Deng Y, Du M, Ren Z, Wan S, Liang KY, et al. Explaining the Genetic Causality for Complex Phenotype via Deep Association Kernel Learning. Patterns (N Y). 2020;1(6):100057.
114. Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, et al. The support of human genetic evidence for approved drug indications. Nature Genetics. 2015;47(8):856-60.
115. Ghoussaini M, Mountjoy E, Carmona M, Peat G, Schmidt Ellen M, Hercules A, et al. Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. Nucleic Acids Research. 2021;49(D1):D1311-D20.
116. Xie Z, Bailey A, Kuleshov MV, Clarke DJB, Evangelista JE, Jenkins SL, et al. Gene Set Knowledge Discovery with Enrichr. Current Protocols. 2021;1(3):e90.
117. Cattin ME, Wang J, Weldrick JJ, Roeske CL, Mak E, Thorn SL, et al. Deletion of MLIP (muscle-enriched A-type lamin-interacting protein) leads to cardiac hyperactivation of Akt/mammalian target of rapamycin (mTOR) and impaired cardiac adaptation. J Biol Chem. 2015;290(44):26699-714.
118. van Oort RJ, Garbino A, Wang W, Dixit SS, Landstrom AP, Gaur N, et al. Disrupted junctional membrane complexes and hyperactive ryanodine receptors after acute junctophilin knockdown in mice. Circulation. 2011;123(9):979-88.
119. Peters N, Opherk C, Zacherle S, Capell A, Gempel P, Dichgans M. CADASIL-associated Notch3 mutations have differential effects both on ligand binding and ligand-induced Notch3 receptor signaling through RBP-Jk. Exp Cell Res. 2004;299(2):454-64.
120. Boulos N, Helle F, Dussaule J-C, Placier S, Milliez P, Djudjaj S, et al. Notch3 Is Essential for Regulation of the Renal Vascular Tone. Hypertension. 2011;57(6):1176-82.
121. Ma L, Wang H, Sun Y, Yang D, Pu L, Zhang X. P53-induced MRVI1 mediates carcinogenesis of colorectal cancer. Scandinavian Journal of Gastroenterology. 2020;55(7):824-33.
122. Arber S, Hunter JJ, Ross J, Jr., Hongo M, Sansig G, Borg J, et al. MLP-Deficient Mice Exhibit a Disruption of Cardiac Cytoarchitectural Organization, Dilated Cardiomyopathy, and Heart Failure. Cell. 1997;88(3):393-403.
123. Shelton DN, Fornalik H, Neff T, Park SY, Bender D, DeGeest K, et al. The Role of LEF1 in Endometrial Gland Formation and Carcinogenesis. PLoS One. 2012;7(7):e40312.
124. Xia W, Liu Y, Jiao J. GRM7 Regulates Embryonic Neurogenesis via CREB and YAP. Stem Cell Reports. 2015;4(5):795-810.
125. Li J, Meng H, Cao W, Qiu T. MiR-335 is involved in major depression disorder and antidepressant treatment through targeting GRM4. Neuroscience Letters. 2015;606:167-72.
126. Liguori L, Andolfo I, De Antonellis P, Aglio V, di Dato V, Marino N, et al. The metallophosphodiesterase Mpped2 impairs tumorigenesis in neuroblastoma. Cell Cycle. 2012;11(3):569-81.
127. Dunnmon JA, Ratner AJ, Saab K, Khandwala N, Markert M, Sagreiya H, et al. Cross-Modal Data Programming Enables Rapid Medical Machine Learning. Patterns (N Y). 2020;1(2).

128. Santos MS, Soares JP, Abreu PH, Araujo H, Santos J. Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches [Research Frontier]. IEEE Computational Intelligence Magazine. 2018;13(4):59-76.
129. Borisov V, Leemann T, Seßler K, Haug J, Pawelczyk M, Kasneci G. Deep Neural Networks and Tabular Data: A Survey2021 October 01, 2021:[arXiv:2110.01889 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2021arXiv211001889B.
130. Lai C-H, Pandey S, Day CH, Ho T-J, Chen R-J, Chang R-L, et al. β-catenin/LEF1/IGF-IIR Signaling Axis Galvanizes the Angiotensin-II- induced Cardiac Hypertrophy. International Journal of Molecular Sciences. 2019;20(17).
131. Xu W, He H, Guo Z, Li W. Evaluation of machine learning models on protein level inference from prioritized RNA features. Briefings in Bioinformatics. 2022;23(3):bbac091.
132. Hensman J, Lawrence ND, Rattray M. Hierarchical Bayesian modelling of gene expression time series across irregularly sampled replicates and clusters. BMC Bioinformatics. 2013;14(1):252.
133. Qiao X. Learning ordinal data. WIREs Computational Statistics. 2015;7(5):341-6.
134. Alam T, Ahmed CF, Zahin SA, Khan MAH, Islam MT. An Effective Recursive Technique for Multi-Class Classification and Regression for Imbalanced Data. IEEE Access. 2019;7:127615-30.
135. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res. 2016;44(D1):D1075-9.
136. Freshour SL, Kiwala S, Cotto KC, Coffman AC, McMichael JF, Song JJ, et al. Integration of the Drug–Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts. Nucleic Acids Research. 2021;49(D1):D1144-D51.
137. Mountjoy E, Schmidt EM, Carmona M, Schwartzentruber J, Peat G, Miranda A, et al. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. Nature genetics. 2021;53(11):1527-33.
138. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Research. 2009;37(suppl_2):W305-W11.
139. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. PLOS Computational Biology. 2015;11(4):e1004219.
140. Li A, Liu S, Bakshi A, Jiang L, Chen W, Zheng Z, et al. mBAT-combo: a more powerful test to detect gene-trait associations from GWAS data. bioRxiv. 2022:2022.06.27.497850.
141. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nature Communications. 2017;8(1):1826.
142. Dickinson ME, Flenniken AM, Ji X, Teboul L, Wong MD, White JK, et al. High-throughput discovery of novel developmental phenotypes. Nature. 2016;537(7621):508-14.
143. Sabharwal R, Yang L, Zimmerman K, Weiss R. Protective actions of angiotensin II type 2 receptors in male mice with muscular dystrophy. The FASEB Journal. 2019;33(S1):746.1-.1.

144. Rasi K, Piuhola J, Czabanka M, Sormunen R, Ilves M, Leskinen H, et al. Collagen XV Is Necessary for Modeling of the Extracellular Matrix and Its Deficiency Predisposes to Cardiomyopathy. Circulation Research. 2010;107(10):1241-52.
145. Li X, Zhang X, Leathers R, Makino A, Huang C, Parsa P, et al. Notch3 signaling promotes the development of pulmonary arterial hypertension. Nature Medicine. 2009;15(11):1289-97.
146. Durgin BG, Cherepanova OA, Gomez D, Karaoli T, Alencar GF, Butcher JT, et al. Smooth muscle cell-specific deletion of Col15a1 unexpectedly leads to impaired development of advanced atherosclerotic lesions. American Journal of Physiology-Heart and Circulatory Physiology. 2017;312(5):H943-H58.
147. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nature Genetics. 2013;45(6):580-5.
148. Lechner SG, Markworth S, Poole K, Smith ESJ, Lapatsina L, Frahm S, et al. The Molecular and Cellular Identity of Peripheral Osmoreceptors. Neuron. 2011;69(2):332-44.
149. Marvar PJ, Harrison DG. Stress-dependent hypertension and the role of T lymphocytes. Experimental Physiology. 2012;97(11):1161-7.
150. Cui C, Fan J, Zeng Q, Cai J, Chen Y, Chen Z, et al. CD4+ T-Cell Endogenous Cystathionine γ Lyase–Hydrogen Sulfide Attenuates Hypertension by Sulfhydrating Liver Kinase B1 to Promote T Regulatory Cell Differentiation and Proliferation. Circulation. 2020;142(18):1752-69.
151. Ikawa T, Watanabe Y, Okuzaki D, Goto N, Okamura N, Yamanishi K, et al. A new approach to identifying hypertension-associated genes in the mesenteric artery of spontaneously hypertensive rats and stroke-prone spontaneously hypertensive rats. Journal of hypertension. 2019;37(8):1644-56.
152. Chen Y-H, Layne MD, Chung SW, Ejima K, Baron RM, Yet S-F, et al. Elk-3 Is a Transcriptional Repressor of Nitric-oxide Synthase 2 *. Journal of Biological Chemistry. 2003;278(41):39572-7.
153. Xin C, Lei J, Wang Q, Yin Y, Yang X, Moran Guerrero JA, et al. Therapeutic silencing of SMOC2 prevents kidney function loss in mouse model of chronic kidney disease. iScience. 2021;24(10):103193.
154. Gong Y, McDonough CW, Beitelshees AL, El Rouby N, Hiltunen TP, O'Connell JR, et al. PTPRD gene associated with blood pressure response to atenolol and resistant hypertension. J Hypertens. 2015;33(11):2278-85.
155. Padmanabhan S, Dominiczak AF. Genomics of hypertension: the road to precision medicine. Nature Reviews Cardiology. 2021;18(4):235-50.
156. Ikawa T, Watanabe Y, Okuzaki D, Goto N, Okamura N, Yamanishi K, et al. A new approach to identifying hypertension-associated genes in the mesenteric artery of spontaneously hypertensive rats and stroke-prone spontaneously hypertensive rats. J Hypertens. 2019;37(8):1644-56.
157. Tonstad S, Tykarski A, Weissgarten J, Ivleva A, Levy B, Kumar A, et al. Efficacy and Safety of Topiramate in the Treatment of Obese Subjects With Essential Hypertension. The American Journal of Cardiology. 2005;96(2):243-51.
158. Li QS, Lenhard JM, Zhan Y, Konvicka K, Athanasiou MC, Strauss RS, et al. A candidate-gene association study of topiramate-induced weight loss in obese patients with and without type 2 diabetes mellitus. Pharmacogenetics and Genomics. 2016;26(2):53-65.

159. Zu HL, Liu HW, Wang HY. Identification of crucial genes involved in pathogenesis of regional weakening of the aortic wall. Hereditas. 2021;158(1):35.
160. Saygin D, Tabib T, Bittar HET, Valenzi E, Sembrat J, Chan SY, et al. Transcriptional profiling of lung cell populations in idiopathic pulmonary arterial hypertension. Pulmonary Circulation. 2020;10(1):???
161. Damkjær M, Isaksson GL, Stubbe J, Jensen BL, Assersen K, Bie P. Renal renin secretion as regulator of body fluid homeostasis. Pflugers Arch. 2013;465(1):153-65.
162. Chamorro-Jorganes A, Grande MT, Herranz B, Jerkic M, Griera M, Gonzalez-Nuñez M, et al. Targeted genomic disruption of h-ras induces hypotension through a NO-cGMP-PKG pathway-dependent mechanism. Hypertension. 2010;56(3):484-9.
163. Blaustein MP, Zhang J, Chen L, Song H, Raina H, Kinsey SP, et al. The pump, the exchanger, and endogenous ouabain: signaling mechanisms that link salt retention to hypertension. Hypertension. 2009;53(2):291-8.
164. Liu K, Liu Z, Qi H, Liu B, Wu J, Liu Y, et al. Genetic Variation in SLC8A1 Gene Involved in Blood Pressure Responses to Acute Salt Loading. American Journal of Hypertension. 2017;31(4):415-21.
165. Whalen S, Schreiber J, Noble WS, Pollard KS. Navigating the pitfalls of applying machine learning in genomics. Nature Reviews Genetics. 2021.
166. Whalen S, Schreiber J, Noble WS, Pollard KS. Navigating the pitfalls of applying machine learning in genomics. Nature Reviews Genetics. 2022;23(3):169-81.
167. Tivesten As, Bollano E, Andersson I, Fitzgerald S, Caidahl K, Sjögren K, et al. Liver-Derived Insulin-Like Growth Factor-I Is Involved in the Regulation of Blood Pressure in Mice. Endocrinology. 2002;143(11):4235-42.
168. Olofsson PS, Steinberg BE, Sobbi R, Cox MA, Ahmed MN, Oswald M, et al. Blood pressure regulation by CD4+ lymphocytes expressing choline acetyltransferase. Nature Biotechnology. 2016;34(10):1066-71.
169. Oliva M, Muñoz-Aguirre M, Kim-Hellmuth S, Wucher V, Gewirtz ADH, Cotter DJ, et al. The impact of sex on gene expression across human tissues. Science. 2020;369(6509).
170. Segarra AB, Prieto I, Villarejo AB, Banegas I, Wangensteen R, de Gasparo M, et al. Effects of antihypertensive drugs on angiotensinase activities in the testis of spontaneously hypertensive rats. Horm Metab Res. 2013;45(5):344-8.
171. Pluznick JL, Protzko RJ, Gevorgyan H, Peterlin Z, Sipos A, Han J, et al. Olfactory receptor responding to gut microbiota-derived signals plays a role in renin secretion and blood pressure regulation. Proceedings of the National Academy of Sciences. 2013;110(11):4410-5.
172. Smith LN. A disciplined approach to neural network hyper-parameters: Part 1--learning rate, batch size, momentum, and weight decay. arXiv preprint arXiv:180309820. 2018.
173. de Smet MD, Castilla M. Ocriplasmin for diabetic retinopathy. Expert Opinion on Biological Therapy. 2013;13(12):1741-7.
174. Zhai S, Zhang H, Mehrotra DV, Shen J. Pharmacogenomics polygenic risk score for drug response prediction using PRS-PGx methods. Nature Communications. 2022;13(1):5278.

175. Oxenkrug GF, Summergrad P. Ramelteon attenuates age-associated hypertension and weight gain in spontaneously hypertensive rats. Ann N Y Acad Sci. 2010;1199:114-20.
176. Bamborough P, Drewry D, Harper G, Smith GK, Schneider K. Assessment of chemical coverage of kinome space and its implications for kinase drug discovery. J Med Chem. 2008;51(24):7898-914.
177. Roohbakhsh A, Parhiz H, Soltani F, Rezaee R, Iranshahi M. Molecular mechanisms behind the biological effects of hesperidin and hesperetin for the prevention of cancer and cardiovascular diseases. Life Sci. 2015;124:64-74.
178. Yamamoto M, Suzuki A, Jokura H, Yamamoto N, Hase T. Glucosyl hesperidin prevents endothelial dysfunction and oxidative stress in spontaneously hypertensive rats. Nutrition. 2008;24(5):470-6.
179. Forte M, Bianchi F, Cotugno M, Marchitti S, De Falco E, Raffa S, et al. Pharmacological restoration of autophagy reduces hypertension-related stroke occurrence. Autophagy. 2020;16(8):1468-81.
180. Lin X, Han T, Fan Y, Wu S, Wang F, Wang C. Quercetin improves vascular endothelial function through promotion of autophagy in hypertensive rats. Life Sci. 2020;258:118106.
181. Chrysanthopoulou A, Gkaliagkousi E, Lazaridis A, Arelaki S, Pateinakis P, Ntinopoulou M, et al. Angiotensin II triggers release of neutrophil extracellular traps, linking thromboinflammation with essential hypertension. JCI Insight. 2021;6(18).
182. Marrachelli VG, Mastronardi ML, Sarr M, Soleti R, Leonetti D, Martínez MC, et al. Sonic hedgehog carried by microparticles corrects angiotensin II-induced hypertension and endothelial dysfunction in mice. PLoS One. 2013;8(8):e72861.
183. Kanoni S, Graham SE, Wang Y, Surakka I, Ramdas S, Zhu X, et al. Implicating genes, pleiotropy and sexual dimorphism at blood lipid loci through multi-ancestry meta-analysis. medRxiv. 2021:2021.12.15.21267852.
184. Datar J, Regassa A, Kim W-K, Taylor CG, Zahradka P, Suh M. Lipid Metabolism is Closely Associated with Normal Testicular Growth Based on Global Transcriptome Profiles in Normal and Underdeveloped Testis of Obese Zucker (fa/fa) Rats. Lipids. 2017;52(11):951-60.
185. Varlamov O, Bethea CL, Roberts CT. Sex-Specific Differences in Lipid and Glucose Metabolism. Frontiers in Endocrinology. 2015;5.
186. Grummer RR, Carroll DJ. A Review of Lipoprotein Cholesterol Metabolism: Importance to Ovarian Function. Journal of Animal Science. 1988;66(12):3160-73.
187. Miller CO, Yang X, Lu K, Cao J, Herath K, Rosahl TW, et al. Ketohexokinase knockout mice, a model for essential fructosuria, exhibit altered fructose metabolism and are protected from diet-induced metabolic defects. American Journal of Physiology-Endocrinology and Metabolism. 2018;315(3):E386-E93.
188. Dron JS, Dilliott AA, Lawson A, McIntyre AD, Davis BD, Wang J, et al. Loss-of-Function CREB3L3 Variants in Patients With Severe Hypertriglyceridemia. Arteriosclerosis, Thrombosis, and Vascular Biology. 2020;40(8):1935-41.
189. Blücher C, Iberl S, Schwagarus N, Müller S, Liebisch G, Höring M, et al. Secreted Factors from Adipose Tissue Reprogram Tumor Lipid Metabolism and Induce Motility by Modulating PPARα/ANGPTL4 and FAK. Molecular Cancer Research. 2020;18(12):1849-62.

190. Hoekstra R, de Vos FY, Eskens FA, Gietema JA, van der Gaast A, Groen HJ, et al. Phase I safety, pharmacokinetic, and pharmacodynamic study of the thrombospondin-1-mimetic angiogenesis inhibitor ABT-510 in patients with advanced cancer. J Clin Oncol. 2005;23(22):5188-97.
191. Febbraio M, Hajjar DP, Silverstein RL. CD36: a class B scavenger receptor involved in angiogenesis, atherosclerosis, inflammation, and lipid metabolism. The Journal of Clinical Investigation. 2001;108(6):785-91.
192. Ramdas S, Judd J, Graham SE, Kanoni S, Wang Y, Surakka I, et al. A multi-layer functional genomic analysis to understand noncoding genetic variation in lipids. bioRxiv. 2021:2021.12.07.470215.
193. Lee MSA, Singh J. The Landscape and Gaps in Open Source Fairness Toolkits. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; Yokohama, Japan: Association for Computing Machinery; 2021. p. Article 699.
194. Huang K, Nogueira R. EpiRL: A Reinforcement Learning Agent to Facilitate Epistasis Detection2018.
195. Wang QS, Kelley DR, Ulirsch J, Kanai M, Sadhuka S, Cui R, et al. Leveraging supervised learning for functionally informed fine-mapping of cis-eQTLs identifies an additional 20,913 putative causal eQTLs. Nature Communications. 2021;12(1):3394.
196. Hemani G, Bowden J, Haycock P, Zheng J, Davis O, Flach P, et al. Automating Mendelian randomization through machine learning to construct a putative causal map of the human phenome. bioRxiv. 2017:173682.
197. Graham SE, Clarke SL, Wu K-HH, Kanoni S, Zajac GJM, Ramdas S, et al. The power of genetic diversity in genome-wide association studies of lipids. Nature. 2021;600(7890):675-9.
198. Castelli WP, Anderson K, Wilson PW, Levy D. Lipids and risk of coronary heart disease. The Framingham Study. Ann Epidemiol. 1992;2(1-2):23-8.
199. Kathiresan S, Manning AK, Demissie S, D'Agostino RB, Surti A, Guiducci C, et al. A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. BMC Medical Genetics. 2007;8(1):S17.
200. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26(17):2190-1.
201. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28(1):112-8.
202. Koscielny G, An P, Carvalho-Silva D, Cham JA, Fumis L, Gasparyan R, et al. Open Targets: a platform for therapeutic target identification and validation. Nucleic Acids Res. 2017;45(D1):D985-d94.
203. Lopes I, Altab G, Raina P, de Magalhães JP. Gene Size Matters: An Analysis of Gene Length in the Human Genome. Frontiers in Genetics. 2021;12.
204. Klarin D, Damrauer SM, Cho K, Sun YV, Teslovich TM, Honerlaw J, et al. Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. Nature Genetics. 2018;50(11):1514-23.
205. Soto M, Cai W, Konishi M, Kahn CR. Insulin signaling in the hippocampus and amygdala regulates metabolism and neurobehavior. Proceedings of the National Academy of Sciences. 2019;116(13):6379-84.
206. Sèdes L, Thirouard L, Maqdasy S, Garcia M, Caira F, Lobaccaro J-MA, et al. Cholesterol: A Gatekeeper of Male Fertility? Frontiers in Endocrinology. 2018;9.

207. Stocco DM, Wang X, Jo Y, Manna PR. Multiple Signaling Pathways Regulating Steroidogenesis and Steroidogenic Acute Regulatory Protein Expression: More Complicated than We Thought. Molecular Endocrinology. 2005;19(11):2647-59.
208. Korou L-M, Agrogiannis G, Koros C, Kitraki E, Vlachos IS, Tzanetakou I, et al. Impact of N-acetylcysteine and sesame oil on lipid metabolism and hypothalamic-pituitary-adrenal axis homeostasis in middle-aged hypercholesterolemic mice. Scientific Reports. 2014;4(1):6806.
209. Sun H, Wang N, Nie X, Zhao L, Li Q, Cang Z, et al. Lead Exposure Induces Weight Gain in Adult Rats, Accompanied by DNA Hypermethylation. PLOS ONE. 2017;12(1):e0169958.
210. Jurgens J, Chen S, Sobreira N, Robbins S, Anzmann AF, Dastgheyb R, et al. Loss of function variants in PCYT1A causing spondylometaphyseal dysplasia with cone/rod dystrophy have broad consequences on lipid metabolism, chondrocyte differentiation, and lipid droplet formation. bioRxiv. 2019:2019.12.19.882191.
211. Amirifar P, Ranjouri MR, Abolhassani H, Moeini Shad T, Almasi-Hashiani A, Azizi G, et al. Clinical, immunological and genetic findings in patients with UNC13D deficiency (FHL3): A systematic review. Pediatric Allergy and Immunology. 2021;32(1):186-97.
212. Vilarinho S, Sari S, Mazzacuva F, Bilgüvar K, Esendagli-Yilmaz G, Jain D, et al. ACOX2 deficiency: A disorder of bile acid synthesis with transaminase elevation, liver fibrosis, ataxia, and cognitive impairment. Proceedings of the National Academy of Sciences. 2016;113(40):11289-93.

8 Appendix

All appendix tables have been submitted separately and can also be downloaded from:

https://doi.org/10.5281/zenodo.7339851

Code for all chapters can be found on: https://github.com/hlnicholls/PhD-Thesis

8.1 Appendix D - Re-application to Prioritise Blood Lipid Traits

8.2 Introduction

A key benefit of applying a ML approach to prioritise genes is its re-applicability – with a framework finalised, prioritising genes for another phenotype only needs bespoke training data curation. This benefit is utilisable for complex traits such as blood lipids which have had large GWAS studies, with one study taking advantage of the Global Lipids Genetics Consortium to perform a GWAS for 1.6 million individuals for five blood lipid traits[197]: high-density lipoprotein (HDL), non-high-density lipoprotein (nonHDL) low-density lipoprotein (LDL), total cholesterol (TC), and triglycerides (TG). This GWAS was followed by two further functional genomic analyses[183, 192], prioritising blood lipid trait genes that could serve as opportune examples in training data for ML prioritisation. In this chapter, I investigate using blood lipid GWAS data in a re-application of the ML framework developed in chapters 2-4 and explore output prioritised genes and their potential roles in blood lipid biology.

8.2.1 Gene Prioritisation for Blood Lipid Traits

When blood lipids reach abnormal levels in circulation they pose a great cardiovascular disease risk[198], with abnormal lipid levels having the ability to disrupt crucial biological functions (such as cell signalling, cell structure integrity, and energy storage). GWAS research has focused on this phenotype since 2007[199], aiming to better understand the genetic component underlying lipid biology that could potentially illuminate therapeutic targets and improve cardiovascular disease treatments. Most recent blood lipid GWAS' have amassed in size and diversity, with the 1.65 million individuals genotyped by Graham et al. (2021) tested across five ancestries in which 350,000 people were of non-European ancestry. They identified 941 loci associated with five blood lipid traits. In total 53,236 associations with lipid

measurements are recorded in the GWAS catalog (with potentially overlapping associations from various studies unaccounted for), with lipid phenotypes again presenting a bottleneck of genetic data that a ML approach can capitalise on.

The genetic associations identified by Graham et al. (2021) had follow-up investigation over two studies, Ramdas et al. (2021) and Kanoni et al. (2021), that found candidate genes and drug targets. Ramdas et al. (2021) found 1,067 genes colocalised between lipid GWAS signals and eQTL signals, and their functional annotations allowed for tissue-specific and regulatory insights for non-coding variants. They focus on exemplar genes prioritised in their approach (*RRBP1* and *CREBRF*), with for example *CREBRF* having 30 candidate variants identified on colocalisation, and the multi-layered functional information then narrowing these candidates to one single variant (chr5:172,566,698) that interacts with the gene's promoter site in adipose tissue[192]. However, it should be noted that the annotation step is crucial for verifying the eQTL prioritisation approach, considering Mostafavi et al. (2022) showed eQTLs skew to unimportant genes[92]. For example, Mostafavi et al. (2022) found that eQTL signals cluster near transcription start sites and have less functional annotations and regulatory complexity[92]. Overall, suggesting that eQTL colocalisation may provide a biased prioritisation, and their importance on functional annotation was a by-chance finding.

Meanwhile, Kanoni et al. (2021) investigated six gene prediction methods (PoPs, DEPICT, closest gene to the sentinel SNP, genes with coding variants in credible sets, eQTL localisation, and transcriptome-wide association study), combining them to

assign a high to low confidence measure for 2,286 sentinel SNP associations across all five blood lipid traits. They also performed a further evaluation of their combinational prioritisation by re-applying their framework to curated gold standard Mendelian dyslipidaemia genes (n=21) and silver standard genes with mouse model knockouts that have lipid phenotypes (n=739). The prioritisation combining all six methods found 118 genes prioritised with high confidence, with 97/118 having the same high confidence across 5/6 prioritisation methods[183]. On the evaluation of their gold standard genes, they found *"the proportion of gold standard genes in the gene list selected by each approach were: 79.4% by TWAS, 78.4% by PoPS, 62.9% by DEPICT, 58.8% by protein coding variants 44.3% by the closest gene and 26.8% by eQTL"*[183]. These results validate the findings of Mostafavi et al. (2022), as the eQTL prioritisation is the least capable of recognising gold standard genes. It also emphasizes the benefits of a combinational prioritisation approach that enables one method's disadvantages to be compensated for, as shown by Kanoni et al. (2021) then using eQTL in their combined approach to identify 'low confidence' genes.

However, while the method developed by Kanoni et al. (2021) investigates several methods, they do not compare a ML prioritisation approach. Furthermore, the gold and silver standard genes curated by Kanoni et al. (2021) present an opportunity for streamlined training data curation. Where the gold and silver standard can be integrated with the training data criteria developed in this thesis, acting as most likely and probable lipid-regulating genes respectively. Providing training data for re-application of the BP gene prioritisation framework to blood lipid traits. In this chapter, I develop a re-application of the gene prioritisation framework previously

applied to BP GWAS, applying the prioritisation to the blood lipid traits GWAS performed by Graham et al. (2021). I investigate how to best adapt the BP framework to larger lipid GWAS data and explore the downstream ML performance and gene prioritisation in comparison to the BP gene prioritisation applied in chapter 4 and other lipid prioritisation techniques.

8.3 Methods

8.3.1 GWAS Description

GWAS results were aggregated from the Global Lipids Genetics Consortium, genotyping 1,654,960 individuals from 201 studies in five ancestry groups (African, East Asian, European, Hispanic, and South Asian)[197]. Each cohort give summary statistics for all five blood lipid traits and underwent imputation using the 1000 Genomes, with those of European ancestry being additionally imputed using the HRC panel – providing 91 million imputed variants. 52 million variants passed quality control to then enter multi-ancestry meta-analysis of the five blood lipid traits.

8.3.2 Data Collection

Re-applying the framework developed in Chapter 4 (Chapter 4 results 4.3.1) to the GWAS for all five blood lipid traits by Graham et al. (2021) required several data pre-processing steps to be tailored. Firstly, on variant annotation, due to the highly missing annotations of pathogenic variants seen in the BP GWAS and the larger size of the total lipid GWAS data (n=52 million variants) requiring more computational time,

ANNOVAR annotation of pathogenic features was not collected. All other databases collected in the BP GWAS pipeline were annotated following the same protocol outlined in Chapter 4 results 4.3.1. Lipid genes were collated from the genes mapping to associated variants with p-values < $5x10^{-8}$ (n=7,209 from all five blood lipid traits[197]).

Phenotypic data for blood lipids was curated from IPA (searching for genes annotated to 'hyperlipidaemia' in the database) and from Exomiser using the HPO search term 'abnormal circulating lipid concentration'. For the summary statistics, effect sizes calculated by Metal[200] for each of the five blood lipid traits were used, exploring taking the absolute maximum value per gene or median value. However, the Metal effect was found to be heavily missing in the training data (as not all of the gold and silver set genes were in the GWAS with a significant association < $5x10^{-8}$, a filter set for computational efficiency when collating and annotating genes from all five blood lipid trait GWAS) and so was removed on feature cleaning (Appendix D Table 1).

Bedtools (v2.28.0) was used to map variants to the hg19/GRCh37 reference genome from Ensembl (release 92, Homo sapiens.GRCh37.87). For each blood lipid trait GWAS individually, a gene was assigned to a variant if the variant was within a 5kb window distance from start and end of transcription of the gene. Variants in the whole GWAS of each trait (n=~20-35 million variants per each blood lipid trait) were annotated to genes within 5kb. To combine all files into one at a size that could be handled locally, genes annotated to each blood lipid trait were filtered to those that

had at least one variant with a p-value < $5x10^{-8}$, n=7,209 genes). Only UCSC variant-level features for methylation sites were collected as additional variant annotations.

8.3.3 Training Data

Genes used in model training data were scored with values between 0 to 1 for regression analysis. The 21 genes labelled as gold standard genes by Kanoni et al. (2021), were scored at 1. These genes were labelled as such due to being Mendelian dyslipidaemia genes, which were only required to have proximity with GWAS loci but did not need to be associated. Genes were assigned a score of 0.75 if they were labelled as silver standard genes by Kanoni et al. (2021) – with 739 total silver genes filtered to 723 due to not including genes that were also in the gold standard gene group. This silver set was annotated by identifying genes with mouse model knockouts that have lipid phenotypes in IMPC or the Mouse Genome Informatics database that also had the closest proximity to any sentinel SNP. Genes scored at 0.1 were least likely to affect blood lipid traits and were identified by having only variants with a p-value >0.05 and no LD with a sentinel SNP across all five blood lipid traits (n=60 genes). LD was calculated for all blood lipid sentinel SNPs using UKBiobank data that has passed QC and PLINK (v.1.9) with an LD threshold $r^2 > 0.1$ and a 1Mb interval region. Due to the low number of least likely genes that met the >0.05 p-value thresholds across all five blood lipid traits, no further PPI filtering was performed. These three scorings provided 804 training genes (21 genes scored at 1, 723 genes scored at 0.75, and 60 genes scored at 0.1) (Appendix D Table 2).

8.3.4 Data Pre-processing and Machine Learning Model Benchmarking

Data pre-processing assessed genetic bias risks (gene length and gene distance), removed heavily missing and correlating features, and involved BorutaShap feature selection. Any variant-level features highly correlated with gene length (>0.3) were removed (Appendix D Table 3), alongside any features that were >25% missing or had a >0.9 correlation (Pearson r^2) (Appendix D Table 4). ML with different correlation thresholds (0.85 and 0.99) were tested (test runs included https://github.com/hlnicholls/PhD-Thesis/tree/main/Chapter5). Further details of these data pre-processing methods can be found in Chapter 2 Methods section 2.2.2 and Chapter 4 Methods section 4.2.3.

The features were imputed using random forest imputation[201] (using the missingpy package, v0.2.0) and underwent feature selection using the BorutaShap package (v1.0.13).

Fourteen regression models were benchmarked, re-applying the ML methods from Chapter 4 Methods section 4.2.3. The top-performing model, underwent interpretation using the SHAP package[26] (v0.36.0), providing feature importance values both globally for overall model performance and individually for each gene. Plots of the feature importance (for both overall predicting and individual predictions) were created alongside feature-feature interactions.

8.3.5 Blood Lipid Traits Gene Prioritisation Analysis

After ML prioritisation, the gene selection per locus algorithm developed in Chapter 4 was applied – with the algorithm steps outlined in Chapter 4 methods section 4.2.4. All genes prioritised (including training genes scored at 1 and 0.75) entered the gene selection algorithm. With genes sorted into 923 loci via having 500kb +/- distance with a sentinel SNP (n=2,624) as defined by Graham et al. (2021). In comparison, the BP GWAS loci ordering used both 500kb+/- distance as well as identifying genes with SNPs in LD ($r^2 > 0.8$) with a sentinel SNP. LD was not provided by Graham et al. (2021) and so was not used in the loci ordering here. However, for confirmation, LD calculated for all sentinel SNPs using PLINK was used to identify genes in a locus by them having SNPs in high LD ($r^2 > 0.8$) with a sentinel SNP. To ensure a full comparison with the loci also prioritised by Kanoni et al. (2021), LD was not used to filter lipid genes or order genes into loci, enabling the prioritisation of as many genes as possible.

I investigated the genes scored > 0.8 by the top-performing model (known as *highly-scored genes*) and the selected genes per locus (known as *selected-genes*) in downstream analysis by investigating their distributional differences for several collected annotations using the Mann-Whitney U test in R and plotting their gene expression across all tissues in GTEx (v8) using ComplexHeatMap (v2.6.2). I further explored GTEx annotations by collecting statistically significant sex-specific gene expression bias annotated in GTEx's most recent data release[169] – in which they identify statistically significant sex effect sizes in gene expression across tissue meta-analysis. This annotation allowed for analysis of sex-specific bias as defined by GTEx

across 44 of their tissues for all prioritised genes and enabled further focus of potential sex-specific biased expression for the selected GTEx features used by the ML model.

The R package GeneOverlap (v1.30.0) was used to perform hypergeometric tests on gene hits in IMPC phenotypes, testing the overlap of gene hits for the *highly-scored genes* and the *selected-genes* against the total number of genes in each phenotype in comparison to a total 1,875 genes annotated in IMPC. For gene enrichment analysis, I used KEGG to compare four groups: the *selected-genes,* the *highly-scored genes*, the genes containing sentinel SNPs[47] GWAS, and the gold and silver set genes used in training (genes scored at 1 and 0.75 respectively). Gene enrichment analyses were visualizsd using the ComplexHeatMap package (v2.6.2) in R.

8.3.6 Priortisation Methods Comparison

The ML prioritisation was compared with other methods applying ML for gene prioritisation (Mantis-ml[33], ToppGene [138], and GPrior[19]), alongside OpenTargets association scores[202], and the combined confidence given from all six prioritisation methods performed by Kanoni et al. (2021). Unlike the BP GWAS, OpenTargets Genetics L2G scores were not available for the GWAS by Graham et al. (2021). Instead, OpenTargets overall association scores for hyperlipidaemia were used for comparison, which aggregates data from 12 databases in a harmonic sum[202]. GPrior, Mantis-ml and ToppGene required an input of positive genes, being given the 21 genes – gold standard lipid genes (scored at 1 for our ML) as positive examples[19, 33, 138]. They then required different parameters to run. ToppGene only required the gene list to be prioritised as input, and all other parameters were set to ToppGene's default training

parameters[138]. Mantis-ml required the phenotype term of interest[33] – for input 'abnormal circulating lipid concentration' and 'hyperlipidaemia' were input and no excluded terms were set. GPrior is the only method that allows the user to input their own features[19], and so for this method I input the selected features also used by our model corresponding with our gene list to be prioritised. The combined confidence assignment by Kanoni et al. (2021) ranged from 'low', 'medium low', 'medium high' and 'high' based on the combinational criteria from the prioritisation methods they investigated:

- High confidence: Mendelian gene[183].
- Medium high confidence: coding variants or mouse knockout gene or PoPS[183].
- Medium low confidence: closest gene or DEPICT or TWAS[183].
- Low confidence: eQTL[183].

8.4 Results

8.4.1 Framework Re-application

Re-applying the framework developed in Chapter 4 (Chapter 4 results 4.3.1) to the GWAS for all five blood lipid traits by Graham et al. (2021) required several data pre-processing steps to be tailored (Figure 5.1). Firstly, ANNOVAR annotation of pathogenic features was not collected due to the larger size of the total lipid GWAS data (n=52 million variants). Secondly, the training data had gold standard and silver standard lipid genes provided by Kanoni et al. (2021), giving gene groups scored at 1.0 and 0.75 respectively (n=744 in total).

Notably unlike the BP GWAS training data, these training genes were not required to be also found within the GWAS with only 395/744 also being identified in the 7,209 lipid genes. For each of the five blood lipid traits, least likely protein-coding genes were then identified by having p-values > 0.05 across all five traits, giving only 60 genes that met this criterion. Overall, this gene curation gave a total of 804 training genes that had 85 annotations that underwent feature cleaning and BorutaShap feature selection to enter model benchmarking of fourteen models.

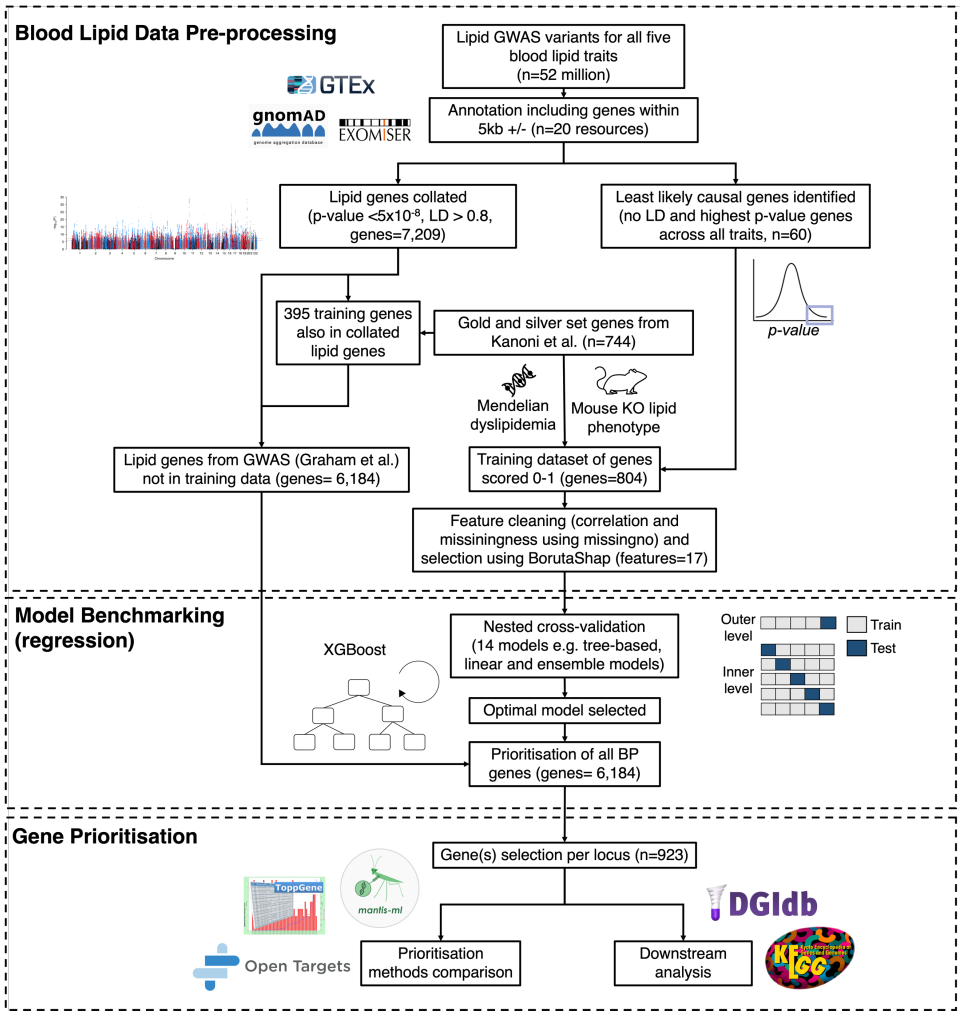


Figure 5.1. Overview of the machine learning framework re-applied to blood lipid trait GWAS. Blood lipid trait genome-wide association study variants from Graham et al. (2021) were annotated to genes and evaluated by benchmarked machine

learning. Data pre-processing involved annotating variants to genes from the whole GWAS and collecting gene-level annotations from several databases. The genes were then filtered to identify lipid genes (with linkage disequilibrium, LD, $r^2 > 0.8$ and a p-value $<5x10^{-8}$) and that are genes least likely to affect blood lipid traits (selected by meeting criteria of having all variants with p-values > 0.05 across all five traits). Training genes of gold and silver standard for impacting blood lipid traits were previously identified by Kanoni et al. (2021). These genes in combination with the least likely lipid genes then created the training dataset and remaining lipid genes were reserved to be predicted by the top-performing model with regression. Model benchmarking was then applied comparing 14 models using 17 selected features. The top-performing trained model (extreme gradient boosting) was then used for gene prioritisation, with the genes and their corresponding scores being assessed within their loci to select the best gene(s) per locus. The prioritised genes underwent downstream analyses and were compared with other prioritisation methods.

8.4.2 Exploratory Data Analysis

Training data curation produced a dataset of 804 genes (21 scored at 1, 723 genes scored at 0.75, and 60 genes scored at 0.1). On exploring the genomic characteristics of the three gene groups, they were found to vary in their chromosome position and gene length (Figure 5.2 and Appendix D Table 5). The 0.1 scored genes had the shortest gene lengths measured by median, maximum and minimum gene lengths in the group (Appendix D Table 5). Feature cleaning removed 20 heavily missing features. Followed by all highly correlating (>0.3) variant-level features being removed to avoid gene length bias (5 features for histone methylation and DNase

cluster signal values). Then, 38 highly correlating features with a > 0.9 correlation threshold (which were all GTEx tissue features) were also removed (Appendix D Table 4), giving 27 features in total passing data cleaning. Finally, this was followed by BorutaShap feature selection identifying 17 important features (12 GTEx tissues, Exomiser human scores, HIPred, SDI, gene damage index - GDI - and pLI ExAC scores) to enter model benchmarking, with all selected features having similar importance (Figure 5.3). It should also be noted that from the 12 selected GTEx tissues, six of them (amygdala, atrial appendage, kidney, lung, small intestine ileum, and ovary) were in correlative pairs with > 0.9 r^2 (Appendix D Table 4), however, the other feature was the one removed (with one feature of each correlating pair being dropped at random).

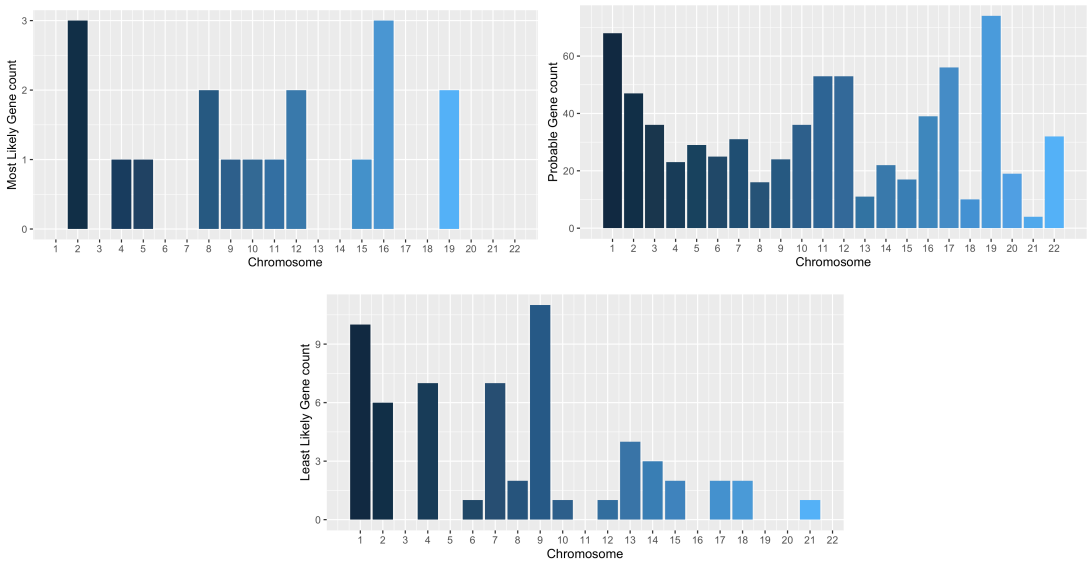


Figure 5.2. Training genes distributions across chromosomes. Counts of training gene groups across the three labels - most likely (gold standard, n=21), probable (silver standard, n=723), and least likely (n=60) - at each chromosome across the genome.

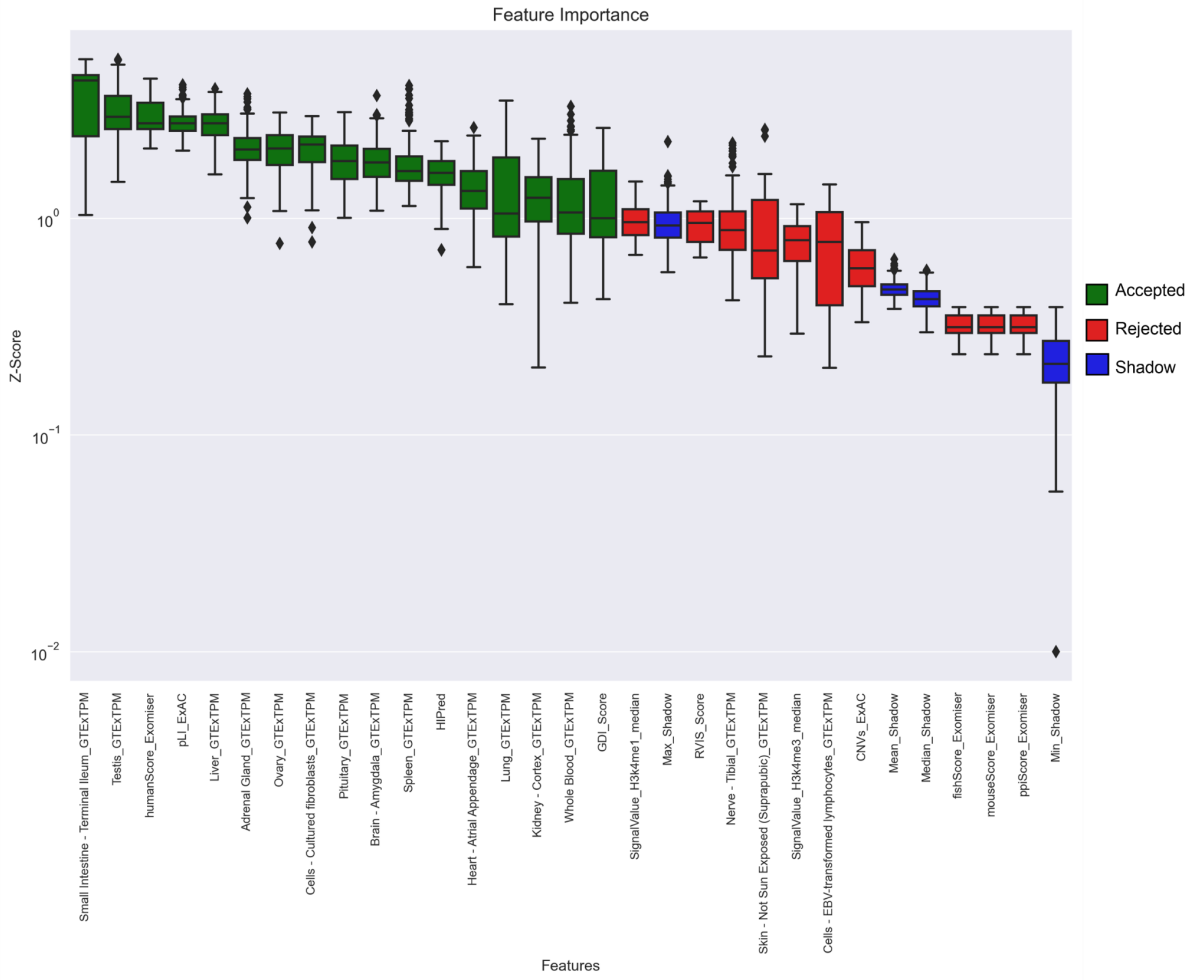


Figure 5.3. Feature importance measures. Box plot of all features that underwent BorutaShap (<25% missingness and <0.9 r^2) and their measured importance by BorutaShap over 100 iterations of the BorutaShap algorithm (using z-scores), ordered from left to right in descending feature importance. Green boxes indicate selected features, red boxes indicate rejected features, and blue boxes indicate shadow features.

8.4.3 Model Benchmarking

The benchmarked models showed similar trends to the performance in Chapter 4 on BP GWAS data (Figure 5.4). All tree-based models performed similarly, with XGB being the top-performing amongst them (0.708 r^2 and 0.826 predicted r^2) (Table 5.1). Meanwhile, from the meta-ensemble models, the bagging XGB model had a higher r^2 and lower predicted r^2 (0.715 and 0.795 respectively). The lower predicted r^2 led to XGB without bagging being the selected model for further investigation.

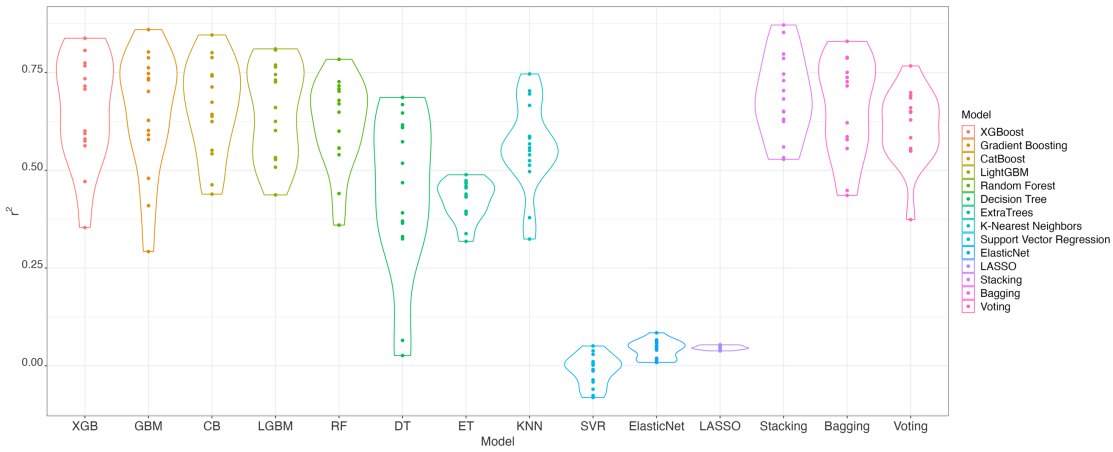


Figure 5.4. Model benchmarking. Fourteen models were benchmarked: extreme gradient boosting (XGB), gradient boosting (GBM), catboost (CB), LightGBM (LGBM), random forest (RF), decision tree (DT), Extratrees (ET), K-nearest neighbors (KNN), support vector regressor (SVR), two linear models using regularization of elastic net and LASSO, respectively, and three meta-ensemble methods – stacking, bagging, and voting models. The model performance was assessed on 5-fold nested cross-validation repeated three times.

Model	Median r^2	Predicted r^2	Mean Square Error	RMSE	Explained Variance	Mean Absolute Error
XGB	0.708	0.826	0.009	0.01	0.716	0.046
GB	0.701	0.938	0.009	0.0973	0.702	0.045
CB	0.674	0.796	0.009	0.095	0.68	0.0457
LGBM	0.66	0.757	0.009	0.0968	0.672	0.0478
RF	0.649	0.706	0.01	0.1054	0.663	0.046
DT	0.468	0.525	0.0167	0.1293	0.4704	0.0471
ET	0.434	0.443	0.0178	0.1334	0.4342	0.0699
KNN	0.557	0.985	0.0135	0.116	0.5675	0.0455
SVR	0.002	-0.048	0.0311	0.1764	0.043	0.1256
LASSO	0.039	-0.007	0.0297	0.1723	0.045	0.0952
ElasticNet	0.046	0.03	0.0298	0.1726	0.046	0.0965
Stacking	0.682	0.793	0.009	0.0973	0.686	0.0389
Bagging	0.716	0.795	0.009	0.01	0.717	0.045
Voting	0.629	0.79	0.01	0.107	0.65	0.058

Table 5.1. Model benchmarking performance. Median performance comparison on nested 5-fold cross-validation. Only predicted r^2 measurements were not median calculations but calculated from each model's performance after hyper-parameter tuning. The fourteen models benchmarked were: extreme gradient boosting (XGB), gradient boosting (GBM), catboost (CB), LightGBM (LGBM), random forest (RF), decision tree (DT), Extratrees (ET), K-nearest neighbors (KNN), support vector regressor (SVR), two linear models using regularization of elastic net and LASSO,

respectively, and three meta-ensemble methods – stacking, bagging, and voting models.

8.4.4 Model Interpretation

SHAP was used to interpret XGB's performance, finding the most important feature was the human Exomiser score, followed by testis gene expression, probability loss of function (pLI), pituitary gene expression and the GDI score as the top five features (Figure 5.5.a). Examining XGB's prioritisation of all gold standard genes (assigned a training score of 1) showed the model was successful in identifying all 21 of these genes had a higher score, with the lowest scored gold standard gene being *CYP27A1* prioritised at 0.743 (Figure 5.5.b). Notably from the 17 selected features, several GTEx tissues' gene expression had minimal SHAP values (Figure 5.5.a and 5.5.b), and they also had minimal feature-feature interaction, with the strongest influencing relationship identified by SHAP being between HIPred and pLI (Figure 5.6).

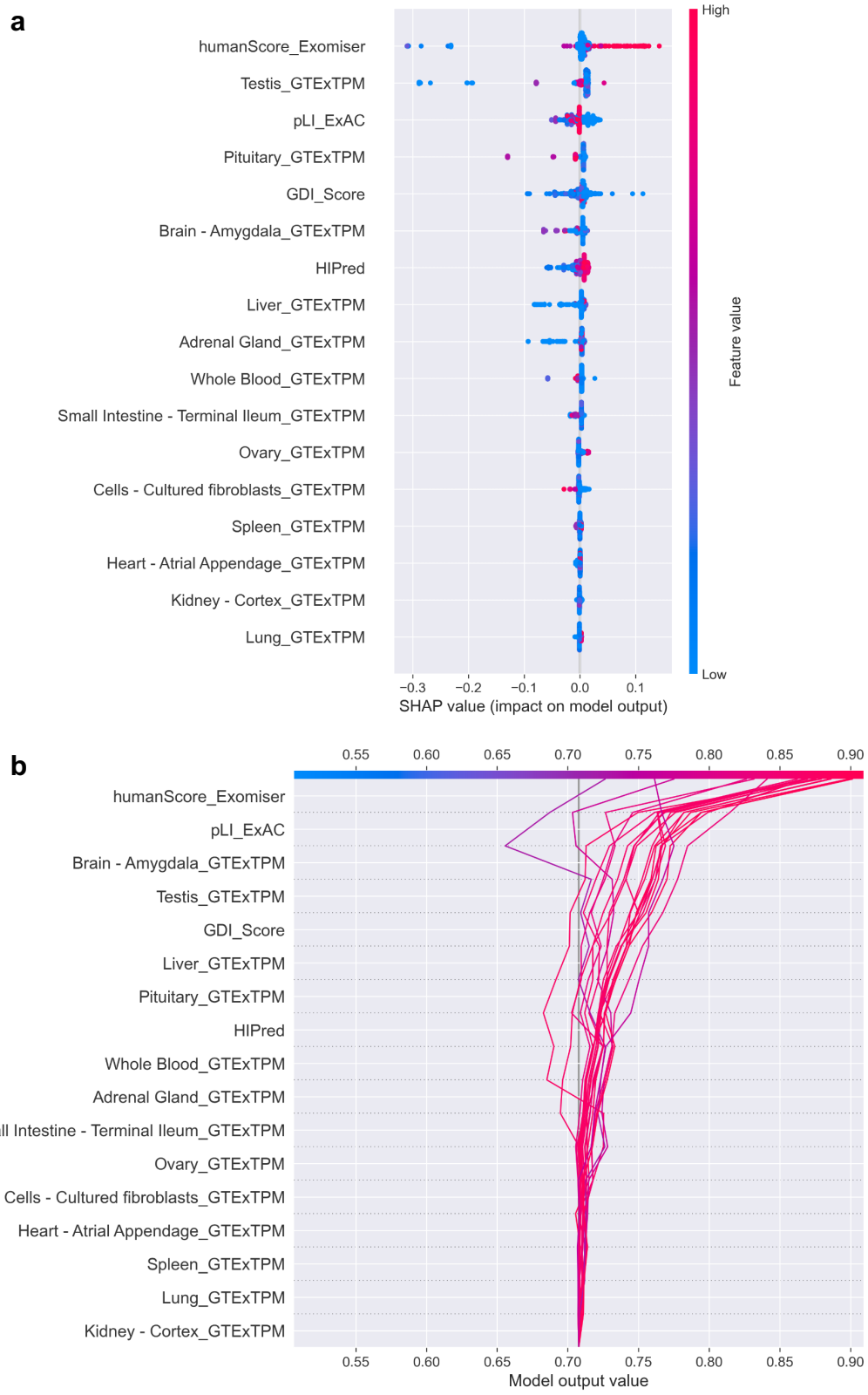


Figure 5.5. Shapley additive explanation summary plots of model decision-making. a SHapley Additive exPlanation (SHAP) summary plot of the top-

performing model (extreme gradient boosting) predictions of all genes and how they were each influenced by each feature. The SHAP value on the x-axis indicates the direction of model influence from that feature for each gene (e.g., a higher SHAP value indicates a more positive output model score). The colour-coding of points (genes) indicates whether their feature value was high (red) or low (blue), and the ordering of features on the y-axis is by descending feature importance. **b** SHAP summary plot of the 21 gold standard gene predictions, visualising the model's use of features for predicting each of the gene's predicted scores (on the x-axis) – with these also being plotted against a black vertical line which is the average model score for all training data (0.7).

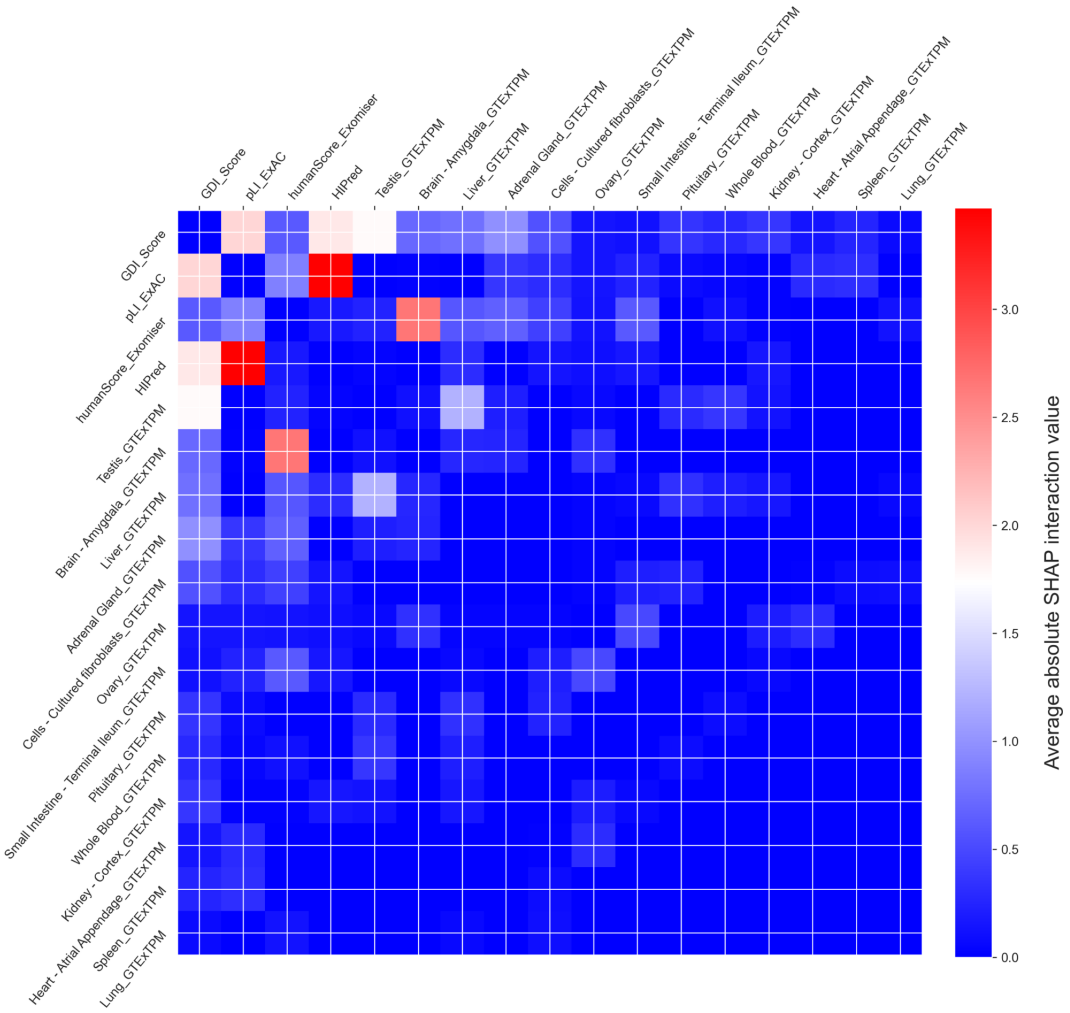


Figure 5.6. Shapley additive explanation heatmap of model decision-making. Heatmap showing the absolute SHAP value of feature-feature interactions, measuring the impact feature interactions had on model decision-making overall, with a red colour gradient indicating a larger influence on the model and a blue colour gradient indicating less to no model influence.

8.4.5 Blood Lipid Traits Gene Prioritisation

Once benchmarked and fit to the training data, the top-performing model, XGB, was used to prioritise all lipid genes (n=7,209) (Appendix D Table 6). I investigated the

highly prioritised genes by assessing intolerance metrics which were not used by the XGB model, using the Mann-Whitney U test. All the genes scored greater than 0.8 by the XGB model (‘*highly-scored genes’*) had significantly different values on Mann-Whitney U tests in comparison to genes with a XGB score < 0.8 for gene essentiality (measured by Avana mean), and drug target probability (measured by SDI) (Appendix D Table 7, Figure 5.7). The most significant difference was the Avana mean with a Mann-Whitney U test adjusted p-value of 4.4×10^{-53} between the *highly-scored genes* and all other genes scored less than 0.8 by XGB (Appendix D Table 7). The *highly-scored genes* had more negative Avana mean values indicating that more essential genes were highly prioritised (Figure 5.7).

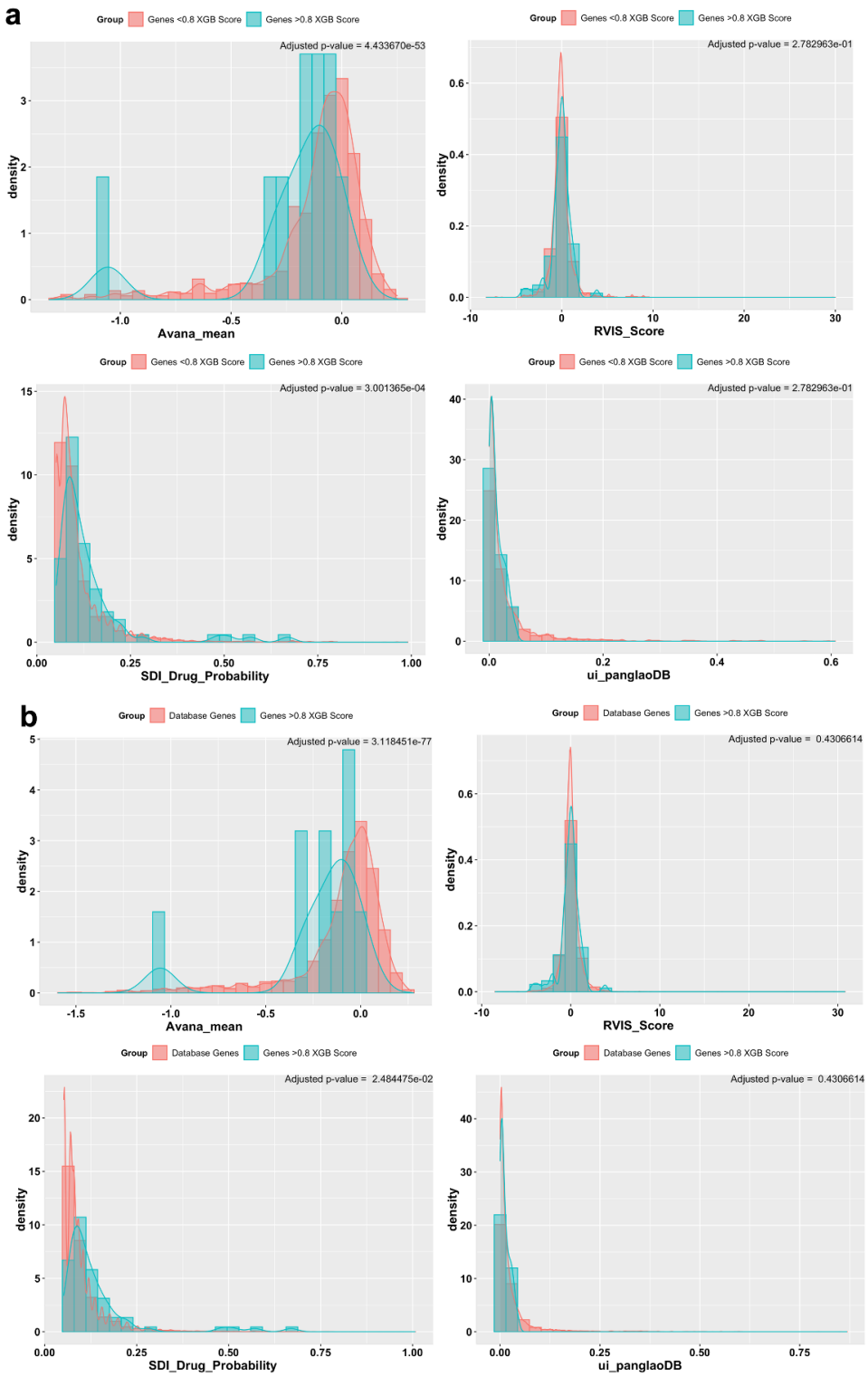


Figure 5.7. Distributions of annotations for genes prioritised > 0.8 versus genes scored < 0.8 (a) and genes > 0.8 versus total database annotations (b). Annotations

not used in machine learning were plotted to compare genesets across several measures: subcellular diversity index (SDI) drug probability, gene essentiality (avana mean), genic intolerance (RVIS), and ubiquitous cell-type expression (panglaoDB). Genes scored > 0.8 had their annotations compared against that of genes < 0.8 (a), and that of the total genes in a database for each annotation (b). The Mann-Whitney U test identified significant differences in distributions.

From the selected gene(s) per locus (the "*selected-genes*"), in total, 2,327/7,209 genes were selected for 923 total blood lipid loci (Appendix D Table 8), with 343 loci having > 1 gene selected at their loci. Genes were sorted into 923 loci via having 500kb +/- distance with a sentinel SNP (n=2,624) as defined by Graham et al. (2021). LD was not provided by Graham et al. (2021) and so was not used in the loci ordering here. However, for confirmation, LD was calculated, which when applied to identify genes in LD with $r^2 > 0.8$ with a sentinel SNP gave genes that filtered into only 457 of the 923 loci.

The *selected-genes* also had significant differences on Mann-Whitney U tests in their annotations across several measures (Appendix D Table 7, Figure 5.8). The *selected-genes* had the most significant difference in comparing their SDI drug probability with that of all other genes in the SDI database (adjusted p-value = 3.28×10^{-23}) (Appendix D Table 7, Figure 5.8), indicating that genes with a higher likelihood of being drug targets were highly prioritised by XGB.

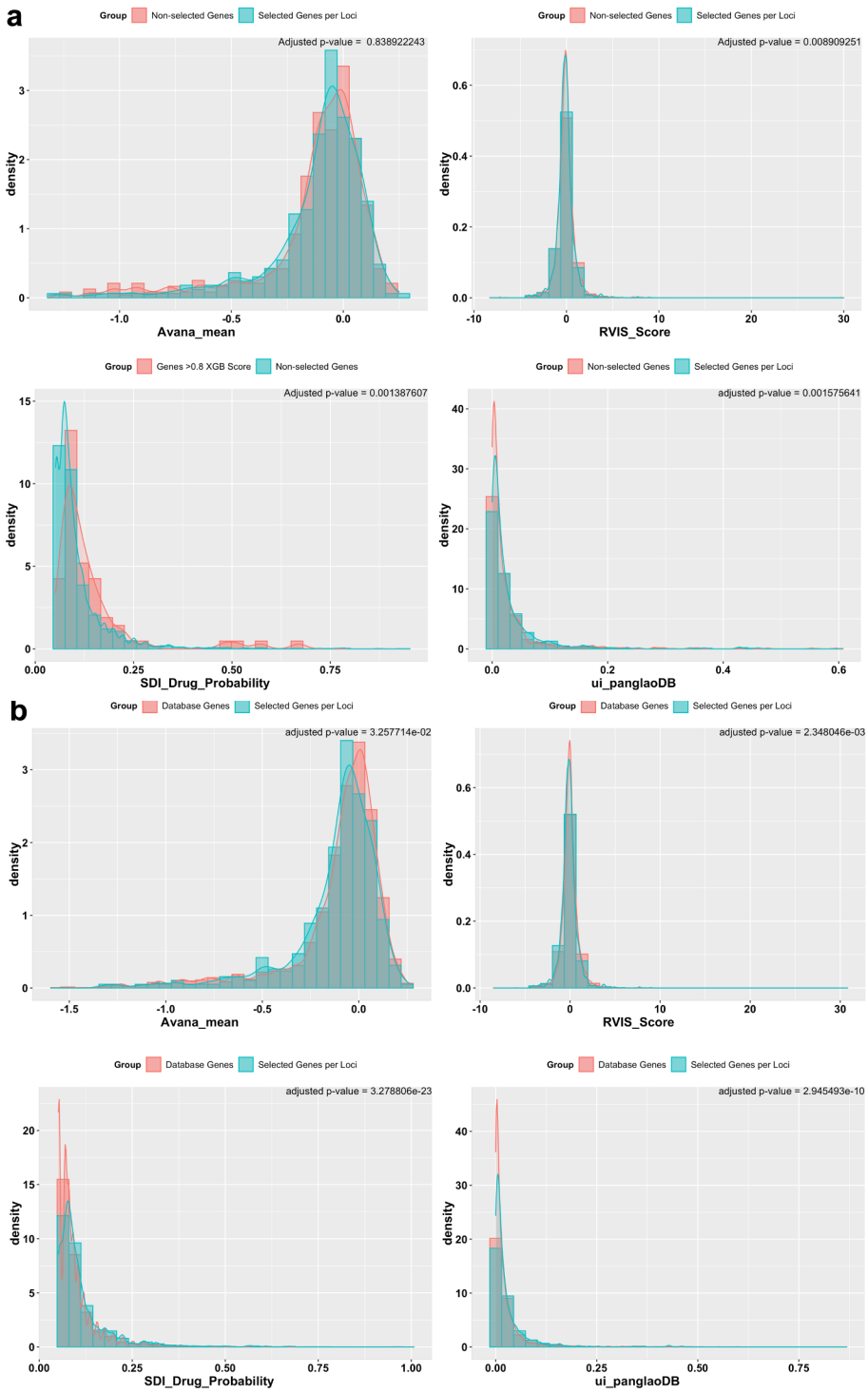


Figure 5.8. Density distributions of annotations for selected genes per locus versus all other scored genes (a) and selected genes per locus versus total database annotations (b). Annotations not used in machine learning were plotted to compare genesets across several measures: subcellular diversity index (SDI) drug probability,

gene essentiality (avana mean), genic intolerance (RVIS), and ubiquitous cell-type expression (panglaoDB). Genes scored > 0.8 had their annotations compared against that of genes < 0.8 (a), and that of the total genes in the database for each annotation (b). The Mann-Whitney U test identified significant differences in distributions.

Overall *KHK* was the top scored novel gene (0.895 XGB score) with other highly prioritised genes by XGB including *GLYCTK, SORD, UNC13D, PPOX, CREB3L3, PIGT, SLC25A20*, *PCYT1A* and *ACOX2* (Table 5.2). Some of these genes have been the focus of *in vivo* and clinical research investigating their roles in lipid metabolism and cardiovascular conditions such as *KHK*[187], and *CREB3L3*[188].

Gene	**XGB Score**	**Gene Description**	**Potential Druggability (DGIdb)**	**Most significant Pathway (KEGG)**	**Median GWAS p-value**
KHK	0.895	Ketohexokinase - catalyses conversion of fructose	Druggable genome	Fructose and mannose metabolism	5.62×10^{-12}
GLYCTK	0.892	Glycerate Kinase - catalyses the phosphorylation of (R)-glycerate	Kinase	Pentose phosphate pathway	5.13×10^{-12}
SORD	0.891	Sorbitol dehydrogenase - catalyses conversion	Druggable genome	Fructose and mannose metabolism	1.02×10^{-8}

		of polyols and ketoses			
UNC13D	0.871	Unc-13 Homolog D – acts in vesicle maturation and regulation of cytolytic granules secretion	NA	NA	$6.85x10^{-15}$
PPOX	0.871	Protoporphyrinogen Oxidase - the penultimate enzyme of heme biosynthesis	Druggable genome	Porphyrin and chlorophyll metabolism	$2.1x10^{-08}$
CREB3L3	0.869	CAMP Responsive Element Binding Protein 3 Like 3 - transcription factor activated by cyclic AMP stimulation	Transcription factor	Vasopressin-regulated water reabsorption	$7.92x10^{-13}$
PIGT	0.868	Phosphatidylinositol Glycan Anchor Biosynthesis Class T - involved in glycosylphosphatidylinositol (GPI)-anchor biosynthesis	Druggable genome	Glycosylphosphatidylinositol (GPI)-anchor biosynthesis	$1.72x10^{-10}$

SLC25A20	0.854	Solute Carrier Family 25 Member 20 - mitochondrial-membrane carrier protein	Druggable genome, transporter	Thermogenesis	$1.85x10^{-11}$
PCYT1A	0.852	Phosphate Cytidylyltransferase 1A, Choline - involved in the regulation of phosphatidylcholine biosynthesis	Enzyme	Phosphonate and phosphinate metabolism	$1.26x10^{-8}$
ACOX2	0.85	Acyl-CoA Oxidase 2 - involved in the degradation of long branched fatty acids and bile acid intermediates in peroxisomes	Enzyme	Primary bile acid biosynthesis	$3.59x10^{-10}$

Table 5.2. Description of the top ten novel prioritised genes. The top ten scored genes by XGBoost (XGB) and descriptions of: their gene function, their druggability as annotated by the Drug-Gene Interaction database, their most significant Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway, their other locus gene(s), and their median GWAS p-value.

8.4.6 Gene Expression

I next explored gene expression clustering across all 54 tissues available in GTEx[147] for both the *highly-scored genes* and *selected-genes*. The *highly-scored genes* had no notable clusters, however, the *selected-genes* identified a group of genes (*RPL13A, RPS11, RPL19, RPS27, RPL5, EEF2, EEF1G, RPS6, RPL4, RPS7, VIM, FN1, and MUC7*) containing predominantly ribosomal proteins with high gene expression across 33 tissues including those relating to lipid metabolism (such as adipose tissue, arteries, and pancreatic tissue).

In considering the sex-specific tissues selected as features and the potential sex-specific bias leaking into model decision-making, I also compared whether the prioritised genes were also identified as having sex-specific bias as defined by GTEx across 44 of their tissues[169] (Appendix D Table 8). This showed that 54.3% of *selected-genes* were annotated as having a sex-specific bias in their gene expression for various tissues by GTEx[169]. From the most important tissues used by XGB, 120 *selected-genes* had a sex-specific bias in their pituitary tissue expression (out of 1,420 genes with sex-specific bias in that tissue in the total GTEx database), 236 genes had a sex-specific bias for amygdala expression (out of 2,398 genes with sex-specific bias in that tissue in total in GTEx) and 65 genes had a sex-specific bias for liver expression (out of 717 genes with sex-specific bias in that tissue in total in GTEx). However, while these genes are annotated as individually being statistically significant for sex-biased expression in GTEx[169], on hypergeometric testing of all prioritised genes biased for each tissue no gene group with overlapping *selected-genes* had statistical significance.

8.4.7 Gene Enrichment Analysis

The prioritised genes were further explored in four gene groups (genes scored > 0.8, selected genes per locus, sentinel genes, and gold and silver standard lipid training genes combined). Enrichment analysis of these gene sets found that all gene groups were most significantly enriched for cholesterol metabolism and lipid metabolism pathways such as PPAR signalling and fat digestion (Figure 5.9). Plotting the gene interaction network between these top five pathways showed only *CD36* (scored 0.763 by XGB) acts within all five, with it having druggability identified by DGIdb, alongside several genes that overlap with at least 2 of the pathways (e.g., *FABP1* scored 0.771 by XGB and *FABP2* scored 0.752 acting in PPAR signalling and fat digestion and absorption) (Figure 5.10).

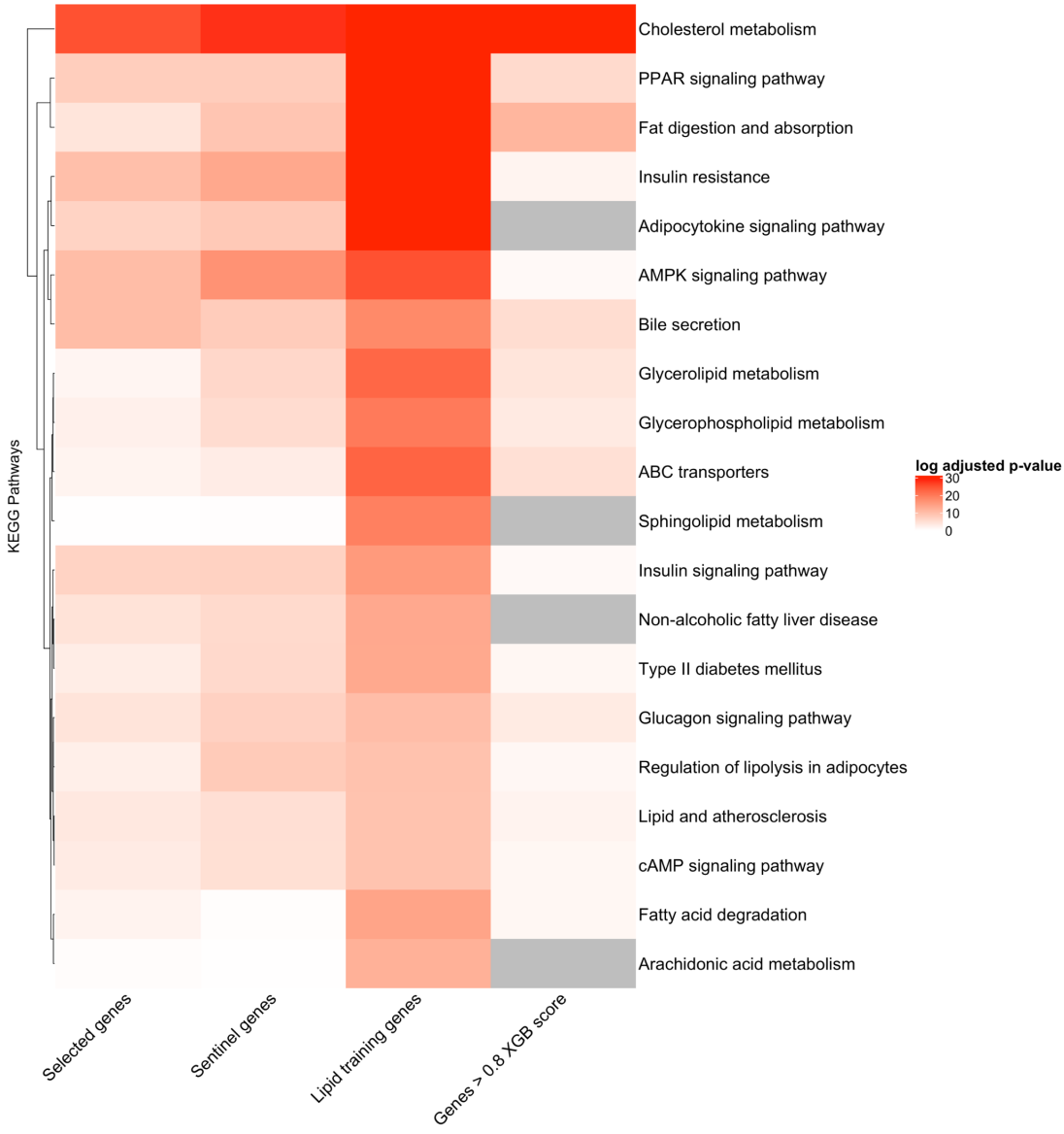


Figure 5.9. KEGG pathway analysis. Heatmap of the top 20 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. The heatmap shows more significant values are indicated by darker shades of red, and no enrichment in grey. Four gene groups are compared, composed of genes with a > 0.8 XGB score (n=60), lipid training genes (gold and silver standard lipid genes) (n=744), sentinel genes (identified in the genome-wide association study by Graham et al. (2021)) (n=1,222), and *selected-genes* (genes selected at their locus) (n=2,327).

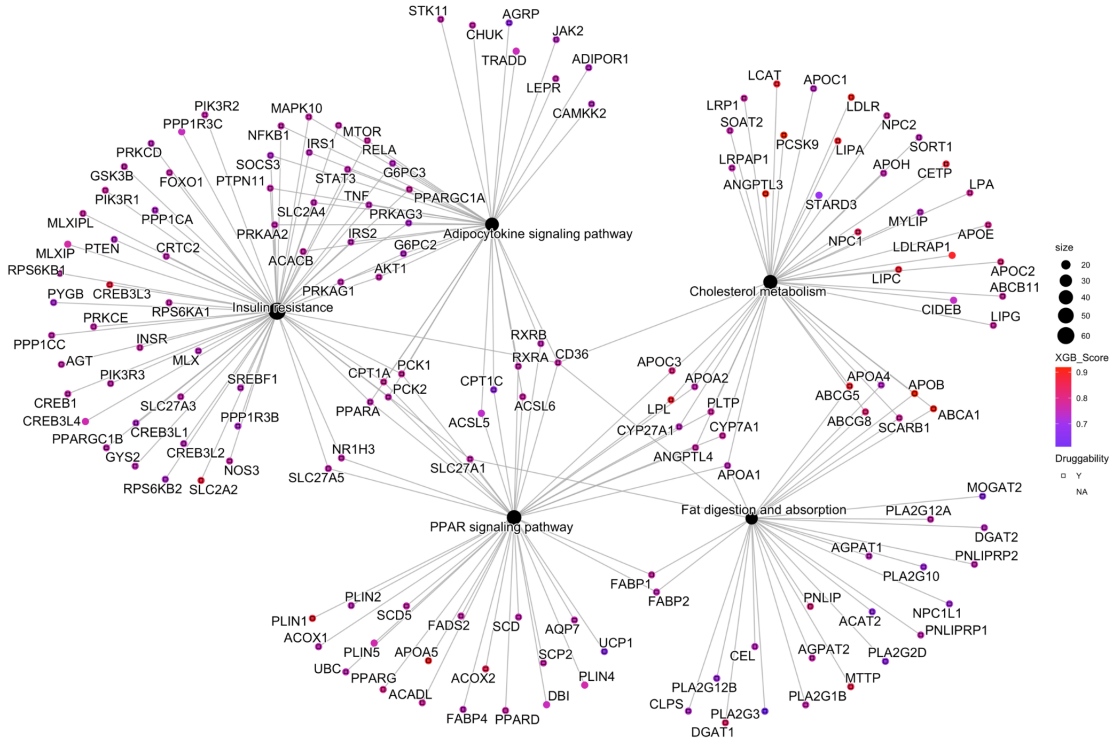


Figure 5.10. Gene interaction network of the top five most significantly enriched pathways. The KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway interactions have gene nodes colour-coded with higher prioritised genes by extreme gradient boosting in dark red and lower-scored genes scored in light red. Pathway node size indicates enrichment log p-value for each pathway node. Square symbols represent whether the gene had druggability recorded in the Drug Gene Interaction Database. For example, *CD36* is denoted in the centre of the plot, interacting with all five pathways, with annotated druggability and a high XGB score (0.763).

8.4.8 Prioritisation Methods Comparison

I selected other gene prioritisation methods to compare against XGB and analyse their predictions for the blood lipid genes. 1) OpenTargets association scores to hyperlipidaemia, 2) Mantis-ml: a positive-unlabelled learning approach that

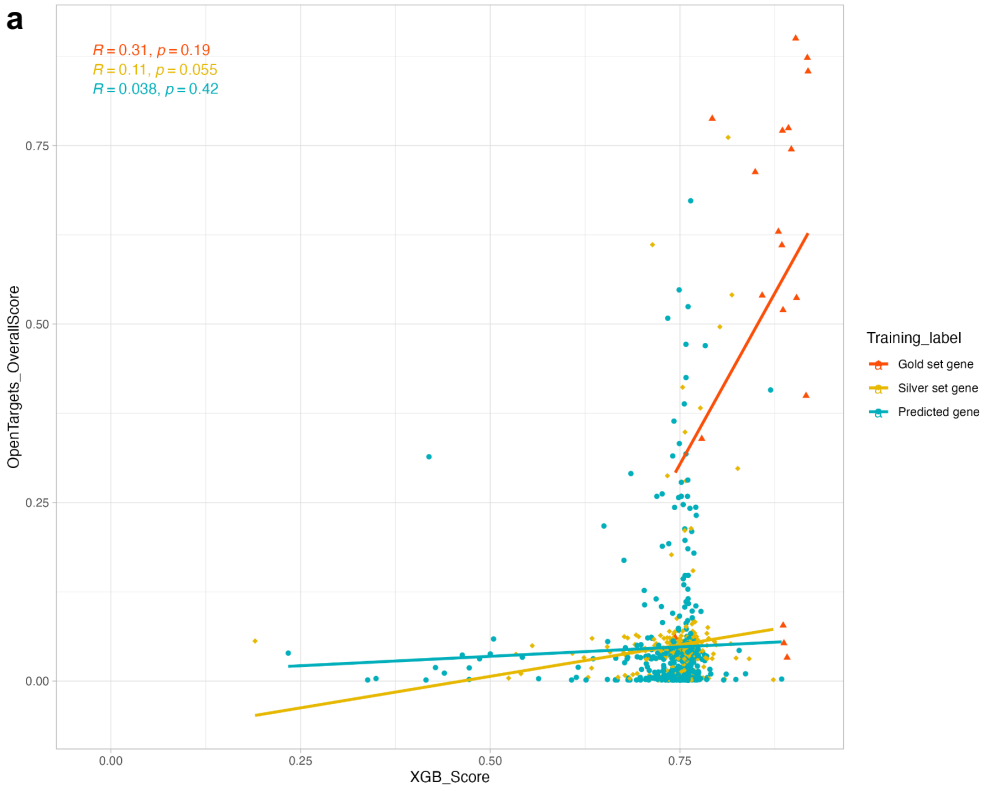
benchmarks several models[33], 3) ToppGene: an unsupervised learning method[138], and 4) GPrior: a positive-unlabelled learning tool that consists of bagging ensemble models[19], 5) all six prioritisation methods used by Kanoni et al. (2021) (PoPs, DEPICT, closest gene to the sentinel SNP, genes with coding variants in credible sets, eQTL localisation, and transcriptome-wide association study). However, it should be noted that prioritisation methods 1-4 output scores between 0-1 whilst the use of six methods by Kanoni et al. (2021) output labels of low to high confidence. All methods 1-4 showed positive correlations for prioritising lipid genes used in our training data (the 21 gold standard lipid genes and 723 silver standard lipid genes) and for all predicted genes (the 7,209 genes prioritised by the trained model) (Table 5.3, Figure 5.11). I then compared the confidence assignments using six prioritisation methods by Kanoni et al. (2021) (Table 5.4) (Appendix D Table 9). All genes assigned 'high' confidence levels were scored > 0.5 by XGB with the lowest gene scored at 0.59 (*UBE2L3*) and a median XGB score being 0.76 (followed by slightly lower median XGB scores for the lower confidence levels albeit only by 0.01 difference, Table 5.4) (Appendix D Table 9).

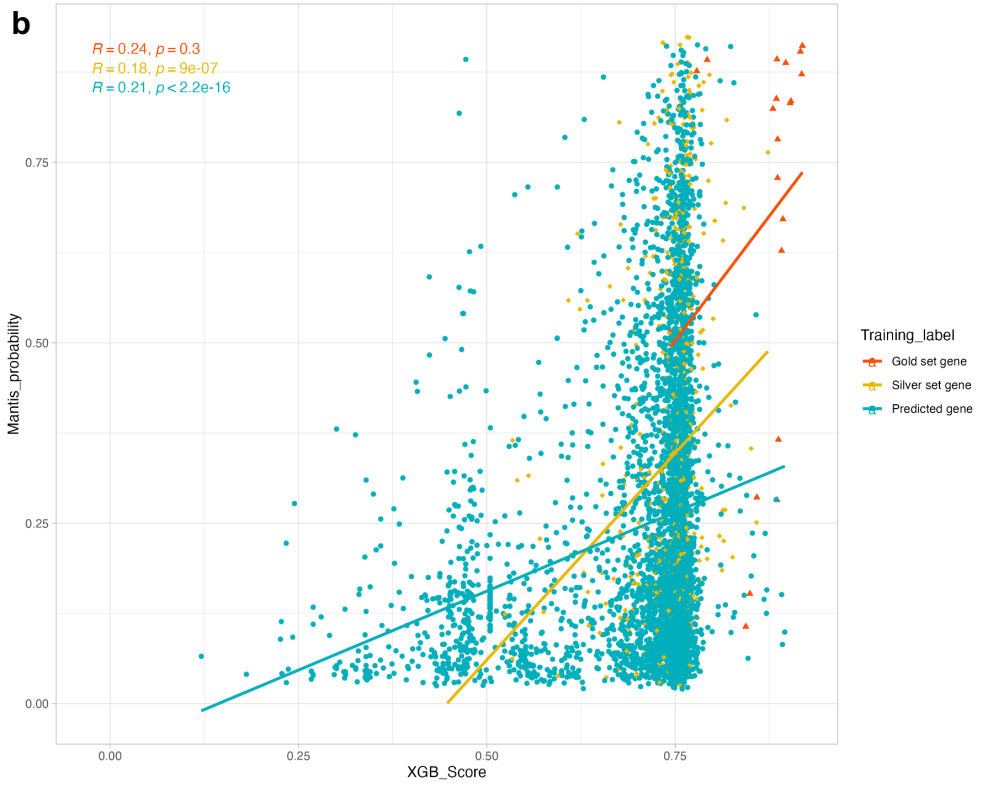
	OpenTargets Association Score	GPrior	Mantis-ml	ToppGene
Gold standard genes (scored 1.0 on training)	0.31	NA	0.24	NA
Silver standard genes (scored 0.75 on training)	0.11	0.24	0.18	0.033
Predicted genes	0.038	0.64	0.21	0.36

Table 5.3. Comparison of gene prioritisation methods. Table comparing the prioritisation of training genes that were scored as the gold standard (scored at 1.0) or probable (scored at 0.75) BP genes (n=744) and predicted genes (n=7,209) by several methods in comparison to extreme gradient boosting, measured by their correlation (R) for their predicted gene scores.

Gene Confidence	Median XGB Score	Maximum XGB Score	Minimum XGB Score
High (n=118)	0.7603	0.9188	0.5936
Medium high (n=1716)	0.75080	0.89540	0.05998
Medium low (n=2,897)	0.74735	0.89672	-0.02496
Low (n=261)	0.7473	0.9188	0.1774

Table 5.4. Prioritisation comparison between confidence levels and XGBoost scoring. Confidence levels used to prioritise genes by Kanoni et al. (2021) were compared with their predicted scores by XGBoost, investigating the gene prioritisations for each of the confidence levels (from low to high).


a
R = 0.31, p = 0.19
R = 0.11, p = 0.055
R = 0.038, p = 0.42
OpenTargets_OverallScore
0.00
0.25
0.50
0.75
XGB_Score
Training_label
Gold set gene
Silver set gene
Predicted gene


b
R = 0.24, p = 0.3
R = 0.18, p = 9e-07
R = 0.21, p < 2.2e-16
Mantis_probability
0.00
0.25
0.50
0.75
XGB_Score
Training_label
Gold set gene
Silver set gene
Predicted gene

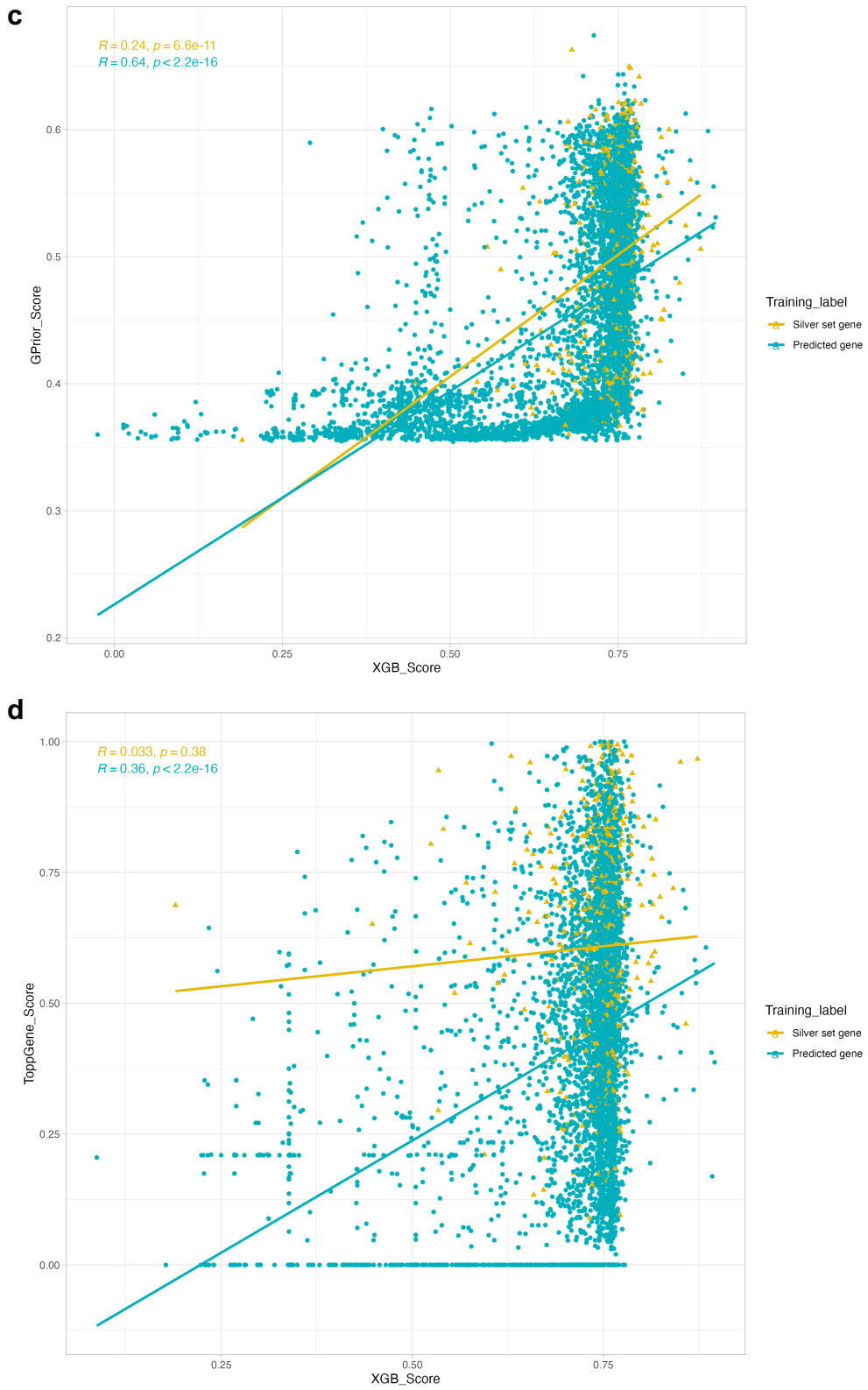


Figure 5.11. Gene prioritisation method comparison. XGB was plotted against four other methods comparing gene prediction; Overall association scores by OpenTargets

(**a**), Mantis-ml (**b**), ToppGene (**c**) and GPrior (**d**) prioritisation methods, investigating how each of the methods scored the lipid training genes (gold and silver standard set genes) and all other lipid genes (predicted genes) from the GWAS by Graham et al. (2021). For each plot, the score our XGB model gave to each gene is plotted along the x-axis in comparison to another method's prediction plotted on the y-axis, with the correlation (R) between the two methods calculated alongside the p-value significance of the R. The 21 gold set genes are coloured in red; the 723 silver set genes are coloured in yellow, and the 7,209 predicted genes are coloured in blue.

8.5 Discussion

Similar to the BP data analysed in chapters 2-4, the curated lipid training data showed genetic characteristics with minimised genetic bias risk. For example, similar to the EDA in chapter 2, only counts of epigenetic sites (CpG islands and methylation sites) had high correlations with gene length (Appendix D Table 3), leading to their removal of features. This result also suggests that the curation of counts of epigenetic sites per gene may not be beneficial to collect as ML features, when considering that these features did not pass gene length correlation in either ML framework application. When analysing gene length alone, the least likely genes curated and scored at 0.1 had the shortest genes – which also potentially impacted the ability to collect PPI data to further filter these genes, as shorter genes have been shown to have less PPIs[203].

The curation of the least likely genes was one of the few key aspects of the ML framework that were altered in this re-application. The change in gene curation was

due to the lack of genes that had all variants with large p-values (set to be greater than 0.05 to get the maximum number of genes) across all five blood lipid traits. Whilst analysing five blood lipid traits at once creates this limitation, making it less likely to have genes passing the filter for all five traits, it was also a necessity to apply the ML framework to all five blood lipid traits combined and not individual. This approach using five traits in a single model was chosen due to several reasons: the genetic and phenotypic correlation of all five phenotypes[197, 204], the gold and silver set genes not being curated from any evidence relating to individual blood lipid traits (but to lipid metabolism more generally), them being the same training examples that would be used in all five individual ML applications, and the Metal effect size feature (which would give the only unique data points per each trait for the gold and silver standard genes) not passing data cleaning. Furthermore, having one approach prioritises genes across all five traits at once enabled a more time-efficient approach, as opposed to running the framework five times.

The performance of the models benchmarked and the top-performing model, XGB, follows the model benchmarking trends discussed in chapter 4. However, GBM had the highest predicted r^2 on this re-application (0.938, Table 5.1) but despite this, XGB was chosen due to the 0.938 being an unexpectedly high increase in predicted r^2 (in comparison to GBM's 0.701 median r^2 and also in comparison to all other models' more conservative predicted r^2 measures). Notably, the algorithmic principles that underly gradient boosting are known to cause an overfitting risk. This is due to the gradient boosted tree optimising its tree over iterations on the training data, using residual errors from its previous trees - unlike for example random forest in which

each tree is trained on subsets of differing data divisions in parallel. Other gradient boosting models aim to minimise the overfitting risk in their design by having additional parameters (e.g., XGB has L1 and L2 regularisation and CB has ordered boosting). Also, the large class imbalance in the blood lipid data (21:723:60 for each of the training gene groups) increases the overfitting risk, requiring a more conservative approach. These reasons lead to XGB being selected with the model having the next best performance with a median r^2 of 0.707 and predicted r2 of 0.826, with all other metrics (mean squared error, explained variance, RMSE, and mean absolute error) having similar performances across gradient boosting ensemble models with less than 0.001 difference.

On exploring SHAP interpretation, it showed the gold standard genes were predicted based on the human Exomiser score followed by pLI measures as the most important features (Figure 5.5b). The use of these genetic intolerance and phenotypic features, similar to that of their importance in chapter 4, validates the re-application of the framework and shows the model highly prioritises genes that have more observed human phenotypes relating to blood lipids in Exomiser (with the human score calculated in Exomiser by the semantic similarity of related HPO terms) and have a higher probability of being loss-of-function intolerant genes. However, the positive SHAP values here are shown to be predominantly driven by the human Exomiser scores (Figure 5.5) suggesting the model could be being overpowered and biased by this one feature. Furthermore, as the Exomiser feature is a phenotypic measure, if used to make the majority of decisions in XGB it may be encouraging a redundancy in comparison to using the Exomiser score itself as a prioritisation measure. However,

similar to the BP GWAS application, HIPred and pLI are measured by SHAP as having the highest influencing interaction, followed by the Exomiser score with the amygdala gene expression (Figure 5.6), suggesting, that despite the large influence of the Exomiser score alone, model-decision making is informed by multiple factors.

The significance of the amygdala, as both its interaction with Exomiser scores and its position as the most important tissue feature for prioritising gold standard genes (Figure 5.5b) highlights its potential impacts on lipid metabolism that could be further investigated. For example, research has shown the amygdala acts to innervate interscapular brown tissue via its insulin receptors[205], which in turn regulates lipid metabolism. Furthermore, XGB also highly valued gene expression in pituitary and testis tissues, which are known to act together to regulate lipid metabolism and free circulating cholesterol in men[206, 207], alongside the pituitary gland also acting in the pituitary-adrenal axis which is known to be affected by a high-fat diet[208]. The importance of these tissues, similar to that of the tissues selected in chapter 4, suggests the XGB model prioritises genes acting in lipid-related signalling pathways as opposed to the sites of action. In contrast, several GTEx tissues were selected that have minimal model influence according to SHAP (e.g., lung, kidney, atrial appendage, spleen, cultured fibroblasts, ovary and the terminal ileum), suggesting the model and the computational efficiency could be further optimised with the removal of these features or that these features were only of notable benefit for certain genes. Six of the selected GTEx tissues were also in feature-feature correlating pairs with an $r^2 > 0.9$ (although their correlating feature counterpart was the removed feature), suggesting

these importance of these features needs further exploration that would ideally be using further testing data to confirm or deny their importance.

Also, from the selected features sex-specific tissues testis and ovary were selected, with the testis being the second most important GTEx feature in the training data (Figure 5.5). This aligns with lipid metabolism research that has identified sex differences in blood lipid traits and the roles played by sex-specific tissues such as testis in lipid metabolism[184, 185]. However, this also poses a bias risk in the ML framework, as the ovary TPM expression is not as highly valued by the model (although it is a selected feature), despite the ovaries also having a studied relationship to lipid metabolism[186]. This interpretation implies that the model, and its higher importance placed on testis gene expression, may have a preference for understanding sex-specific biological patterns. To further investigate this, I performed additional annotation of the model's gene ranking to indicate which genes are more likely to have biased sex-specific GTEx tissue expression (Appendix D Table 8). From the training genes, 126/744 genes were annotated as having sex-specific biased expression (across any of the tissue including those that were selected features), suggesting the ML model may be influenced by biased expression measures. The top prioritised gene, *KHK*, was annotated in GTEx as having sex-biased expression in 17 tissues including skeletal muscle tissue, which was a tissue focused on by Miller et al. (2018) who explored the gene's impact on fructose metabolism in knockout mice. Miller et al. (2018) found significant levels of fructose metabolites in the mice's skeletal tissue, suggesting *KHK* impacts fructose metabolism in skeletal muscle. However, whilst these findings link *KHK* to lipid regulation via fructose metabolism[187], the potential for biased sex-

specific expression as identified in GTEx indicates that caution is needed in follow-up research. Functional research focusing on sex-specific impacts of genes on lipid metabolism would also confirm whether the XGB model is truly identifying sex-specific biological patterns in its genes and selected features, which is difficult to ascertain from the model output alone.

Unlike the results of chapter 4, this model was more conservative in prioritising genes greater than 0.8, with the majority of genes being scored ~0.7. This result reflects the underlying larger class imbalance in comparison to that of the BP GWAS data, requiring more gold standard genes to overcome the lack of higher prioritisations. In total, the model prioritised 72 genes greater than 0.8, which on the Mann-Whitney U tests showed that these genes were more likely to be essential genes but that they had insignificant distributions in other metrics (RVIS, cell-type expression, and SDI drug probability). Furthermore, these genes also had predominantly insignificant enrichment for IMPC mouse phenotypes, with only 10/72 genes having a phenotype related to lipid or cardiovascular disease. Overall, a conservative model is beneficial as the higher-scored genes become a more refined list that may be easier to functionally investigate one by one. However, future research requires an increased size of gold standard genes to increase a ML model's understanding of what makes a most likely causal gene.

Downstream analysis validated the ML approach, highlighting supporting evidence for the highest-scored genes and their potential roles in lipid metabolism. Notably, *KHK*[187] and *CREB3L3*[188] have had direct lipid metabolism research. The top

prioritised gene *KHK*, for example, has knockout mouse model research showing the model has fructose intolerance that is known to impact lipogenesis, dyslipidemia, and insulin resistance[187]. Meanwhile, *CREB3L3* has had research into relation to hypertriglyceridemia, finding patients with a loss-of-function of the gene were significantly more likely to have severe hypertriglyceridemia[188]. Several of the other top 10 genes (*GLYCTK*[209], *SORD*[184], *SLC25A20*[189], *PCYT1A*[210], *UNC13D*[211], *ACOX2*[212]) have been involved in research related to lipid regulation but have not been the focus of any lipid metabolism study. For example, *GLYCTK* has been shown to be hypermethylated in rat models that were studied focusing on lead exposure leading to weight gain[209]. Furthermore, *UNC13D* mutations in patients with familial hemophagocytic lymphohistiocytosis have had clinical study showing they also have hypertriglyceridemia[211].

In total, 2,327 genes were selected per loci as most likely lipids genes with 2,312 of them having a model score greater than 0.5. The majority of *selected-genes* being scored greater than 0.5 again highlights the class imbalance using 723 silver set genes scored at 0.75 out of a total 804 training genes, however, it also highlights a limitation of the gene per locus selection algorithm as having a large pool of highly scored prioritised genes (with also fewer PPIs than the BP prioritisation which used collated GWAS' for 7,705 BP gene direct and secondary PPIs in total, versus 3,266 direct and secondary lipid gene PPIs found here) led to more opportunity for multiple genes to be selected per loci. In total, 343 loci had more than one gene selected, suggesting there is still further opportunity for refinement in these loci that ML and the selection algorithm could not provide. However, future work could overlay prioritisation

approaches, such as assessing prioritisation with other methods, e.g., DEPICT or PoPs as seen by Kanoni et al. (2021), which would enhance the gene selection algorithm developed here.

This would also optimise the *selected-gene* list that went into gene enrichment analysis. For example, the GTEx analysis showed only one cluster of 13 genes from the *selected-genes* with higher expression across 33 tissues than other *selected-genes*. Whilst the increased strength of their expression in lipid-related tissues suggests they may impact lipid regulation, the majority of these 13 genes encode ribosomal proteins and had only protein-translational roles identified in STRINGdb, suggesting their impacts are likely to be ubiquitous and their increased GTEx expression in 33 tissues may have only been identified by chance. Furthermore, the most significantly enriched pathways for the highest prioritised genes were known lipid metabolism pathways (e.g., cholesterol metabolism and PPAR signalling – Figure 4.5, Appendix D Table 10), also leading to fewer novel discoveries. However, further analysis of the genes inside these pathways may hold new insights. For example, the identification of *CD36,* which interacted with the top five enriched pathways and is also druggable (Figure 5.10), also interacts with the angiogenesis inhibitor and cancer drug ABT-510[190] and it has known roles in atherosclerosis and lipid metabolism[191], suggesting it may be a worthwhile target for further investigation in known lipid pathways.

When comparing the XGB to other methods, there were positive correlations for all gene predictions overall (Table 5.3, Figure 5.11), validating the re-application of this ML framework. For example, the OpenTargets overall association scores had the

highest positive correlations for any training gene group across blood lipid traits at 0.31 correlation for the gold standard genes (Table 5.3), however, this can only be compared with Mantis-ml which had a 0.24 correlation for the gold standard genes. In comparison to the prioritisation by Kanoni et al. (2021), the genes prioritised with high to low confidence also showed that genes with higher XGB prioritisation were more often assigned higher confidence by Kanoni et al. (2021). However, this was only with a slight difference in XGB scores (Table 5.4), emphasizing XGB's tendency to score genes ~0.7 and the effect of the underlying class imbalance influencing the model. Furthermore, XGB scored some genes that were labelled as having medium-low confidence with lower prioritisation scores than those with a low confidence assigned by Kanoni et al. (2021). This could indicate that XGB is not finding a great difference between the gene-characteristics of those confidence labels as clearly as in the other categories. For example, there is a clear difference between high and medium-high XGB scores where the minimum score for the high category only reaches 0.59, whilst the median and maximum prioritisations shift accordingly to their categorisation. Future work could combine both the medium-low and low-confidence groups into one category, using the evidence provided by both groups (e.g., DEPICT or TWAS that was used to assign medium-low confidence combined with the eQTL data used to assign low confidence) to have a multi-layered assessment of low confidence genes.

Investigating the genes focused on by Ramdas et al. (2021) (*RRBP1* and *CREBRF*) showed XGB scored *RRBP1* at 0.73 (while Kanoni et al. (2021) gave the gene a high confidence label), and XGB scored *CREBRF* at 0.76 while Kanoni et al. (2021) assigns the gene low confidence. *CREBRF* received its low confidence from eQTL data,

implying there may be a bias that is skewing prioritisation[183]. Overall, these differences between these methods and their output prioritisations emphasize the importance of comparing the prioritisation of any individual gene of interest to be sure their biological evidence is in strong enough agreement to warrant their follow-up experimentation.

The re-application of this ML framework also highlighted limitations with the method's general use. For example, the curation of training genes derived not from the GWAS itself prevented the use of GWAS summary statistics, such as effect size, to be used as a feature. Furthermore, the least likely gene curation was limited by a minimal number of genes having large p-values (caused in part by requiring this to be the case for five blood lipid traits). This challenge was compensated for with the removal of their further PPI curation, which potentially weakens the strength of these genes being truly least likely to affect blood lipid traits. In addition, the collected annotations from 20 databases, the benchmarking of fourteen models, and the size of the training data being more than double that of the BP GWAS, made the framework more computationally expensive – making the framework less accessible. The data size and time to run also led to regression scoring intervals not being tested as it was in chapter 4, with the gene scores being set to 1.0, 0.75, and 0.1 with no further scoring tests. This scoring interval was chosen due to its successful performance in chapter 4; however, further comparative tests are still needed to confirm the ML framework is optimised to this blood lipid dataset. Additionally, the gold standard and silver standard training genes provided gene curations for efficient re-application of the ML framework, however they were not curated with ML in mind, with the silver set

providing a large class imbalance in a ML context - suggesting future work would need to refine this gene group to optimise the ML performance.

Overall, this re-application of the ML framework developed in this thesis highlights the potential of the methodology to act as a disease-agnostic framework that develops automated disease-specificity as part of its pipeline (by specified training gene curation and collection of phenotypic features using relevant search terms in tools such as Exomiser) to successfully prioritise most likely causal genes. The re-application to blood lipid traits also indicates novel genes with potential roles in lipid metabolism, that are further verified by the supporting prioritisation evidence gathered in previous studies[183, 192]. However, the limitations of this re-application (quality of least likely gene curation, class imbalance, computational efficiency, etc.) add to the challenges already faced by ML applied to gene prioritisation and emphasize that re-application of this ML framework needs to be applied with care.