Modelling human complex traits with regression and neural-network based methods

Marton Kelemen

Wellcome Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Darwin College November 2020

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee. This dissertation contains fewer than the limit of 60,000 words set by the Biology Degree Committee.

Marton Kelemen
November 2020

Modelling human complex traits with regression and neural-network based methods

Marton Kelemen

Identifying how epistasis, non-linear genetic effects, contribute to phenotypic variance in humans has been an enduring challenge. So far neither the computational resources that could accommodate higher-order interactions at scale nor the large-scale population cohorts with adequate statistical power were available up until recently. With the advent of graphics processing unit computing farms and neural-network based methods, together with large biobank-scale data sets, such as the UK Biobank which offers a sample size of ~500K, this has been changing. These developments offer opportunities for the development of novel approaches that could provide insights into the genetic underpinnings of complex disease risk and trait variation.

After reviewing the necessary background material, this work consists of three research chapters. The organising theme of these is the building of genotype-phenotype maps, which grow from the simple additive, through the two-way interactions, up to higher-order interactions in the last chapter.

I begin by covering the common quality control steps and basic additive association analyses I carried out that explored the information boundaries of my data which serves as the foundation for the rest of my work. I managed to recover primary association signals described in the literature for my cohorts confirming the validity of my data processing steps. I also describe a novel method that exploits shared genetic effects to improve risk prediction for related traits. Relative to baselines, this improved squared correlations between observed and predicted sub-phenotypes by ~25% and ~19% for ulcerative colitis and Crohn's disease, respectively.

Building on the previously prepared data sets, I searched for two-way interactions using standard statistical methods belonging to the regression framework. In the UK Biobank cohort I pursued a hypothesis-free approach to consider interactions both within and between the genomic domains of SNP, transcription and protein derived predictors. For the much smaller inflammatory bowel disease studies, I followed a hypothesis driven strategy to reduce search space which only considered haplotype-specific interactions between biologically plausible loci to increase power. I found that the results from both of these approaches were consistent with the null hypothesis of no significant contribution to phenotypic variance from non-linear genetic effects.

Parallel to my search for epistasis using regression based models, I also considered the neural-network framework to find indirect evidence for non-linear effects contributing to phenotypic variance. I confirmed via a large-scale simulation study the potential of neural-networks to be able to identify interactions at a higher accuracy than standard regression based methods. In the real datasets, I searched for individual epistatic interactions using both experimental approaches from the literature, together with methods that I developed for this purpose. However, I was unable to find convincing evidence for statistical interactions contributing to complex trait variance.

In summary, I found that despite the large cohorts I had access to and the modern non-linear methods I deployed, evidence for non-linear genetic effects contributing to complex human trait variance remained elusive.

Acknowledgements

I thank my supervisors Carl Anderson and Chris Wallace, without whom this work would not exist. Thank you for putting up with my stubbornness, I hope that my future work will make you proud. I am immensely grateful for the opportunity to pursue a research topic of my choice, which was made possible by the Wellcome Trust who supported my work financially, and the Mathematical Genomics and Medicine administrators who organised this exceptional programme. I also thank Nadav Brandes for kindly agreeing to share the data and provided help to navigate the PWAS method.

This thesis is dedicated to my mother, Judith Dimitrova, and my father, Lajos Kelemen.

Table of contents

List of figures xvii

List of tables xxi

Nomenclature xxv

1 Introduction 1

1.1 Non-linear encoding of genetic information 1

1.1.1 Epistasis 2

1.1.2 The two main forms of epistasis 3

1.1.3 The importance of epistasis in understanding biology 5

1.1.4 Examples of statistical epistasis in humans 5

1.1.5 Variance component analyses of epistasis 6

1.1.5.1 An alternative explanation to the apparent lack importance of epistatic variance 7

1.1.5.2 Is non-linear population genetic variance needed for non-linear information encoding? 8

1.1.6 Challenges of statistical epistasis detection 9

1.1.6.1 Statistical and computational challenges 9

1.1.6.2 Linkage Disequilibrium 10

1.1.6.3 Thresholding effects for traits with a limited recorded range 12

1.1.7 General approaches to epistasis detection 12

1.2 Heritability 14

1.2.1 Genetic prediction and heritability 15

1.2.2 Overview of methods that estimate variance components 16

1.3 Genome-wide association studies 17

1.3.1 GWAS background 17

1.3.2 GWAS framework 18

1.3.2.1 GWAS quality-control 18

1.3.2.2 The GWAS model and statistical considerations 20

1.3.3 GWAS insights and recent trends 21

1.4 Transcriptome-wide association study 22

1.4.1 TWAS framework 22

1.4.2 The potential benefits of the TWAS framework 23

1.4.3 Limitations of TWAS 23

1.5 Protein burden score tests 24

1.5.1 Protein burden test method outline 24

1.5.1.1 Generating the protein burden scores 24

1.5.1.2 Protein burden association tests 25

1.5.2 Potential benefits of the protein burden test 25

1.6 Genetic risk prediction 25

1.6.1 Polygenic scores 25

1.6.2 The origin of PRS 26

1.6.3 Current methods for building PRS 26

1.6.3.1 Univariate regression based models 27

1.6.3.2 Whole-genome regression based models 28

1.6.3.3 LDpred 30

1.6.4 Recent applications of PRS 32

1.6.4.1 Limitations 33

1.6.5 Genetic prediction incorporating non-additive effects 34

1.7 Neural-network based methods 35

1.7.1 The origins of neural-networks 35

1.7.2 What are neural-networks exactly? 36

1.7.3 How are neural-networks fit? 38

1.7.3.1 Stochastic gradient descent 38

1.7.3.2 Weight initialisation 39

1.7.4 Advanced neural-network concepts 39

1.7.4.1 ADAM optimizer 40

1.7.4.2 Layers that address the vanishing gradient problem . . . 40

1.7.4.3 ReLU and batch normalization 41

1.7.4.4 SELU 41

1.7.4.5 Regularization of neural-networks 42

1.7.4.6 L1 and L2 norms 42

1.7.4.7 Early stopping 42

1.7.4.8 Dropout layer 42

1.8 Thesis objectives 43

2 Additive models and common quality-control steps 45

2.1 Chapter 2 outline 45

2.2 Datasets 45

2.2.1 Overview of the phenotypes considered 45

2.2.1.1 UK Biobank traits: height, BMI, fluid intelligence and asthma 46

2.2.1.2 IBD and its subphenotypes 48

2.2.2 UK Biobank genotype and phenotype data diagnostics 49

2.2.3 IBD datasets 50

2.3 Quality Control 51

2.3.1 Common quality control steps 51

2.3.1.1 Converting genotype probabilities to hard calls 51

2.3.1.2 Post-imputation quality-control for the UKBB genotypes 52

2.3.1.3 Post-imputation quality-control for the IBD genotypes . . 53

2.3.1.4 Phenotype quality control 54

2.3.1.5 Further filtering of genotypes for the TWAS and protein burden score tests 55

2.4 Experimental setup for later analyses 56

2.4.0.1 Cohort organisation in the UK Biobank 56

2.4.0.2 Dataset organisation for the IBD datasets 56

2.5 Additive association tests 57

2.5.1 GWAS 57

2.5.1.1 Post-association QC 57

2.5.1.2 UKBB association test results 61

2.5.1.3 IBD association test results 61

2.5.2 Summary of the additive association experiments 63

2.6 Leveraging shared genetic effects to improve genetic risk prediction for IBD 66

2.6.1 Establishing baselines 66

2.6.2 Estimating SNP heterogeneity of effect in the IBD studies 67

2.6.3 Finding the balance between the subphenotypes and IBD 68

2.6.4 Results for predicting IBD subphenotypes 71

2.6.5 Discussion of the improved IBD subphenotype PRS 71

3 Regression based models of statistical epistasis 73
3.1 Chapter 3 outline 73
3.2 Dimensionality reduction in the UKBB 73
3.2.1 Transcriptome and protein score data-sets 74
3.2.1.1 FIRM protein scores 74
3.2.1.2 BLUEPRINT transcriptome data 74
3.2.2 TWAS for asthma in the UKBB 75
3.2.2.1 Imputing the transcriptome 75
3.2.2.2 Expression association to the phenotype 76
3.2.2.3 UKBB asthma TWAS dimensionality reduction results . 76
3.2.3 Protein burden score tests in the UKBB 77
3.2.3.1 Protein burden score dimensionality reduction results . . 77
3.2.4 Filtering the protein burden and gene expression scores 78
3.2.5 GWAS data 79
3.3 Interaction tests 79
3.3.1 Post-association QC 80
3.3.2 Interaction test results 81
3.4 Cross-domain interaction tests 86
3.4.1 Cross-domain filtering 86
3.4.1.1 Gene filter for asthma TWAS and protein burden scores . 86
3.4.1.2 SNP-Gene cross-filtering 86
3.4.2 Cross-domain interaction results 87
3.4.3 Summary of the UKBB interaction test experiments 88
3.5 Interaction tests in the IBD datasets 89
3.5.1 Biological insight to reduce search-space 90
3.5.2 Statistical haplotype phasing 90
3.5.2.1 The definition and the utility of haplotype phase 90
3.5.2.2 Overview of phasing methods 90
3.5.2.3 Statistical Methods 91
3.5.2.4 Trio and pedigree based phasing 91
3.5.2.5 Hidden markov model based phasing 92
3.5.2.6 Phasing summary 93
3.5.3 Genotype and summary data 93
3.5.3.1 Collating summary statistics for IBD 94
3.5.3.2 Obtaining haplotype configurations 94
3.5.4 Two statistical models to evaluate haplotype-specific interaction effects 95

3.5.4.1 '#Bad haplo' model 95
3.5.4.2 'haplo regression' model 96
3.5.4.3 Results for the haplotype-specific interaction tests 96
3.5.4.4 Post-association QC and discussion of haplotype-specific interaction tests 97
3.6 Concluding remarks 98

4 Prediction and inference on non-linear genetic effects using neural-networks 101
4.1 Chapter 4 outline 101
4.2 Neural-networks in genomics 101
4.2.1 Relevant previous work 101
4.2.2 Opportunities and challenges for NNs in genomics 104
4.2.3 Neural-network models and data preparation 105
4.2.3.1 Choosing the model architecture 105
4.2.4 NN methods used in this chapter 107
4.2.4.1 Using NNs to evaluate the evidence for non-linearity . . . 107
4.2.5 Inference via neural-networks 108
4.2.6 Overview of my NN inference strategy 108
4.2.6.1 Uncertainty estimation via dropout 109
4.2.6.2 Estimating the importance of input features 110
4.2.6.3 Examining the learned weights of the network directly . . 110
4.2.6.4 My NID implementation 111
4.2.6.5 Inference-via-prediction 111
4.2.6.6 My inference-via-prediction implementation 112
4.2.6.7 Common search space reduction strategy 114
4.2.6.8 OLS baseline 115
4.3 Simulation experiments on synthetic data 115
4.3.1 Genotype dataset 115
4.3.2 Phenotype simulation details 115
4.3.3 Prediction results 118
4.3.4 Inference results 119
4.3.5 Discussion of the simulation experiments 121
4.3.5.1 Prediction performance 124
4.3.5.2 Inference performance 125
4.4 Neural-network tests on real data 127
4.4.1 Data preparation and model selection 127
4.4.2 Prediction results on real data 128

4.4.3 Inference results on the asthma cross-domain data 128
4.4.4 Summary and limitations 130
4.4.5 The outlook of NNs for building PRS 132

5 Conclusion 137
5.1 Overview and limitations 137
5.2 Reflections on non-linear genetic effects 138
5.3 Outlook and future work 140

Appendix A Simulation results supplementary 189

Appendix B Neural-network supplementary 191
B.0.0.1 Convolutional neural-networks 191

List of figures

1.1 **Hypothetical genotype matrix of n individuals at p loci.** Functional epistasis can take place between any of the p loci. However, statistical epistasis can only take place between loci 1, 3 and p, which are SNPs. The two loci, 2 and 4, do not vary in a population, therefore interactions between them, or even between these and SNPs, cannot contribute to phenotypic variance. Figure and terminology adapted from Angermueller et al. (2016). 4

1.2 **Illustration of the haplotype effect as an artefact generator for statistical epistasis.** The two SNPs imperfectly tag ($r^2 > 0$) the untyped SNP3, which is the causal variant affecting the phenotype Y. This pattern can arise even if the r^2 between SNP1 and SNP2 is zero. Statistical epistasis may only be generated by this pattern if the SNP1-SNP2 haplotype is a better tag for SNP3 than either SNPs on their own. 11

2.1 **Distributions of the three quantitative phenotypes in the UKBB**. Height, body mass index (BMI) and fluid intelligence score (FIS). 47

2.3 **Manhattan plot visualising the GWAS1 study without applying post-association QC to consider LD patterns.** There are many associations above the genome-wide significance level with no LD structure to support them, a property that marks them out as potential false positives. 58

2.4 **Four examples that illustrate common cases where the application of the automated filtering either eliminated potential false positive associations, or alternatively, retained those consistent with the nearby signal.** 60

2.5 **Manhattan plots visualising the UKBB GWAS.** y-axis shows the $-log10$ of the additive association p-values and the x-axis displays the genomic coordinates. The red line represents the genome-wide significance level of $5 * 10^{-8}$. 61

2.6 **Manhattan plot visualising the GWAS3 dataset IBD association result.** y-axis represents the $-log10$ of the additive association p-values and the x-axis displays the genomic coordinates. The red line represents the genome-wide significance level of $5*10^{-8}$. 62

2.7 **Manhattan plots visualising the IBD, CD and UC GWAS.** y-axis represents the $-log10$ of the additive association p-values and the x-axis displays the genomic coordinates. The red line represents the genome-wide significance level of $5*10^{-8}$. 62

2.8 **Plots comparing the GWAS z-scores of my results against relevant studies in the literature.** x-axis ('Thesis zscore') represents the z-scores from my analyses, and the y-axis represents z-scores for the same variants I obtained from reference studies in the literature. 65

2.9 **Manhattan visualising the adjusted Q values that measured SNP heterogeneity of effect between CD and UC.** Left y-axis shows the adjusted Q-values and right y-axis shows 1-lFDR. x-axis represents genomic coordinates. 69

2.10 **Dot-plots for the IBD subphenotype compose PRS hard threshold experiments.** y-axis represents the r^2 between the predicted and observed phenotypes. The dots represent bootstrap samples and the coloured bar is the mean across all bootstrap samples. The grey dotted line represents the mean across all experiments. The suffix after each plot's name indicates the lFDR threshold used to swap between subphenotype and IBD SNP summary statistics. 70

2.11 **Dot-plots for the IBD subphenotype composite and blended PRS experiments.** y-axis represents the r^2 between the predicted and observed phenotypes. The dots represent bootstrap samples and the coloured bar is the mean across all bootstrap samples. The naming convention is as follows. The first line of each PRS represents the target phenotype on which the PRS was evaluated on and the second line represents the source on which the PRS was trained on. For example, "*predicted: CD trained: Blend*" is the PRS that was evaluated on the CD phenotype and was trained using the blended PRS approach. 71

3.1 **Manhattan plots visualising all three tissues in the UKBB asthma TWAS.** y-axis represents the $-log10$ of the additive association p-values. x-axis shows the genomic coordinates. Red line represents the (Bonferroni corrected) genome-wide significance level of $5*10^{-6}$. 77

3.2 **Manhattan plots visualising the PWAS test results for the four UKBB traits.** y-axis represents the $-log10$ of the additive association p-values. x-axis shows the genomic coordinates. Red line represents the $-log10$ p-value threshold of $5*10^{-6}$. 78

3.3 **QQ-plots visualising the p-values of the two-way interaction term for the height SNP and protein burden score domain and asthma SNP domain.** Grey area represents 95% confidence intervals. 81

3.4 **QQ-plots visualising the p-values of the two-way interaction term for the post-QC analyses for the four UKBB traits in the SNP domain.** Grey area represents 95% confidence intervals. 83

3.5 **QQ-plots visualising the p-values of the two-way interaction term for the post-QC analyses for the four UKBB traits in the protein score domain.** Grey area represents 95% confidence intervals 84

3.6 **QQ-plots visualising the p-values of the two-way interaction term for the post-QC analyses for the asthma phenotype in the TWAS domain.** Grey area represents 95% confidence intervals 85

3.7 **QQ-plots visualising the p-values of the two-way interaction term for the four UKBB trait cross-domain analyses.** Grey area represents 95% confidence intervals. 88

3.8 **Missense-eQTL schematic diagram** Top: Illustration of the haplotype-specific interaction effect between a missense variant and a cis-regulatory SNP. In the deleterious haplotype configuration, the missense and the eQTL upregulatory alleles are on the same chromosome which results in an increase of the faulty gene product. Bottom: a benign haplotype configuration, where the hypothetical individual carries the same alleles, but not on the same chromosome, which would result in a greater abundance of the normal gene product. 91

3.9 **Hypothetical example for a phenotype column vector and design matrix for the '*#bad haplo*' model for *n* individuals.** Intercept omitted for clarity but was present in the model fit. 95

3.10 **Hypothetical example for a phenotype column vector and design matrix for the haplotype regression model for *n* individuals.** Intercept omitted for clarity but was present in the model fit. 96

4.1 **Neural-network performance on obtaining non-linear solutions under varying conditions for the experiments with a causal fraction of 0.25 of SNPs involved in statistical epistasis.** x-axis represents the % of sample size used and y-axis represents the r^2 of predicted vs observed phenotypes on the test set. Facets display experiments of genetic architectures that involve either additive, second, third and fourth-order interactions. 120

4.2 **Neural-network performance on obtaining non-linear solutions under varying conditions for the experiments with a causal fraction of 0.95 of SNPs involved in statistical epistasis.** x-axis represents the % of sample size used and y-axis represents the r^2 of predicted vs observed phenotypes on the test set. Facets display experiments of genetic architectures that involve either additive, second, third and fourth-order interactions. 121

4.3 **The performance of the three evaluated algorithms for statistical epistasis detection for the fourth-order series of experiments for the five sample sizes evaluated.** The average AUCs for OLS, NID and NNPred are shown blue, purple and green, respectively. *n* is the number of experiments from which the curves were drawn from. 122

4.4 **NN performance in the six experiments where a non-linear solution was preferred.** Results given in the format of *phenotype - domain*. y-axis represents r^2 of predicted vs observed phenotypes on the Test Set. For CD and UC the Test Sets were GWAS1 and GWAS2, respectively. 134

4.5 **Diagnostic plots for the asthma cross-domain experiments for putative interaction pairs involving gene-level predictors.** Red and blue represent cases and controls, respectively. 135

A.1 **NN performance on obtaining non-linear solutions under varying conditions for the experiments with a causal fraction of 0.5 of SNPs involved in statistical epistasis.** x-axis represents the % of sample size used and y-axis represents the r^2 of predicted vs observed phenotypes on the test set. Facets display experiments of genetic architectures that involve either additive, second, third and fourth-order interactions. 190

List of tables

2.1 **Correlations between the four occasions the FIS UKBB phenotype was recorded.** 'time1', 'time2' and 'time3' are the three different time points where the participants were assessed via in-person tests. 'online' represents the online follow-up test. 49

2.2 **UK Biobank summary of phenotypes.** 'SNP' h^2 is the LDSC estimated SNP heritability, 'Neff' is the effective sample size. Data was obtained from the Neale lab's 'SNP-Heritability Browser' online service from https://nealelab.github.io/UKBB_ldsc/index.html, accessed on 01/03/2020. 50

2.3 **Platform and study size details for the three IBD datasets.** 'GWAS1', 'GWAS2' and 'GWAS3' refer to the WTCCC1, WTCCC2 and the internal GWAS dataset, respectively. 51

2.4 **List of significant covariates for both the UKBB and IBD datasets.** Covariates were selected by a two stage backward selection process to be considered for each dataset and phenotype combination. 55

2.5 **The number of individuals in the various data splits for each experiment for the UKBB phenotypes.** The validation set sizes are shown as approximate, as the number of unique individuals not sampled into the training set varied slightly in each bootstrap sample due to the random nature of the resampling process. 57

2.6 **Landmark associations for my IBD analyses.** Comparisons of associations between the GWAS3 dataset and the study by de Lange et al. (2017). 'de Lange p' is the p-value from the de Lange et al. study, and 'chrom' indicates the chromosome. 63

2.7 **The four traits I selected for a quantitative comparison against reference studies from the literature.** The values in the correlation column are Pearson correlation coefficients between the z-scores from my association results and those of the literature. The values in the column 'correlation' ($p < 5 * 10^{-8}$), are Pearson correlation coefficients computed between z-scores that were restricted to have an additive association $p < 5 * 10^{-8}$. . . 64

3.1 **Summary of the number of predictors and interaction tests performed in the UKBB cohort.** The columns 'pre/post filtering' display the number of SNPs or PWAS scores pre and post LD filtering out of the total number of < 0.05 FDR corrected predictors. The 'number of tests' columns show the total number of interaction tests performed post-filtering using either the SNPs or the protein burden scores. 79

3.2 **Summary of the number of TWAS scores and interaction tests performed for the asthma phenotype.** The column 'pre/post filtering' displays the number of TWAS scores pre and post LD filtering out of the total number of $FDR < 0.05$ corrected predictors. The 'number of tests' column shows the total number of interaction tests performed post-filtering. 80

3.3 **Summary of post-QC results for the two-way interaction tests for all four UKBB phenotypes for both SNP and protein scores.** The 'minimum FDR' column represents the lowest FDR observed in a given experiment, and the 'number of tests' column displays the total number of tests performed. 82

3.4 **Summary of post-QC results for the three TWAS tissues for the asthma phenotype** The 'minimum FDR' column represents the lowest FDR observed in a given experiment and the 'number of tests' column displays the total number of tests performed. 82

3.5 **Summary of the model terms of the linear regression between SNPs rs117290331 and rs115122203 for the asthma phenotype.** Values in the 'beta' column represent the regression coefficient. 82

3.6 **Genotype count tables for the asthma phenotype for cases and controls.** The values in parentheses are proportions. 85

3.7 **Summary of the cross-domain filtering process.** 87

3.8 **The results of the cross-domain two-way interaction tests for all four UKBB phenotypes.** The 'minimum FDR' column shows the lowest FDR observed in a given experiment, and the 'number of tests' column displays the total number of tests performed. 87

3.9 **Results for the two-way interaction tests between the missense and eQTL SNPs for both the '*haplo regression*' and '*#bad haplo*' models.** Values in the 'p' column show association p-values for the haplotype-specific interaction term and values in the 'coef' column show their corresponding coefficient estimates. 97

4.1 **Summary of the differences between typical image classification and genetic prediction tasks.** n is the number of observations and p is the number of input features. H^2 is broad-sense heritability. *Accuracies taken from He et al. (2016) and Lee et al. (2018) for images and PRS, respectively. 105

4.2 **Summary of the search-space covered by the hyperopt tool.** The SELU activation function is described in the Introduction in section 1.7.4.4. . . . 106

4.3 **Fractions of experiments where a non-linear solution was found by the NN out 100 simulations with a causal fraction of 0.25 of SNPs involved in statistical epistasis.** The values in the 'additive' column represent experiments where the ground truth genetic architecture was purely additive. . . 118

4.4 **Fractions of experiments where a non-linear solution was found by the NN out 100 simulations with a causal fraction of 0.95 of SNPs involved in statistical epistasis.** The values in the 'additive' column represent experiments where the ground truth genetic architecture was purely additive. . . 119

4.5 **Inference results for the simulation experiments for the three methods (NNPred, NID and OLS) at different percentages of the total sample size.** '*AUC*', '*SE*' and '*n*' denote the area under the curve, its standard error and the number of experiments the preceding values were calculated from, respectively. Values under 'all results' represent inference results from all 100 experiments. In case a method did not report a result, its accuracy was substituted by an AUC of 0.5. Values under 'successful results' represent inference results conditioned on individual methods successfully reporting a result. Values under 'intersection results' represent inference results conditioned all three methods reporting a result. Bold text highlights the best method in a given scenario. 123

4.6 **Comparison between the significance metrics of the NN and standard statistical methods for the variants identified as potentially interacting.** $\mathbf{NN_{IS}}$ is the importance score produced by the NID algorithm (arbitrary scale). $\mathbf{p_{train}}$ is the raw interaction p-value from Chapter 3 that considered all predictors which survived the filtering process in the Training Set. $\mathbf{p_{test}}$ is the raw interaction p-value for the same pairs in the Test set. $\mathbf{p^{cases}_{testCorr}}$ is the p-value of the correlation between the predictors in the cases only test in the Test Set. 130

4.7 **Diagnostic statistics for each variant potentially involved in interactions.** The values in parentheses in the 'MAF' columns are the standard error of the mean. 131

4.8 **Training Set genotype fraction tables for the asthma phenotype for cases and controls for the three putative two-way interactions that involved SNP pairs.** 132

A.1 **Fractions of experiments where a non-linear solution was found by the NN out 100 simulations with a causal fraction of 0.5 of SNPs involved in statistical epistasis.** The values in the 'additive' column represent experiments where the ground truth genetic architecture was purely additive. . . 189

Nomenclature

Acronyms / Abbreviations

AUC	Area under the curve
BMI	Body mass index
CD	Crohn’s disease
CNN	Convolutional neural-network
FDR	False discovery rate
FIS	Fluid intelligence score
FNN	Fully-connected neural-network
GPU	Graphics processing unit
GWAS	Genome-wide association study
HLA	Human leukocyte antigen
HMM	Hidden Markov model
HWE	Hardy–Weinberg equilibrium
IBD	Inflammatory bowel disease
KRR	Kernel-ridge regression
LD	Linkage disequilibrium
LMM	Linear mixed-effects model
MAF	Minor allele frequency

NN	Neural-network
OLS	Ordinary least squares
PRS	Polygenic risk score
QC	Quality control
REML	Restricted maximum likelihood
RKHS	Reproducing kernel Hilbert space
ROC	Receiver operating characteristic
RR	Ridge regression
SNP	Single-nucleotide polymorphism
T1D	Type 1 diabetes
TF	Transcription factor
TWAS	Transcriptome-wide association study
UC	Ulcerative colitis
UKBB	UK Biobank
WGS	Whole-genome sequencing