LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT
TECHNISCHE UNIVERSITÄT MÜNCHEN TUM

Institute of Computational Biology
Helmholtz-Zentrum Munich

Masterarbeit

in Bioinformatik

Machine learning for predicting type 1 diabetes with high-throughput data

Angermüller, Christof

Aufgabensteller:	Prof. Dr. Dr. Fabian Theis
Betreuer:	Dr.Florian Büttner & Dr. Jan Krumsiek
Abgabedatum:	15.09.2013

Ich versichere, dass ich diese Masterarbeit selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel verwendet habe.

15.09.2013 _____

Angermüller, Christof

Abstract

Type 1 diabetes (T1D) is characterised by an autoimmune destruction of insulin producing pancreatic beta cells with varying progression rate. At the point when disease symptoms become evident and T1D is clinically diagnosed, most beta cells are already destroyed. Insulin replacement therapy is then often the only remaining treatment to counter the lacking insulin production. Early detection of rapid progression to T1D is thus of great medical interest.

The first aim of this thesis was to employ established machine learning methods to improve predicting the risk of T1D with genotyping data, and to quantify the effect of genes on T1D risk. The second aim was to use proteomics data for discriminating between rapid and slow progression to T1D, and to discover potential biomarkers.

In the first half of this thesis, I studied a genomics dataset with 41 genetic risk factors of T1D positive and T1D negative subjects. I fitted classification models, which I evaluated by the area under the ROC curve (AUC) on an independent test set. I tested for gene-gene interactions and employed feature selection to rank genes by their usefulness for classification. A logistic regression model fitted with all 41 genetic risk factors and no gene-gene interactions achieved the highest AUC of 0.84. Most genes had a small effect on T1D risk, but improved prediction in combination.

In the second half of this thesis, I studied a proteomics dataset with 4384 blood serum peptides of 15 rapid T1D progressors and 15 slow T1D progressors. I employed feature selection to rank peptides by their usefulness for predicting T1D, and evaluated classification performance by the cross-validation AUC. Only four peptides perfectly discriminated between slow and rapid progressors. For several highly discriminative peptides, recent studies reported a critical role in the cardiovascular system and the innate immune system.

My results can help to develop new prognostic tests using genotyping data, as well as diagnostic tests using blood serum biomarkers. Cost-effective high-throughput genotyping as well as non-invasive, inexpensive, biomarkers, can allow large-scale screening of newborns and early detection of high-risk children for personalised prevention therapies.

Zusammenfassung

Typ 1 Diabetes (T1D) zeichnet sich durch eine kontinuierliche, unterschiedlich schnell fortschreitende, Zerstörung von Insulin-produzierender beta Zellen der Pankreas aus. Wenn Krankheitssymptome zum Vorschein treten und T1D klinisch diagnostiziert werden kann, ist bereits ein Großteil der beta Zellen irreversibel zerstört. Schnell fortschreitendes T1D früh zu erkennen, ist daher von großer medizinischer Relevanz.

Das erste Ziel dieser Arbeit war der Einsatz von Methoden des maschinellen Lernens, um die Vorhersage von T1D mit Genotypisierungsdaten zu verbessern, und um den Einfluss verschiedener Gene auf das Entstehungsrisiko von T1D zu quantifizieren. Das zweite Ziel war die Verwendung von Proteomdaten, um zwischen schnellem und langsamen T1D zu unterscheiden, und um potentielle Biomarker zu entdecken.

In der ersten Hälfte dieser Arbeit habe ich einen Genomdatensatz mit 41 genetischen Risikofaktoren von T1D positiven und T1D negativen Patienten ausgewertet. Ich habe Klassifikationsmodelle optimiert, und diese auf einem unabhängigen Test Datensatz anhand der Fläche unter der ROC Kurve (AUC) evaluiert. Ferner habe ich auf mögliche Gen-Gen Interaktionen geprüft, und Methoden zur Feature Selektion verwendet, um Gene nach ihrer Nützlichkeit zur Klassifikation zu ordnen. Ein logistisches Regressionsmodell, das ich mit 41 genetischen Risikofaktoren ohne Einschluss von Gen-Gen Interaktionen optimiert habe, hat die höchste AUC von 0.84 erzielt. Die meisten Gene beeinflussten das Risiko von T1D nur geringfügig, konnten allerdings die Vorhersagegenauigkeit verbessern.

In der zweiten Hälfte dieser Arbeit habe ich einen Proteomdatensatz mit 4384 Peptiden ausgewertet, welche für Patienten gemessen wurden, die T1D schnell oder langsam entwickelten. Ich habe Methoden zur Feature Selektion verwendet, um Peptide nach ihrer Nützlichkeit zur Vorhersage von T1D zu ordnen. Die Klassifikationsgenauigkeit habe ich anhand der kreuzvalidierungs AUC gemessen. Mit nur vier Peptiden konnte vollständig zwischen Patienten unterschieden werden, die T1D schnell oder langsam entwickelten. Für mehrere Peptide wiesen aktuelle Studienergebnisse auf eine wichtige Rolle im Herz-Kreislauf-System und im angeborenen Immunsystem hin.

Meine Ergebnisse können dazu beitragen, neue prognostische Tests basierend auf Genotypisierungsdaten zu entwickeln, als auch diagnostische Tests basierend auf Blutserum Biomarker.

Acknowledgements

First and foremost, I would like to express my sincere thanks to Prof. Dr. Dr. Fabian Theis for giving me the opportunity to work on this intriguing topic, and to experience a new excellent research group at the end of my Master's studies.

I would like to thank my supervisor Dr. Florian Büttner, who supported me throughout this thesis, and whose expertise in machine learning and statistics helped me to learn about new methods for data analysis.

Likewise, I would like to thank my supervisor Dr. Jan Krumsiek, who never ran out of questions, and whose excellent background in Bioinformatics helped me to identify pertinent computational methods for addressing biomedical problems.

I would like to thank Dr. Christiane Winkler from the Institute of Diabetes Research at the Helmholtz-Zentrum Munich for providing the genomics dataset and for supporting me from a biological point of view.

I would also like to thank Dr. Christine von Törne from the Research Unit Protein Science at the Helmholtz-Zentrum Munich for providing the proteomics dataset, for giving me an introduction to mass-spectrometry, and for answering all my questions.

Finally, I would like to thank my family for their love and ongoing support throughout my studies and forgiving me for not finding enough time to return home in the course of this thesis.

Contents

Abstract I

Zusammenfassung II

Acknowledgements III

List of Figures VI

List of Tables VIII

1 Introduction 1

2 Type 1 diabetes 3

3 Methods 6

3.1 Models for classification . . . 6

3.1.1 Logistic Regression . . . 6

3.1.2 LASSO . . . 8

3.1.3 Support vector machine . . . 8

3.1.4 Random Forest . . . 10

3.2 Evaluation . . . 11

3.2.1 Performance measures . . . 11

3.2.2 Performance estimation . . . 12

3.3 Feature selection . . . 13

3.3.1 FSCORE . . . 15

3.3.2 SVM RFE . . . 15

3.3.3 MSVM RFE . . . 17

3.3.4 LASSO-R . . . 18

3.3.5 RF-R . . . 19

3.3.6 VARBVS . . . 19

4 Genomics dataset 22

4.1 Introduction . . . 22

4.2 Dataset . . . 22

4.2.1 Training set . . . 23
4.2.2 Test set . . . 23
4.2.3 Features . . . 23
4.3 Results . . . 24
4.3.1 HLA model . . . 24
4.3.2 HLA/non-HLA model . . . 25
4.3.3 non-HLA model for high-risk subjects . . . 28
4.3.4 Interactions . . . 30
4.3.4.1 HLA interactions . . . 30
4.3.4.2 HLA/non-HLA interactions . . . 32
4.3.5 Feature selection . . . 32
4.4 Discussion . . . 34

5 Proteomics dataset **38**
5.1 Introduction . . . 38
5.2 Dataset . . . 39
5.3 Results . . . 39
5.3.1 Classification without feature selection . . . 40
5.3.2 Classification with feature selection . . . 41
5.3.3 Biochemical function of selected proteins . . . 43
5.3.3.1 Proteins of the cardiovascular system . . . 43
5.3.3.2 Proteins of the complement system . . . 45
5.4 Discussion . . . 48

6 Discussion and Conclusions **50**

Bibliography **54**

List of Figures

2.1 Stages in the development of type 1 diabetes. 4

3.1 Logistic function $g(z) = \frac{1}{1+\exp(-z)}$ 7
3.2 SVM learning objective. The goal is to find the hyperplane defined by θ (solid black line) that maximises the margin (dashed lines) to positive (green) and negative (red) training instances. $\xi_i > 0$ (orange arrow) if the distance $\theta^T x^{(i)}$ to the hyperplane is less than one. 9
3.3 A decision tree (left) and the corresponding partitioning of the input space (right). Every node tests an attribute and every test represents the outcome of a test. Every leaf node is associated with a class. Tests on attributes partition the input space by axis-parallel splits. The attributes $x_1 = 4$ and $x_2 = 8$ of a new instance x are tested from the root to a leaf, whose class $h(x) = y_2$ is predicted. 10
3.4 Contingency table for binary classification and common performance measures. . 12
3.5 5-fold stratified cross-validation. Data is randomly partitioned into 5 folds (boxes) of equal size and class distribution. In every iteration, one fold is used for testing (grey shaded), the remaining for training. The process is iterated 5 times, such that every fold is used exactly once for testing. Multiple cross-validation rounds can be performed. 13
3.6 Examples for relevant, irrelevant, and useful features. Relevant features are discriminative alone (projection onto axis). Useful features are predictive in combination, but may be non-predictive alone. (a) x_1 relevant, x_2 irrelevant. (b) Xor problem: x_1 irrelevant, x_2 irrelevant. x_1 and x_2 useful in combination. (c) Correlation does not imply redundancy: x_1 almost irrelevant, x_2 completely irrelevant. x_1 and x_2 useful in combination. 14
3.7 LASSO path for an artificial dataset. Optimal LASSO parameters $\theta(\lambda)$ along the LASSO path are shown from left (high λ) to the right. On the top: number of selected features. In colour: parameters of the ten most useful features for prediction. 18

4.1 Estimated logistic regression coefficients of HLA risk scores 24
4.2 True positive rate versus false positive rate for logistic regression trained on all samples using HLA risk score as feature 25

4.3 True positive rate versus false positive rate for logistic regression trained on all samples using HLA and 40 non-HLA risk scores as features 26
4.4 Estimated logistic regression coefficients of HLA and 40 non-HLA risk scores . . 27
4.5 True positive rate versus false positive rate for logistic regression trained on high-risk samples using 40 non-HLA risk scores as features 28
4.6 Logistic regression coefficients of 40 non-HLA risk scores estimated for high-risk samples 29
4.7 Estimated logistic regression coefficients of 40 HLA/non-HLA interaction terms . 31
4.8 Cross-validation area under the curve (AUC) versus number of top ranked features used to train logistic regression model 33
4.9 Feature rankings clustered by Kendall tau rank correlation coefficient 33
4.10 Kendall tau rank correlation coefficients and complete linkage clustering of feature rankings 34

5.1 True positive rate versus false positive rate estimated by 5-fold stratified cross-validation on the proteomics data without feature selection 40
5.2 Cross-validation area under the curve (AUC) versus number of top ranked features used to train support vector machine 41
5.3 Feature rankings clustered by Kendall tau rank correlation coefficient 42
5.4 Kendall tau rank correlation coefficients and complete linkage clustering of feature rankings 43
5.5 Mean normalised peptide concentrations of slow and rapid progressors 45
5.6 Mean normalised peptide concentrations of slow and rapid progressors 46
5.7 Complement system and the innate immune response. Three different pathways can initiate complement: the classical pathway, the lectin pathway, and the alternative pathway. Initiation pathways merge at the formation of C3 and C5 convertase. Cleavage of C3 and C5 leads to a cascade of further cleavage reactions. Complement activates inflammatory processes, marks pathogens for degradation by lymphocytes (opsonization), and expresses membrane attack complex for cell lysis. Figure taken from [1]. 47

List of Tables

4.1 Statistics about the genomics training set and test set. Training samples were from the Type 1 Diabetes Genetics Consortium (T1DGC), test samples were parents of BABYDIAB children or participated in the DiMelli study. Cases developed type 1 diabetes, controls did not. High-risk samples had at least two risk alleles (HLA DR3-DQA1*0501-DQB1*0201 or HLA DR4-DQA1*0301-DQB1*0302). Data are in absolute numbers (%), or means±SD. 23

4.2 Area under the curve (AUC) for models trained with HLA and 40 non-HLA risk scores. Training (Train), 10-fold stratified cross-validation (CV), and test (Test) AUC is shown. Models were logistic regression (LR), LASSO, SVM with linear kernel (LIN SVM), SVM with RBF kernel (RBF SVM), and random forest (RF). 26

4.3 Area under the curve (AUC) for models trained with 40 non-HLA risk scores of high-risk subjects. Training (Train), 10-fold stratified cross-validation (CV), and test (Test) AUC is shown. Models were logistic regression (LR), LASSO, SVM with linear kernel (LIN SVM), SVM with RBF kernel (RBF SVM), and random forest (RF). 28

4.4 Area under the curve (AUC) for models trained with HLA, 40 non-HLA risk scores, and 40 HLA/non-HLA interaction terms. Training (Train), 10-fold stratified cross-validation (CV), and test (Test) AUC is shown. Models were logistic regression (LR) and LASSO. 31

4.5 Area under the curve (AUC) for models trained with HLA, 40 non-HLA risk scores, and 820 interaction terms. Training (Train), 10-fold stratified cross-validation (CV), and test (Test) AUC is shown. Models were logistic regression (LR), and LASSO. 32

5.1 5-fold stratified cross-validation AUC and standard deviation (SD) on the proteomics dataset without feature selection. Classifiers were LASSO, SVM with linear kernel (LIN SVM), SVM with RBF kernel (RBF SVM), and random forest (RF). 40

5.2 Maximum 5-fold cross-validation AUC of feature ranking methods. Number of features, mean cross-validation AUC, and standard deviation (SD) is shown. . . . 42

5.3 Peptides ranked by SVM RFE and MSVM RFE. Gene symbol and description refers to the protein, to which the peptide was allocated. 44

1 Introduction

Novel high-throughput experimental platforms have reshaped biological research within the last decade: next generation sequencing, DNA microarray, ChIP sequencing, or mass-spectrometry made it not only possible to study the genome, transcriptome, proteome, and metabolome, but also to investigate interactions between different levels of organisation. Due to rapidly dropping costs, these technologies are now practical research tools for analysing a large number of individuals or even single cells, and thus enrich the pool of available biomedical research data. These data help to acquire knowledge about complex human diseases, to develop new therapeutic approaches, and to discover diagnostic and prognostic biomarkers. Being able to sequence genomes and to perform large-scale blood screens within a short time frame allows to predict the risk for diseases and to develop personalised treatments. [2, 3]

Biological research is no longer limited by data, but by statistical and computational models to analyse them. The explosion of data goes along with advances in machine learning and data mining. Machine learning methods automatically learn models from observed data, which can then be used to infer new knowledge. Classification models predict the value of a target variable from a set of observed variables. For instance, logistic regression, support vector machines, or decision trees are employed to predict the function of genes, to discover protein-protein interactions, or to discriminate between different types of cancer. In the biomedical domain, classification models are valuable for prediction and diagnosis. The risk for developing a disease can be predicted from observed causal disease factors, such as genetic mutations or environmental exposures. Diseases can be diagnosed from disease symptoms (biomarkers), such as blood serum proteins or metabolites. [4, 5]

The availability of huge amounts of data not only poses new opportunities, but also challenges. Biomedical data are often high-dimensional, and not all variables are useful for a particular task such as classification. Mass spectrometry can measure the concentration of thousands of proteins, but only few of them may be associated with the disease of interest. Feature selection methods select subsets of variables which optimise a particular performance measure. They improve interpretability of the data by removing noisy as well as redundant variables, and thus facilitate the discovery of the most useful variables. [6, 7, 8]

Type 1 diabetes (T1D) is characterised by an autoimmune destruction of insulin producing pancreatic beta cells with varying progression rate. At the point when disease symptoms be-

come evident and T1D is clinically diagnosed, most beta cells are already destroyed. Insulin replacement therapy is then in most cases the only remaining treatment to counter the lacking insulin production. Early detection of rapid progression to T1D is thus of great medical interest. [9, 10, 11, 12]

The first aim of this thesis was to employ established machine learning methods to improve predicting the risk of T1D with genotyping data, and to quantify the effect of genes on T1D risk. The second aim was to use proteomics data for discriminating between rapid and slow progression to T1D, and to discover potential biomarkers.

In the first half of this thesis, I studied a genomics dataset with 41 genetic risk factors of T1D positive and T1D negative subjects. I fitted classification models, which I evaluated on an independent test set. I tested for gene-gene interactions and employed feature selection to rank genes by their usefulness for classification. In the second half of this thesis, I studied a proteomics dataset with 4384 blood serum peptides of 15 rapid T1D progressors and 15 slow T1D progressors. I employed feature selection to rank peptides by their usefulness for discriminating between rapid and slow progressors, and evaluated classification performance by k-fold stratified cross-validation.

This thesis will first review the disease mechanisms of T1D, and describe classification as well as feature selection methods. Next, it will outline the analysis of the genomics and proteomics dataset in separate chapters. Each chapter describes the data, presents the classification and feature selection results, and closes with a biological discussion of the results. At the end, it discusses the results from a methodological point of view, which applies to both datasets.

2 Type 1 diabetes

Diabetes mellitus is a chronic disease, characterised by high blood glucose levels (hyperglycemia) as a consequence of by either a lack of insulin (type 1 diabetes), or the resistance to insulin (type 2 diabetes). Symptoms are frequent urination, increased thirst, increased hunger, fatigue, and weight loss [13]. Long term effects of high blood glucose levels include kidney failure, cardiovascular diseases, accumulation of keto acids (ketoacidosis), and damages to the retina [11].

Insulin is a key hormone for regulating blood glucose levels. It is released from beta cells of the pancreas when blood glucose levels are elevated, and stimulates cells to absorb glucose by conversion to glycogen, triglycerides, or proteins. As a result, blood glucose levels are lowered to normal levels. Since T1D patients are unable to produce insulin, their blood glucose levels are increased.

The human immune system becomes normally active, when cells are infected by a pathogen. Antigen presenting cells then transport fragments specific to the pathogen (antigens) to their cell surface, where they are presented on receptors, such as the human leukocyte antigen (HLA) complex. Immune cells recognise the presented antigens, trigger an immune response, and destroy the infected cells.

T1D is an autoimmune disease, i.e. the immune system targets cells of the own body. Islet beta cells of the pancreas produce islet autoantigens, namely insulin, glutamate decarboxylase (GAD), zinc transporter 8 (Znt8), or islet antigen 2 (IA-2) [10]. Although being non-pathogenic and essential for glucose homeostasis, immune cells recognise these islet autoantibodies as non-self and initiate an immune response, which destroys beta cells. Different types of immune cells contribute the pathological processes of T1D. B-cells produce islet autoantibodies, which bind to insulin (IAA), GAD (GADA), ZnT8 (ZnT8A), and IA-2 (IA-2A), and thus mark beta cells for destruction by phagocytes [14]. CD4+ T-helper cells amplify the immune response by recruiting and activating further immune cells [15]. CD8+ T-killer cells directly destroy islet autoantigen presenting beta cells [16]. T-regulatory cells normally modulate the immune response by suppressing CD4+ and CD8+ cells. Their role has been shown to be disturbed in patients with T1D, which may account for the uncontrolled attack of beta cells by CD4+ and CD8+ cells [17]. The escape of islet autoantigen attacking T-cells from negative selection in the thymus is another explanation for the autoimmune response in diabetic patients [18].

Although the mechanisms that lead to the destruction of islet beta cells are well studied, the

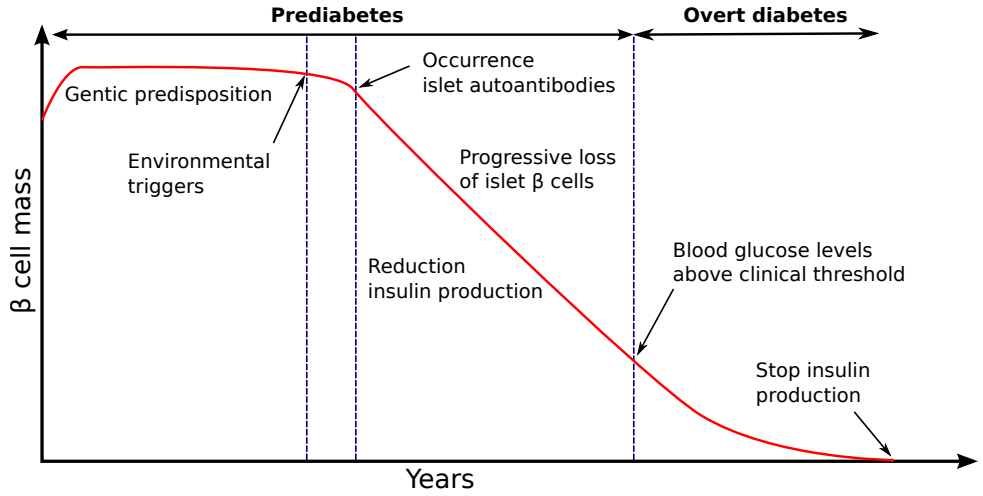


Figure 2.1: Stages in the development of type 1 diabetes.

causal factors for T1D are only poorly understood. Studies suggest a combination of various genetic and environmental factors.

Human leukocyte antigen (HLA) genes are the major genetic risk factors of T1D, which account for 30-50% of the overall genetic risk for T1D [14]. Proinsulin (INS) and protein tyrosine phosphatase N22 (PTPN22) are the two non-HLA genes with the highest effect on T1D [19]. Genome-wide association studies identified about 40 further non-HLA genes, but each gene contributes only little to the overall risk for T1D [20].

The coxsackie virus is one possible environmental trigger for T1D [21]. Since the virus exhibits a high sequence similarity to the GAD islet autoantigen, T-cells against the coxsackie virus may also attack GAD presenting beta cells. Studies further found an association between a lack of beneficial gut bacteria and T1D [22]. Sterile living conditions and the excessive use of antibiotics may therefore account for the high incidence rate of T1D in industrial countries. Further possible environmental triggers are the early consumption of cow milk [23] or certain wheat proteins [24], and a lack of vitamin D [25].

The development of T1D can be split into two stages: prediabetes and overt medical diabetes (see figure 2.1). A period of normal immune function precedes a point in time when environmental factors trigger autoimmunity in genetically predisposed people. The first islet autoantibodies occur (seroconversion), which destroy pancreatic islet beta cells, hence leading to a linear decline of beta cell mass. Simultaneously, the insulin production decreases and the blood glucose levels elevate until they exceed a medical threshold and T1D is clinically diagnosed (overt diabetes). Eventually, all beta cells will be destroyed and insulin production ceases.

At the point of overt diabetes, already a large fraction of beta cells are irreversibly destroyed. The administration of synthetically produced insulin is then in most cases the only treatment to

counter the lacking insulin production. Instead, T1D may be prevented or delayed if detected early in the prediabetic stage. Füchtenbusch et al. [26] showed that an early insulin therapy can delay the onset of T1D. Vaccinating children with islet autoantigens, such as GAD or insulin, is another promising approach to prevent T1D [27]. Methods to early detect children at a high risk for developing T1D are therefore of huge medical interest.

3 Methods

This chapter will outline classification, evaluation, and feature selection methods that were used for data analysis.

3.1 Models for classification

The classification problem is to approximate an unknown function $f : X \to Y$, which assigns a class $y \in Y$ to a sample x from population X. A sample x has n variables (features) x_j. Given is a training set $\mathcal{D} = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$ with m training examples, where $y^{(i)} = f(x^{(i)})$. A classification model defines a hypothesis space $\mathcal{H}$, which contains hypotheses h with parameters $\theta \in \mathbb{R}^d$. Training (learning) a classifier on training set $\mathcal{D}$ is defined as finding the hypothesis $h \in \mathcal{H}$, which approximates the target function f best.

3.1.1 Logistic Regression

In binary logistic regression, the output variable $y \in \{0, 1\}$ is modelled as a Bernoulli random variable with success probability $g(\theta^T x)$:

$$P(y|x, \theta) = g(\theta^T x) = \frac{1}{1 + \exp(-\theta^T x)} \tag{3.1}$$

Here, $P(y|x, \theta) := P(y = 1|x, \theta)$, and g is the logistic function, which maps real numbers to probabilities between zero and one (see figure 3.1).

The hypothesis function h predicts the class with the highest probability:

$$h(x) = \begin{cases} 1 & \text{if } P(y = 1|x) \geq 0.5 \\ 0 & \text{if } P(y = 1|x) < 0.5 \end{cases} \tag{3.2}$$

The logistic regression model is a white-box model, since the parameters θ_j can be interpreted as the effect size of input variable x_j on y. To make this point clear, equation 3.1 can be solved

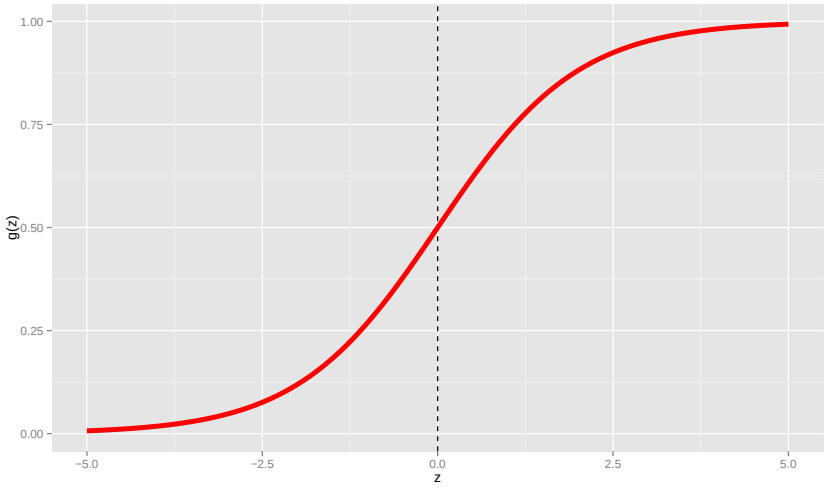


Figure 3.1: Logistic function $g(z) = \frac{1}{1+\exp(-z)}$.

for

$$\text{odds}(x) := \frac{P(y|x)}{1 - P(y|x)} = \exp(\theta^T x), \tag{3.3}$$

which are the odds for $y = 1$. The odds ratio for $x_j = 1$ and $x_j = 0$, holding all remaining variables x_{-j} constant, is

$$\frac{\text{odds}(x_j = 1, x_{-j})}{\text{odds}(x_j = 0, x_{-j})} = \frac{\exp(\theta_0 + \theta_1 x_1 + \cdots + \theta_j + \ldots \theta_n x_n)}{\exp(\theta_0 + \theta_1 x_1 + \cdots + 0 + \ldots \theta_n x_n)} = \exp(\theta_j). \tag{3.4}$$

Hence, $\exp(\theta_j)$ is the odds ratio for $y = 1$, after increasing x_j by one unit, holding all other variables x_{-j} constant. By applying the logarithm to equation 3.4, one can see that θ_j is the log odds per unit increase in x_j. If y indicates a disease, and x_j are potential risk factors, then every unit increase in x_j increases the log odds for that disease by θ_j, controlling for all other variables.

Using equation 3.1, the log conditional likelihood of m independent training examples is

$$\begin{aligned} ll(\theta, \mathcal{D}) &= \log \prod_{i=1}^{m} P(y^{(i)}|x^{(i)}, \theta) = \log \prod_{i=1}^{m} P(y = 1|x^{(i)}, \theta)^{y^{(i)}} P(y = 0|x^{(i)}, \theta)^{1-y^{(i)}} \\ &= \sum_{i=1}^{m} y^{(i)} \log g(\theta^T x^{(i)}) + (1 - y^{(i)}) \log(1 - g(\theta^T x^{(i)})), \end{aligned} \tag{3.5}$$

which is a convex function and can be maximised by gradient descent to learn the maximum likelihood estimator θ^{ML} of the logistic regression model parameters.

3.1.2 LASSO

The least absolute shrinkage selection operator (LASSO) [28], or regularized logistic regression, has the same hypothesis function than logistic regression (see equation 3.2), but the model parameters θ are learnt differently. The maximum likelihood estimator θ^{ML} for logistic regression tends to overfit the data if the number of parameters n is high relative to the number of training examples m. Overfitting can be prevented by making prior assumptions about θ and computing the posterior probability of θ given the training data $\mathcal{D}$:

$$P(\theta|\mathcal{D}) = \frac{P(D|\theta)}{P(D)}P(\theta) \tag{3.6}$$

LASSO assumes a Laplacian distribution with mean zero and precision λ:

$$P(\theta) = \prod_{j=1}^{n} \frac{1}{2}\lambda \exp(-\lambda|\theta_j|) \tag{3.7}$$

The posterior of θ is then proportional to

$$P(\theta|D) \propto \log P(\mathcal{D}|\theta) + \log \prod_{j=1}^{n} \frac{1}{2}\lambda \exp(-\lambda|\theta_j|) \tag{3.8}$$

$$\propto ll(\theta|\mathcal{D}) - \lambda \sum_{j=1}^{n} |\theta_j|. \tag{3.9}$$

Here, $ll(\theta|\mathcal{D})$ is the log conditional likelihood of θ (see equation 3.5), and $-\lambda \sum_{j=1}^{n} |\theta_j|$ is the LASSO regularization term, which penalises parameters that deviate from zero. Maximising equation 3.9 gives the maximum a posteriori estimator θ^{MAP}, which is less prone to overfit the data than θ^{ML}. High λ can prevent overfitting.

3.1.3 Support vector machine

Support vector machines (SVMs) [29] are recent classification models, which can perform efficient non-linear classification via the so called Kernel trick. The hypothesis function of a SVM is:

$$h(x) = \mathrm{sign}\left(\sum_{i=1}^{k} \alpha_i y^{(i)} K(x^{(i)}, x)\right) \tag{3.10}$$

Here, sign is the sign function, and $x^{(i)}$ is a support vector with class $y^{(i)} \in \{+1, -1\}$ and weight α_i. Support vectors are instances of the training set $\mathcal{D}$, and their weights α_i are determined by training the SVM, as described later. K is a kernel function, which quantifies the similarity between $x^{(i)}$ and x. Hence, $h(x)$ is obtained by averaging the classes of support vectors, weighted

by α_i and the similarity $K(x^{(i)}, x)$.

The linear kernel is the scalar product of two vectors u and v:

$$K(u, v) = u^T v \tag{3.11}$$

The Gaussian radial basis function (RBF) kernel is defined as

$$K(u, v) = \exp(-\gamma \|u - v\|_2^2), \tag{3.12}$$

where γ is the precision of the Gaussian distribution. Low γ can prevent overfitting.

If K is a valid Kernel, then Mercer's theorem states:

$$K(u, v) = \phi(u)^T \phi(v) \tag{3.13}$$

ϕ is a kernel dependent function, which maps u onto a p dimensional vector $\phi(u)$. For the linear Kernel, $\phi(u) = u$ is the identify function. For the RBF kernel, ϕ is unknown and p can be infinite, depending on the choice of γ. The Kernel trick is to choose a kernel, such that the transformed training examples $\phi(u)$ become linearly separable. Although $\phi(u)$ and $\phi(v)$ can be vectors of an infinite dimension p, their scalar product $K(u, v)$ can be computed proportionally to d, the dimension of the original space.

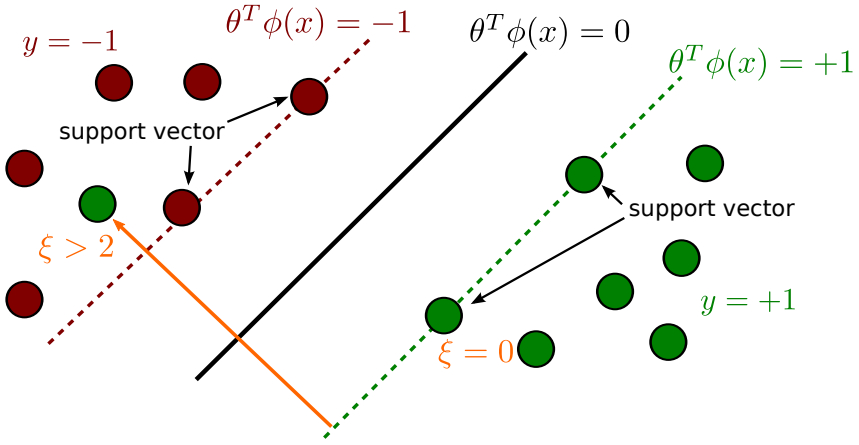


Figure 3.2: SVM learning objective. The goal is to find the hyperplane defined by θ (solid black line) that maximises the margin (dashed lines) to positive (green) and negative (red) training instances. $\xi_i > 0$ (orange arrow) if the distance $\theta^T x^{(i)}$ to the hyperplane is less than one.

By inserting equation 3.13 into equation 3.10, the hypothesis function becomes

$$h(x) = \text{sign}\left(\sum_{i=1}^{k} \alpha_i y^{(i)} \phi(x^{(i)})^T \phi(x)\right) = \text{sign}\left(\left[\sum_{i=1}^{k} \alpha_i y^{(i)} \phi(x^{(i)})\right]^T \phi(x)\right) = \text{sign}\left(\theta^T \phi(x)\right). \tag{3.14}$$

θ are the parameters of the SVM, which are learnt by solving the following constraint optimisation problem:

$$\text{minimise } J(\theta) = \|\theta\|_2 + C\sum_{i=1}^{m} \xi_i \quad \text{, s.t.} \quad \forall i \in \{1, \ldots, m\} : \theta^T \phi(x^{(i)}) y^{(i)} > 1 - \xi_i \tag{3.15}$$

Equation 3.15 aims to find the hyperplane defined by θ, which separates positive and negative examples (see figure 3.2). The slack variables ξ_i allow to violate the large margin conditions $\theta^T \phi(x^{(i)}) y^{(i)} > 1$ if classes are not linearly separable. The cost parameter C penalises unsatisfied constraints. Low C can prevent overfitting.

3.1.4 Random Forest

Decision trees [30, 31] belong to the most common machine learning algorithms. Every internal node of a decision tree tests an attribute, and every edge represents the outcome of a test. Every leaf node is associated with a class. The hypothesis $h(x)$ is the class of the leaf, to which x is assigned by testing the attributes of x starting at the root (see figure 3.3).

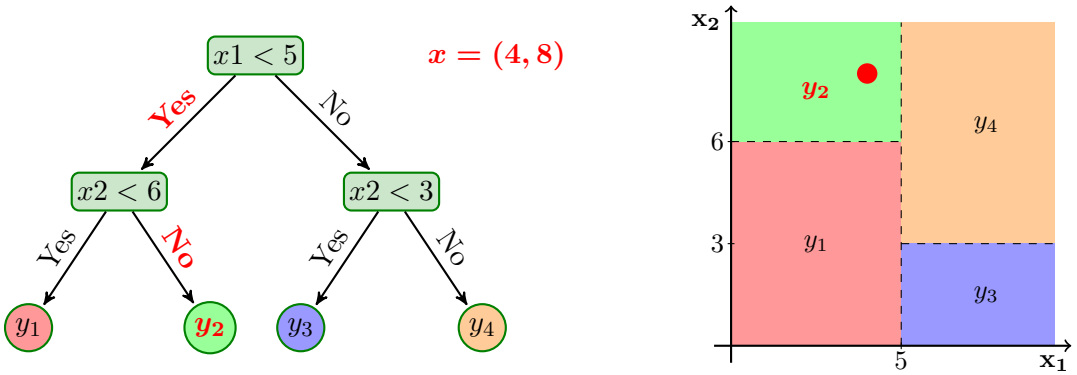


Figure 3.3: A decision tree (left) and the corresponding partitioning of the input space (right). Every node tests an attribute and every test represents the outcome of a test. Every leaf node is associated with a class. Tests on attributes partition the input space by axis-parallel splits. The attributes $x_1 = 4$ and $x_2 = 8$ of a new instance x are tested from the root to a leaf, whose class $h(x) = y_2$ is predicted.

Since learning the optimal decision tree is NP hard [32], heuristic algorithms are used. Divide

and conquer methods start with all training examples in a single root node, and recursively split leaf nodes, which contain samples of different classes. The split attribute optimises some evaluation measure, such as the information gain or the Gini index, which evaluates how well an attribute separates classes. Decision tree pruning is applied to prevent overfitting. Decision trees are easy to interpret, allow to fit complex functions, and scale well with the size of the training set.

A random forest is an ensemble of K randomly generated decision trees $\{h_1, \ldots, h_K\}$. Every tree h_k votes the most popular class $h_k(x)$, and the random forest hypothesis $h(x)$ is the class that receives the most votes:

$$h(x) = \text{argmax}_{y \in Y} |\{k|h_k(x) = y\}| \tag{3.16}$$

Different methods exist for generating K ensemble trees from the training data $\mathcal{D}$ with m examples. Breiman's original random forest algorithm [33] uses bagging and random sampling of split attributes. For every tree k, a bootstrap sample $\mathcal{D}_k$ is generated by sampling m examples with replacement from $\mathcal{D}$. For every split, L out of n possible attributes are sampled, from which the optimal split attribute is selected. The variance of a random forest is proportional to K and L: the higher the number of ensemble trees K and possible split attributes L, the higher the classification accuracy and the risk to overfit the data.

3.2 Evaluation

The hypothesis h of a classifier is evaluated by some performance measure P, which quantifies the agreement between h and the target function f(see section 3.1). The generalisation performance of h is the expected performance $E[P(h(X), f(X))]$ on the population X, which need to be estimated with the training sample $\mathcal{D}$. The following sections will discuss performance measures and estimation methods for binary classification.

3.2.1 Performance measures

A binary classifier predicts the positive class ($h(x) = +1$), or negative class ($h(x) = -1$). The prediction is a true positive (TP), if $h(x) = +1$ and $f(x) = +1$, a false positive (FP), if $h(x) = +1$ and $f(x) = -1$, a true negative (TN), if $h(x) = -1$ and $f(x) = -1$, or a false negative (FN), if $h(x) = -1$ and $f(x) = +1$. Different performance combine TP, FP, TN, and FP (see figure 3.4). The sensitivity, or true positive rate (TPR), is the fraction of positive examples ($TP + FN$) that were predicted as positive. The specificity, or true negative rate (TNR), is the fraction of negative examples ($TN + FP$) that were predicted as negative. A good classifier should have

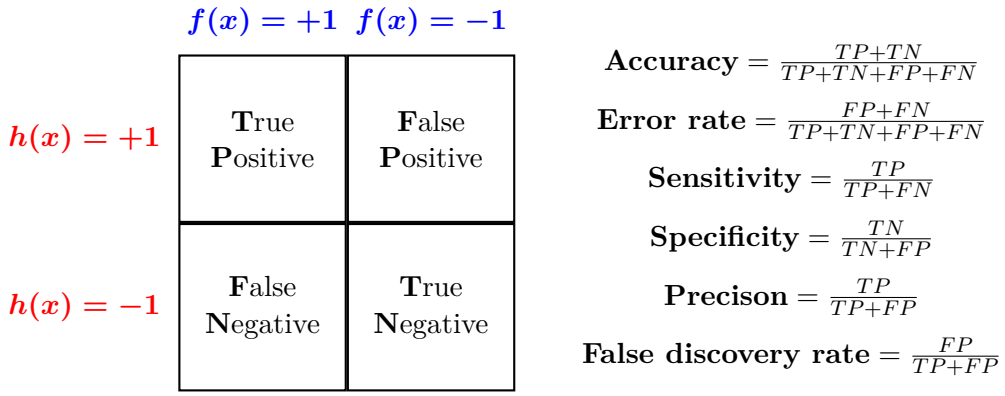


Figure 3.4: Contingency table for binary classification and common performance measures.

both a high sensitivity, and a high specificity. Sensitivity and specificity are independent of the class distributions, since they only depend on the number of positive or negative examples, respectively. Same holds true for the false positive rate (FPR) and false negative rate (FNR). [34]

A scoring classifier predicts a score $g(x)$, which quantifies the degree to which x is member of the positive class. If $g(x)$ is above a threshold τ, the positive class is predicted, formally $h(x) = +1$ if $g(x) > \tau$. For logistic regression, $g(x) = P(y = 1|x)$, i.e. the score is the probability that x is positive (see equation 3.1). A binary SVM can be turned into a scoring classifier by fitting a logistic regression model to the decision values of the training examples (see equation 3.10) [35]. The score of a random forest can be the fraction of ensemble trees that voted for the positive class (see equation 3.16) [33].

Receiver operating characteristic (ROC) curve is commonly used to evaluate scoring classifiers, in particular in the medical domain. For increasing values of τ, it shows the fraction of true positives (TPR) on the y-axis, depending on the fraction of false negatives (FPR) on the x-axis. The area under the ROC curve (AUC) captures the ROC performance in a single value. The AUC corresponds to the probability of ranking a true positive ahead of a false positive when ordering them by their score [34]. A random classifier has an AUC of 0.5, and a perfect classifier an AUC of 1.0. Since TPR and FPR are independent of the class distribution, the ROC curve and the AUC is resistant to changes in the class distribution.

3.2.2 Performance estimation

Different methods exist to estimate the generalisation performance $\epsilon := E[P(h(X), f(X)]$ given a training sample $\mathcal{D}$.

The training set, or resubstitution, estimator of ϵ is computed on the same examples that were

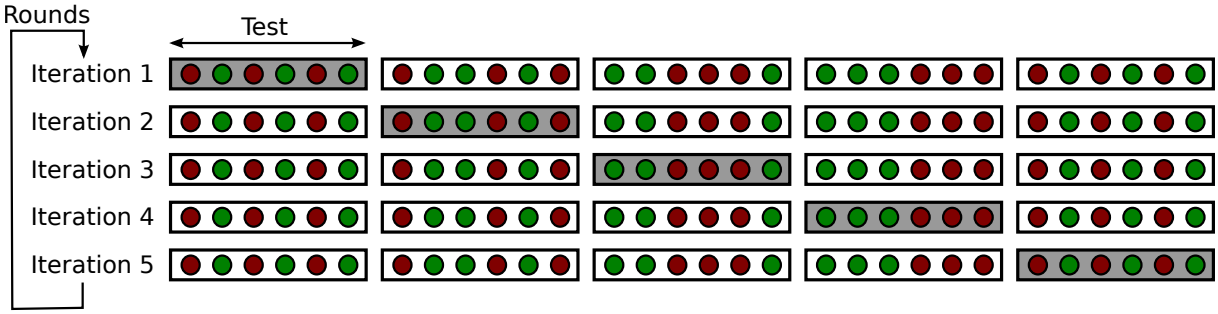


Figure 3.5: 5-fold stratified cross-validation. Data is randomly partitioned into 5 folds (boxes) of equal size and class distribution. In every iteration, one fold is used for testing (grey shaded), the remaining for training. The process is iterated 5 times, such that every fold is used exactly once for testing. Multiple cross-validation rounds can be performed.

used to train h. Since it does not capture a classifier's performance on unseen examples, it will be too optimistic if the training set is small.

In holdout validation, the sample $\mathcal{D}$ is randomly partitioned into a training and test set. h is chosen on the training set and evaluated on the test set. Since the holdout estimator is computed on unseen examples, it is less biased than the resubstitution estimator.

Resampling techniques such as bootstrapping or cross-validation can be applied if $\mathcal{D}$ is small. They use the available data more efficiently than holdout validation, by using examples both for training and testing. In k-fold cross-validation, $\mathcal{D}$ is randomly partitioned into k folds of equal size (see figure 3.5). $k-1$ folds are used to train h, whose performance is evaluated on the remaining test fold. The process is iterated k times, such that every fold is used exactly once as test set. The estimators of all iterations are averaged to obtain a final estimator for ϵ. If k equals the number of training examples $|\mathcal{D}|$, this is called leave-one-out cross-validation. In stratified k-fold cross-validation, $\mathcal{D}$ is partitioned such that the class distribution of every fold is approximately the same as for the complete sample $\mathcal{D}$. Kohavi [36] showed that stratification can reduce the bias and variance of the estimator. Repeating k-fold cross-validation for several rounds is another approach to reduce the variance if $\mathcal{D}$ is small.

3.3 Feature selection

Feature selection aims to reduce the number of features (variables) n in high-dimensional datasets by searching for subsets of features that optimise some evaluation measure. Fewer features can improve the accuracy and runtime of a predictor, make medical tests more cost-effective, and facilitate interpretability as well as knowledge discovery. [6]

Feature selection methods can be classified into filter, wrapper, and embedded methods. Filter, or feature ranking methods, rank features by some evaluation measure that is independent of a specific predictor. The evaluation measure can depend on a single variable (univariate filter), or multiple variables (multivariate filter).

Wrapper methods search for feature subsets that optimises the performance of a specific predictor, which is treated as a black box. Since the feature space contains 2^n possible subsets, exhaustive search is intractable for most practical settings. Forward selection, backward elimination, or floating search are search heuristics, which can explore the feature space more efficiently.

Embedded methods integrate feature selection into the learning process of a specific predictor. Compared with wrapper methods, embedded methods are more efficient, and less prone to overfit the data, but more biased towards the embedded predictor.

This study focuses on filter methods, since feature rankings are attractive for many biological problems: to discover groups of genes belonging to the same biochemical pathway, to select a small set of candidate proteins for follow-up studies, or to develop new diagnostic tests based upon the most predictive and cost-effective combination of biomarkers.

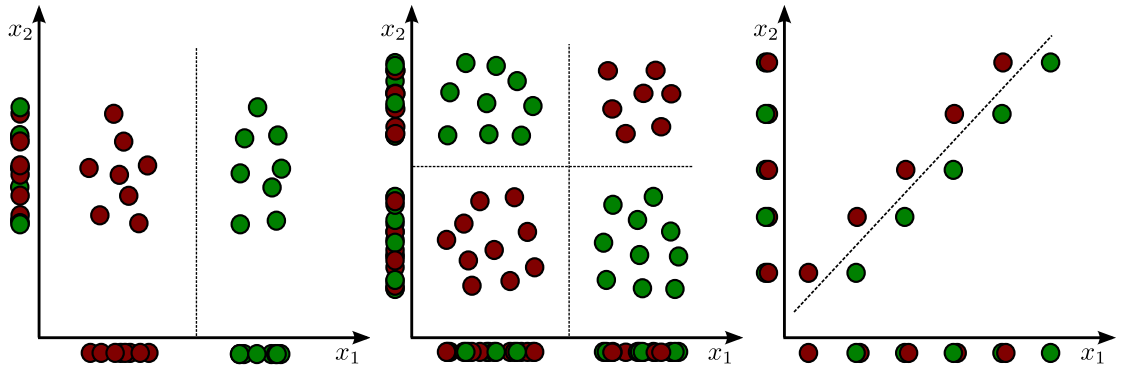


Figure 3.6: Examples for relevant, irrelevant, and useful features. Relevant features are discriminative alone (projection onto axis). Useful features are predictive in combination, but may be non-predictive alone. (a) x_1 relevant, x_2 irrelevant. (b) Xor problem: x_1 irrelevant, x_2 irrelevant. x_1 and x_2 useful in combination. (c) Correlation does not imply redundancy: x_1 almost irrelevant, x_2 completely irrelevant. x_1 and x_2 useful in combination.

Univariate filter can select features x_j that are associated with the output variable y. Kohavi and John [37] defined x_j as relevant, if $P(y|x_j) \neq P(y)$, that means, if x_j contains information about y. In figure 3.6a, x_1 is relevant, since it perfectly discriminates both classes when considered individually (projection onto x-axis). In contrast, x_2 is irrelevant and contains no information about y. Univariate filter are the fastest feature selection methods, but fail to select features

that are irrelevant, but useful for prediction in combination with other features. Figure 3.6b illustrates the xor problem: although x_1 and x_2 perfectly discriminate the classes in combination, they are irrelevant individually and would therefore not be selected by an univariate filter.

Multivariate filter can select relevant, but also and useful features. Additionally, they allow for more compact rankings by removing redundant features. Feature x_2 is redundant given x_1 if $x_2 = f(x_1)$ for some function f. Note that correlated features are not necessarily redundant. In figure 3.6c, x_1 and x_2 are strongly correlated, but not redundant. x_1 is irrelevant, and x_2 carries little information about y. In combination, they are useful for prediction and would be selected by a multivariate filter.

The following sections will outline the feature ranking methods that were used for this study. Every method takes the sample $\mathcal{D}$ with n features (see section 3.1) as input, and outputs a list of ranked features $x_{r_1}, ..., x_{r_n}$, where x_{r_1} is the most useful feature. The best k features $x_{r_1}, ..., x_{r_k}$ can than be used to train any predictor.

3.3.1 FSCORE

The Fisher-score (F-score) [38] measures the dependency between a continuous variable x_j and a categorical variable y. For a binary class variable $y \in \{+1, -1\}$, the F-score of feature x_j is defined as:

$$F(j) = \frac{(\mu(+) - \mu)^2 + (\mu(-) - \mu)^2}{\sigma^2(+) + \sigma^2(-)} \tag{3.17}$$

μ, $\mu(+)$, $\mu(-)$ is the mean of all, positive, and negative samples, respectively. $\sigma^2(+)$ and $\sigma^2(-)$ is the variance of positive and negative samples. The numerator corresponds to the inter-class variance, and the denominator to the sum of within-class variances. If x_j differs between classes, i.e. is dependent on y, the $F(j)$ will be high. FSCORE therefore ranks features in descending order by their F-scores.

The F-score can be computed efficiently and is independent of any classifier. However, it does not take dependencies between features into account, such that irrelevant but useful features will obtain a low rank.

3.3.2 SVM RFE

SVM recursive feature elimination (SVM RFE) [39] is a wrapper for SVM, that performs backward elimination to generate a ranking of feature subsets. It had been applied to identify predictive features in high-dimensional data for gene expression analysis [40], or cancer classification [39, 41].

The cost function $J(\theta)$ (see equation 3.15) of a SVM converges to the expected misclassification rate, if it is evaluated on an infinite number of examples. Features that do not increase $J(\theta)$ when they are removed, can therefore be considered as unimportant for classification. It can be shown that removing feature x_j, i.e. setting θ_j to zero, changes $J(\theta)$ proportionally to $(\theta_j)^2$. Hence, $(\theta_j)^2$ corresponds to the usefulness of feature x_j, and is used as ranking criterion in SVM RFE.

SVM RFE starts with all features $F = \{1, \ldots, n\}$ (see algorithm 1). In every iteration, SVM weights θ are learnt on the training data $\mathcal{D}$ only with the features of the current feature subset F. The ranking criterion $c_j = (\theta_j)^2$ is computed, and the feature with the smallest c_j is eliminated from F. The process is repeated until no features are left. Features which are eliminated at the end have the highest rank.

Algorithm 1: Recursive feature elimination algorithm for support vector machine (SVM RFE).

Input: $\mathcal{D}$ training set with n features
begin
  $F \leftarrow \{1, \ldots, n\}$ is current set of features
  $R \leftarrow []$ is list of ranked features
  while $F \neq \emptyset$ **do**
    Train SVM parameters θ on $\mathcal{D}$ with features F
    $c_j \leftarrow (\theta_j)^2$ is ranking criterion of feature j
    $r \leftarrow \text{argmin}_{j \in F_k} c_j$ is least useful feature, which is eliminated from F
    $F \leftarrow F \setminus \{r\}$ is feature set of next iteration
    $R \leftarrow [r\ R]$ is new list of ranked features
  end
  return R
end

SVM RFE generates a ranking of nested feature subsets $F_1 \subset F_2 \subset \cdots \subset F_n$, which minimise $J(\theta)$. That means, F_k is the subset with k features, which performs best on the training data. However, features in F_k may not be the most relevant features when considered individually.

$(\theta_j)^2$ is proportional to the change in $J(\theta)$, independent of the SVM kernel [39]. SVM with nonlinear kernel can therefore be used to capture nonlinear relationships in the data. In iteration k, SVM RFE algorithm learns a SVM with k features. The runtime is thus highest for the first iterations. Instead of eliminating only one feature per iteration, SVM RFE can be accelerated by eliminating chunks of features as long as the number of current features is above a certain threshold. However, this can result in less accurate rankings.

Following example [39] illustrates the ability of SVM RFE to eliminate redundant features. Assume x_1 and x_2 are equally useful for prediction, and that x_1 and x_2 is copied two and three times, respectively. Since both groups of features are equally useful, SVM RFE learns in every

iteration the same weight for both groups of features:

1. $x_1(\frac{1}{4})$ $x_1(\frac{1}{4})$ $x_2(\frac{1}{6})$ $x_2(\frac{1}{6})$ $x_2(\frac{1}{6})$
2. $x_1(\frac{1}{4})$ $x_1(\frac{1}{4})$ $x_2(\frac{1}{4})$ $x_2(\frac{1}{4})$
3. $x_1(\frac{1}{4})$ $x_1(\frac{1}{4})$ $x_2(\frac{1}{2})$
4. $x_1(\frac{1}{2})$ $x_2(\frac{1}{2})$

In the first iteration, SVM RFE eliminates one copy of x_2, since it has the lowest weight. In the second iteration, it eliminates either x_1 or x_2. In the latter case, x_1 has a lower weight in the third iteration and one copy is thus eliminated. Hence, SVM RFE correctly selects one copy of x_1 and one copy of x_2, whereas an univariate filter such as FSCORE may have select x_1 or x_2 twice.

3.3.3 MSVM RFE

Multiple SVM RFE (MSVM RFE) [41] is similar to SVM RFE, but trains multiple SVMs on different samples from the training data, and combines their weight vectors to a feature ranking score.

In every iteration, t subsamples $\mathcal{D}^{(l)}$ are resampled from $\mathcal{D}$, for instance by bootstrapping of k-fold cross-validation. The SVM weights $\theta^{(l)}$ are learnt on $\mathcal{D}^{(l)}$ and normalised:

$$\theta^{(l)} = \frac{\theta^{(l)}}{\|\theta^{(l)}\|} \tag{3.18}$$

$v_j^{(l)} = (\theta_j^{(l)})^2$ is the sample ranking score of feature x_j, and μ_j and σ_j^2 is the sample mean and sample variance:

$$\mu_j = \frac{1}{t}\sum_{l=1}^{t} v_j^{(l)} \qquad \sigma_j^2 = \frac{1}{t-1}\sum_{l=1}^{t}\left(v_j^{(l)} - \mu_j\right)^2 \tag{3.19}$$

The final ranking criterion is

$$c_j = \frac{\mu_j}{\sigma_j} \tag{3.20}$$

and the feature with the smallest c_j is eliminated.

The authors compared MSVM RFE with the original SVM RFE on different datasets, and reported a lower misclassification rate, a higher sensitivity, and a higher specificity. MSVM RFE also generated more diverse feature sets F_k, which contained functionally unrelated features.

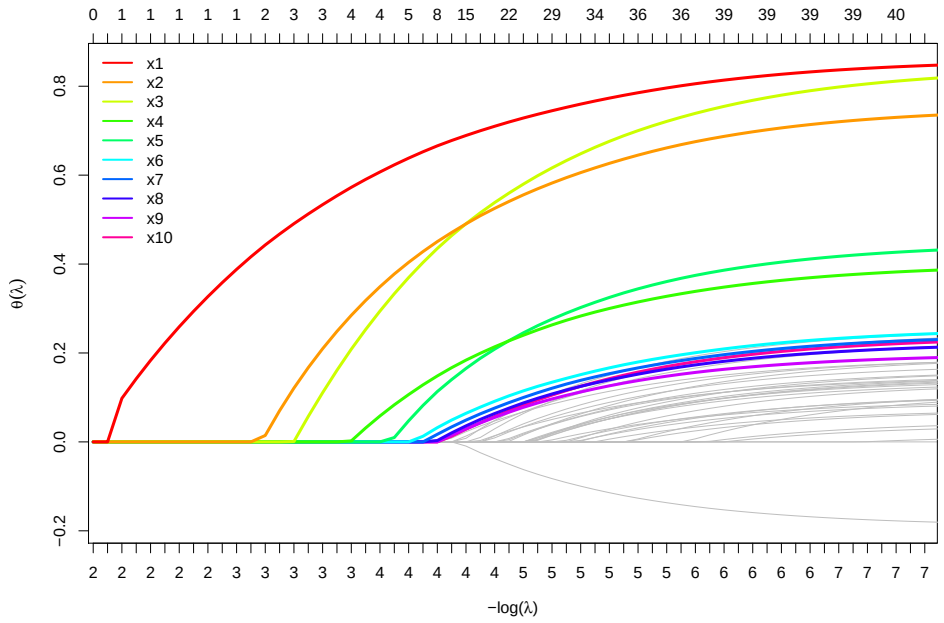


Figure 3.7: LASSO path for an artificial dataset. Optimal LASSO parameters $\theta(\lambda)$ along the LASSO path are shown from left (high λ) to the right. On the top: number of selected features. In colour: parameters of the ten most useful features for prediction.

However, learning t SVMs per iteration, increases the runtime by a factor of t compared with SVM RFE.

For this study, r rounds of k-fold cross-validation were used to generate subsamples $\mathcal{D}^{(l)}$.

3.3.4 LASSO-R

The L1 penalty (see equation 3.9) will shrink parameters θ to zero, if the regularization parameter λ is sufficiently high [28]. This leads to sparse models with few nonzero parameters, where λ controls the model sparsity. Features x_j whose corresponding parameter θ_j is nonzero, are selected. LASSO is thus an embedded method, which selects features while learning the model parameters.

The LASSO path represents the optimal LASSO parameters $\theta(\lambda)$ as a function of λ (see figure 3.7). For high λ, i.e. at the beginning of the LASSO path, LASSO selects only the most useful features, whereas it allows for more features, if λ is small. The earlier θ_j of a feature x_j becomes nonzero along the LASSO path, the more useful is x_j for prediction. Based on this observation, I devised the LASSO-R feature ranking heuristic. For every feature x_j, LASSO-R computes the highest value λ_j, such that $|\theta(\lambda_j)_j| > 0$, and ranks features by λ_j in descending

order.

The LASSO path can be computed efficiently, even for high-dimensional datasets with many samples [42]. LASSO-R is therefore an efficient feature ranking algorithm, applicable to large datasets. Since LASSO is a multivariate model, LASSO-R can captures dependencies between features. If the data contains groups of correlated features, LASSO will select the group member that correlates most with the class variable. Features which carry non-orthogonal information about the class variable, are thus removed. However, all group members may be of interest in some setting, e.g. for studying all proteins that belong to the same biochemical pathway. Here, ELASTIC-NET [43] or RANDOM-LASSO [44] may be more appropriate than LASSO for computing the LASSO path since they can select groups of correlated features.

3.3.5 RF-R

Breiman [33] took following approach to approximate the importance of feature x_j. Let $\mathcal{D}_k$ be the bootstrap sample of tree k (see section 3.1.4) and $O_k = \mathcal{D} \setminus \mathcal{D}_k$ be the out-of-bag sample of tree k, i.e. all training examples that were not used to grow tree k. Let further e_k be the misclassification rate on the out-of-bag sample O_k. Generate O'_k by permuting the values of x_j in O_k, such that x_j becomes independent of the class variable. Let e'_k be the misclassification rate on O'_k. The importance of feature x_j is then defined as:

$$\text{Importance(j)} := \frac{1}{K} \sum_{k=1}^{K} e'_k - e_k \tag{3.21}$$

Importance(j) is the expected increase in the misclassification rate, after resolving the dependency between x_j and the class variable y. Features do no increase the misclassification rate are unimportant. RF-R therefore ranks features by Importance(j) in descending order.

The costs for computing the importance of all variables depends on the number of ensemble trees and the size of the training data. Importance(j) captures dependencies between features and is low for redundant features: Assume that x_j and x_k contain the same information about y. Then each ensemble tree will either choose x_j or x_k, and x_j will be a split attribute in at most half to the trees. Permuting x_j will increase e'_k of at most half of the trees (see equation 3.21), and x_j becomes half as important as it would be without x_k.

3.3.6 VARBVS

Variational Bayesian variable selection (VARBVS) [45] takes a Bayesian approach to select the most predictive features. In *Bayesian regression*, the model parameters θ are randomly

distributed with some prior density $P(\theta)$. Given a data matrix $\boldsymbol{X}$, the joint distribution of class variable y and model parameters θ factorises as follows:

$$P(y, \theta|\boldsymbol{X}) = P(y|\theta, \boldsymbol{X})P(\theta) \tag{3.22}$$

Bayesian variable selection allows for different models γ, which include different variables. $\gamma_j \in \{0, 1\}$ indicates, if the model includes variable x_j. The joint distribution for Bayesian variable selection factorises as follows:

$$P(y, \theta, \gamma|\boldsymbol{X}) = P(y|\theta, \boldsymbol{X})P(\theta|\gamma)P(\gamma) \tag{3.23}$$

The posterior inclusion probability (PIP) of variable x_j is inferred with Bayes'Theorem:

$$\text{PIP}(j) := P(\gamma_j = 1|y, \boldsymbol{X}) = \frac{\sum_{\gamma_{-j}} \int P(y, \theta, \gamma_j = 1, \gamma_{-j}|\boldsymbol{X})d\theta}{\sum_{\gamma} \int P(y, \theta, \gamma|\boldsymbol{X})d\theta} \tag{3.24}$$

Here, γ_{-j} denotes all variables but j. PIP(j) is the fraction of models that selected x_j as predictor variable, and can be used as ranking criterion: the higher PIP(j), the more useful is x_j for predicting the class y.

Computing PIP(j) requires marginalising out the parameters θ of all models γ. For n variables, there are 2^n possible assignments to vector γ. Thus, summing over all models is intractable in practice, even for datasets with few variables. Different methods have been proposed to address this problem.

One way is to sample models from the joint distribution (see equation 3.23), and to approximate PIP(j) by the frequency of models that include x_j. Green [46] proposed reversible jump MCMC (RJ-MCMC), which extends the Metropolis Hastings criterion [47] for sampling from a transdimensional model space. Lunn et al. [48] applied RJ-MCMC to investigate genome-wide association studies. The Swendsen-Wang algorithm [49] and Bayesian adaptive sampling [50] are further sampling methods for Bayesian variable selection. However, sampling remains computationally costly and is not applicable to datasets with thousands of variables.

Hoeting et al. [51] suggested two approximations to make Bayesian variable selection more tractable. First, the Bayesian information criterion (BIC) [52], to approximate the integral over the model parameter θ in equation 3.24. Second, the *Occam's window* method [53], to reduce the number of models which need to be marginalised out.

VARVBS employs variational inference [54] to approximate the joint distribution (see equa-

tion 3.23) by a simpler distribution, which assumes independence of the model parameters θ:

$$P(y, \theta, \gamma | \boldsymbol{X}, \sigma_\theta^2, \pi) = P(y|\theta, \boldsymbol{X}) \prod_{j=1}^{n} P(\theta_j|\gamma_j, \sigma_\theta^2) \prod_{j=1}^{n} P(\gamma_j|\pi) \tag{3.25}$$

Here, $P(\theta_j|\gamma_j, \sigma_\theta^2) = N(\theta_j|0, \sigma_\theta^2)$ and $P(\gamma_j|\pi) = \text{Bernoulli}(\pi)$. The hyperparameters σ_θ^2 and π need to be specified by the user. σ_θ^2 controls the spread of model parameters and can prevent overfitting. π is the prior inclusion probability and controls the sparsity of the model.

VARBVS scales linearly with the number of variables n, such that variables can be ranked efficiently by PIP(j) in descending order. Since it is multivariate, dependencies between variables are taken into account. The approximated posterior inclusion probabilities may be inaccurate, if model parameters are dependent, for instance if two genetic loci are linked, or if peptides belong to the same biochemical pathway. Yet, the authors reported useful inference even for datasets that violated the independence assumption [45].

4 Genomics dataset

This chapter will outline the data analysis of the genomics dataset.

4.1 Introduction

Autoantibodies against pancreatic islet beta cells are currently used to predict T1D [55, 56], but they are expensive to measure and sample dependent. Recent studies therefore focus on cost-effective genetic markers for predicting T1D.

Existing prediction methods based on genetic risk factors [57, 58, 59] have following limitations. First, they only use HLA genes or few (< 10) non-HLA genes and thus ignore the cumulative effect of many non-HLA genes. Second, they ignore possible gene-gene interactions, which can moderate the effect of individual genes. Third, they do not rank genes by their usefulness for predicting T1D, which can provide insights into the disease mechanisms of T1D and help to define a cost-effective set of disease markers.

In the first half of my thesis, I contributed to a research project of the Institute of Diabetes Research at the Helmholtz-Zentrum Munich, which aimed to a) improve predicting the risk for T1D based on HLA and 40 non-HLA genes, and b) to estimate their effect on the development of T1D. I trained classification models of different complexity for discriminating between healthy and diabetic subjects, and I evaluated them on an independent test set. I tested for possible gene-gene interactions, and used feature ranking to quantify genes by their usefulness for predicting T1D. The dataset on which I carried out the analysis will henceforth be referred to as *genomics dataset*.

4.2 Dataset

The following sections will outline the genomics training set for fitting model parameters, the independent test for evaluating classification performance, and the features of the data.

	Training set			Test set		
	Total	Cases	Controls	Total	Cases	Controls
Samples	5795	4587	1208	1188	765	423
High-risk	2387 (41.2)	2295 (50.0)	92 (7.6)	295 (24.8)	278 (36.3)	17 (4.0)
Females	2825 (48.7)	2167 (47.2)	658 (54.5)	601 (50.6)	443 (57.9)	158 (37.4)
Age of onset	-	7.9 ± 4.0	-	-	14.2 ± 7.6	-

Table 4.1: Statistics about the genomics training set and test set. Training samples were from the Type 1 Diabetes Genetics Consortium (T1DGC), test samples were parents of BABYDIAB children or participated in the DiMelli study. Cases developed type 1 diabetes, controls did not. High-risk samples had at least two risk alleles (HLA DR3-DQA1*0501-DQB1*0201 or HLA DR4-DQA1*0301-DQB1*0302). Data are in absolute numbers (%), or means±SD.

4.2.1 Training set

The Type 1 Diabetes Genetics Consortium (T1DGC) aims to identify genes that are associated with T1D [60]. It recruits families from four regional networks: North America, United Kingdom, Europe, and Asia. Three types of families exist: 1) affected sib-pairs with at least two siblings (brothers and/or sisters), 2) trios of a mother, father, and a single child, 3) controls without a family history of T1D. Family members are genotyped and classified as T1D positive or T1D negative. Data are provided to the T1D research community.

The genomics dataset for model training was derived from the T1DGC.2011.03 Taqman dataset. Only subjects from the network Europe and United Kingdom were included. Cases had no history of T1D. The resulting training set contained 4587 cases (average age of onset of 7.9 years), and 1208 controls (see table 4.1).

4.2.2 Test set

Test samples either participated in the BABYDIAB [61] or DiMelli study [62]. 423 unrelated cases and 437 unrelated controls were parents of BABYDIAB children from Germany. 328 further cases were taken from the DiMelli register, which contains T1D patients with age below 20 years, diagnosed in Bavaria, Germany. The resulting test set contained 751 cases (average age of onset of 14.2 years), and 437 controls.

4.2.3 Features

Features of training and test samples were 40 non-HLA risk scores and a single HLA risk score.

A non-HLA risk score was derived from the genotype of a single nucleotide polymorphism (SNP)

in a non-HLA gene. It was 0, 1, or 2, if there were no (aa), one (Aa), or two (AA) copies of the risk allele A.

The HLA risk score was derived from the presence of the risk alleles HLA DR3-DQA1*0501-DQB1*0201 (A) and HLA DR4-DQA1*0301-DQB1*0302 (B) at position rs2187668 and rs2187668, which are the major genetic determinants of T1D [63, 57]. The HLA risk score was 0 for aabb, 1 for Aabb, 2 for aaBb, 3 for AAbb, 4 for aaBB, and 5 for AaBb, AABb, AaBB, or AABB. High-risk subjects were defined as those with a HLA risk score > 2.

4.3 Results

The following sections present the classification performance of models that were trained with HLA and non-HLA risk scores, as well the performance of feature selection methods on the genomics dataset.

4.3.1 HLA model

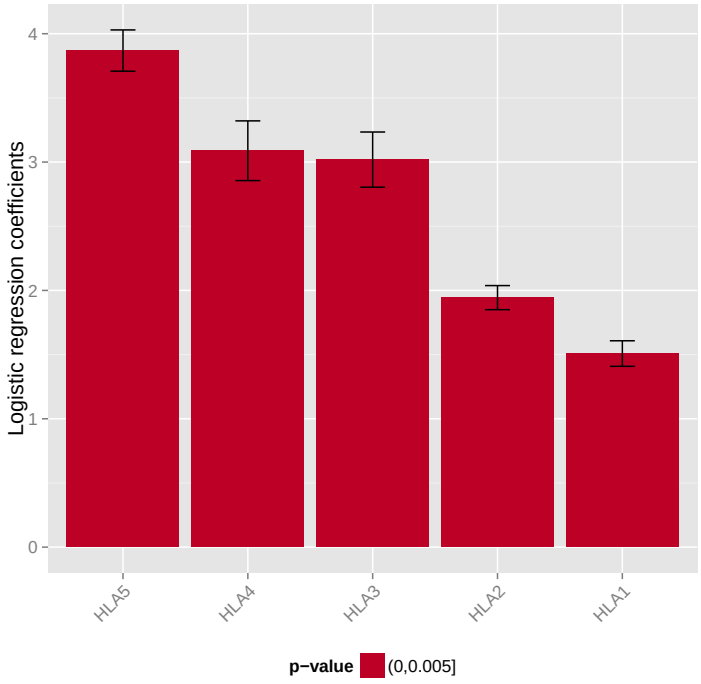


Figure 4.1: Estimated logistic regression coefficients of HLA risk scores. Error bars show 95% confidence interval. Colours refer to p-value of t-test.

First, I trained a logistic regression model using the HLA risk score as the only predictor vari-

able. This model served as baseline, whose prediction performance I later aimed to improve by including non-HLA risk score.

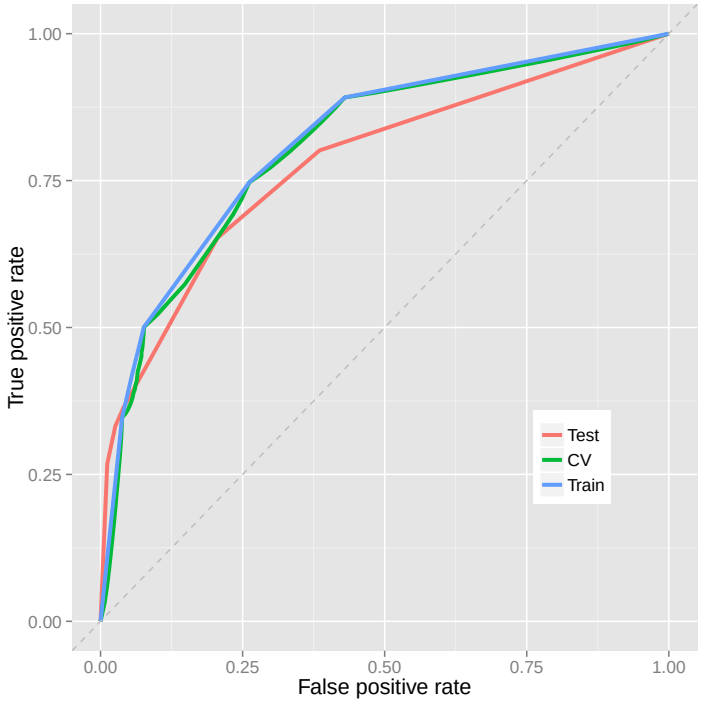


Figure 4.2: True positive rate versus false positive rate for logistic regression trained on all samples using HLA risk score as feature. Training set (Train), 10-fold stratified cross-validation (CV), and test set (Test) performance is shown.

HLA was treated as a categorical variable and dummy coded. Figure 4.1 shows the estimated regression coefficients. HLA 1, 2, 3, 4, and 5 significantly (p-value < 0.005) increased the log odds of T1D by 1.5, 1.9, 3.0, 3.1, and 3.7, respectively. The mean increase in log odds was 1.04 for low risk (HLA$\leq$ 2), and 3.32 for high-risk (HLA$>$ 2) subjects.

The training, cross-validation, and test AUC was 0.82, 0.81, and 0.78, respectively (see figure 4.2).

4.3.2 HLA/non-HLA model

I included non-HLA risk scores for learning models to a) improve the prediction performance of the baseline model that contained only HLA as predictor variable, and b) to estimate the effect on non-HLA genes on T1D risk.

Logistic regression (LR), LASSO, and SVM with linear kernel (LIN SVM), were used as linear models, SVM with RBF kernel (RBF SVM), and random forest (RF) as nonlinear models.

	LR	LASSO	LIN SVM	RBF SVM	RF
Train	0.87	0.87	0.87	1.00	1.00
CV	0.86	0.86	0.86	0.79	0.85
Test	0.84	0.84	0.84	0.75	0.82

Table 4.2: Area under the curve (AUC) for models trained with HLA and 40 non-HLA risk scores. Training (Train), 10-fold stratified cross-validation (CV), and test (Test) AUC is shown. Models were logistic regression (LR), LASSO, SVM with linear kernel (LIN SVM), SVM with RBF kernel (RBF SVM), and random forest (RF).

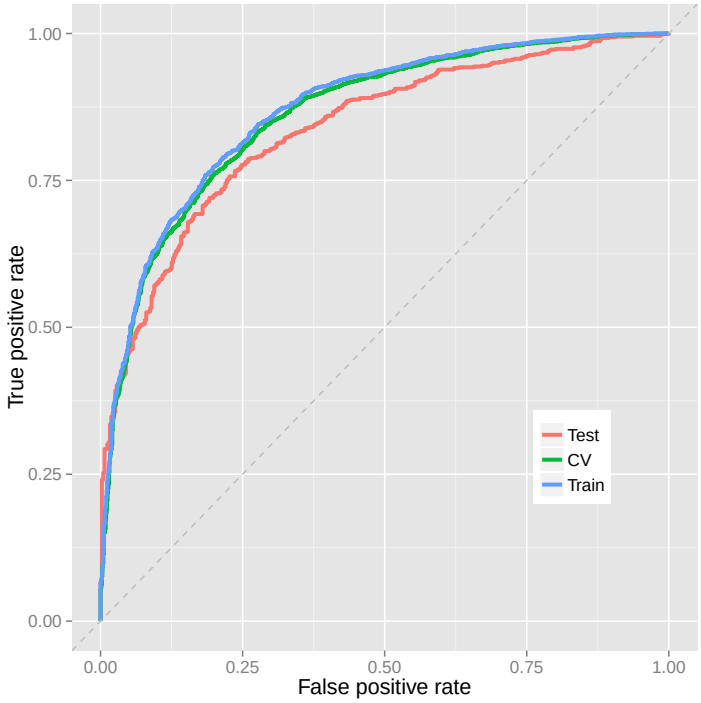


Figure 4.3: True positive rate versus false positive rate for logistic regression trained on all samples using HLA and 40 non-HLA risk scores as features. Training set (Train), 10-fold stratified cross-validation (CV), and test set (Test) performance is shown.

Models were trained with all 5795 samples and 41 features, which were HLA and 40 non-HLA risk scores. Since it was assumed that risk scores increase the risk for T1D linearly, risk scores were modelled as discrete variables. Using dummy coding to model risk scores as categorical allowed for a better fit of the training data due to the higher number features (85 instead of 41), but the cross-validation and test AUC decreased. Additionally, the standard error of the regression coefficients increased.

Model parameters were optimised by 10-fold stratified cross-validation on the training set. $\lambda = 0.0012$ was optimal for LASSO, $C = 50$ for LIN SVM, $C = 50$ and $\gamma = 0.025$ for RBF SVM, $K = 50$ and $L = 10$ for RF (see section 3.1 for model details).

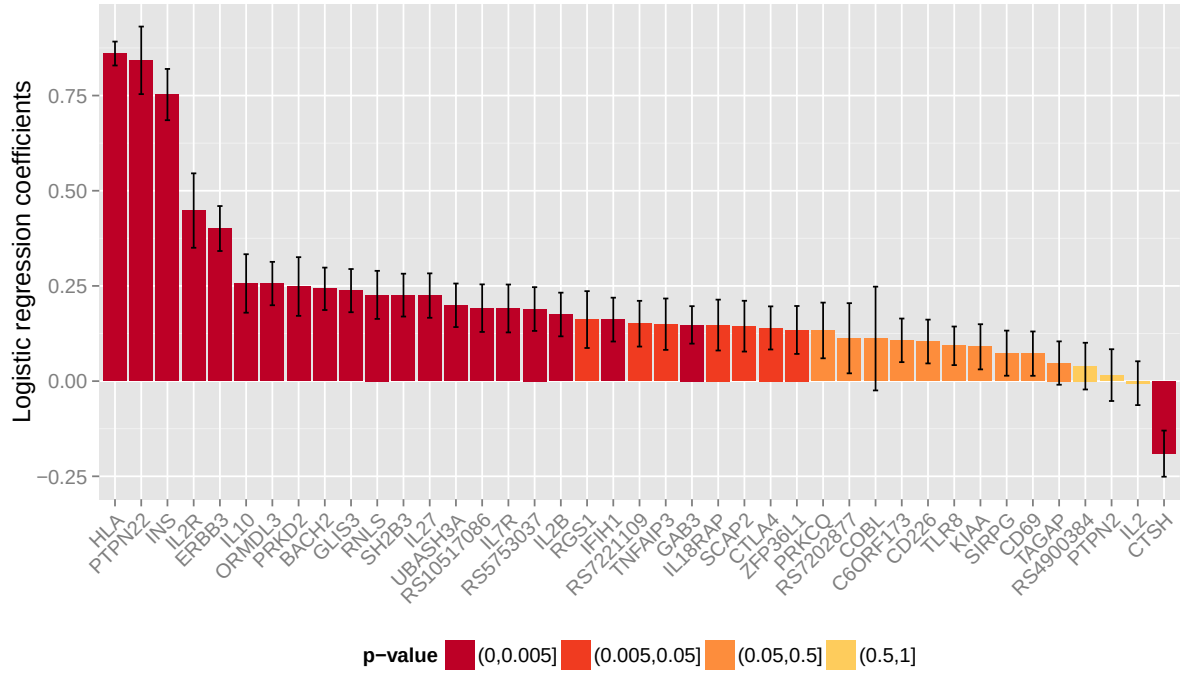


Figure 4.4: Estimated logistic regression coefficients of HLA and 40 non-HLA risk scores. Error bars show 95% confidence interval. Colours refer to p-value of t-test.

To account for the higher fraction of cases than controls in the training set (see table 4.1), LIN SVM and RBF SVM were trained with different class weights: 0.8 for cases and 0.2 for controls (one minus the relative class frequency).

Table 4.2 shows the prediction performance of all models. The training, cross-validation and test AUC was 0.87, 0.86, and 0.84, for linear models. Using 40 non-HLA risk scores in addition to the HLA risk code thus increased the AUC by 5%, 5%, and 6% (absolute difference). Nonlinear models perfectly fitted the training data, but had a lower cross-validation and test AUC.

Figure 4.3 shows the ROC curves for logistic regression. 40 non-HLA risk scores in addition to the HLA risk score decreased the model deviance from 4538 to 4029. The model thus fitted the training data significantly (likelihood ratio test p-value < 0.005) better than the baseline model.

Figure 4.4 shows the estimated regression coefficients. HLA had the highest effect size ($\beta = 0.86$), followed by PTPN22 ($\beta = 0.84$) and INS ($\beta = 0.75$). Note that the maximum level of HLA was 5, and only 2 for non-HLA risk scores. HLA=5 thus increased the log odds of T1D by 4.3, which is 2.6 times higher than 1.7 for PTPN22=2. The negative coefficient of CTSH indicated a protective effect of the allele that was used to define its risk score.

4.3.3 non-HLA model for high-risk subjects

The Institute of Diabetes Research aimed to develop a genetic test specifically for high-risk subjects (HLA$>$ 3). I therefore trained and evaluated models only on high-risk subjects using 40 non-HLA risk scores as features, without the HLA risk score.

	LR	LASSO	LIN SVM	RBF SVM	RF
Train	0.79	0.79	0.80	1.00	1.00
CV	0.69	0.69	0.68	0.56	0.63
Test	0.71	0.71	0.72	0.61	0.73

Table 4.3: Area under the curve (AUC) for models trained with 40 non-HLA risk scores of high-risk subjects. Training (Train), 10-fold stratified cross-validation (CV), and test (Test) AUC is shown. Models were logistic regression (LR), LASSO, SVM with linear kernel (LIN SVM), SVM with RBF kernel (RBF SVM), and random forest (RF).

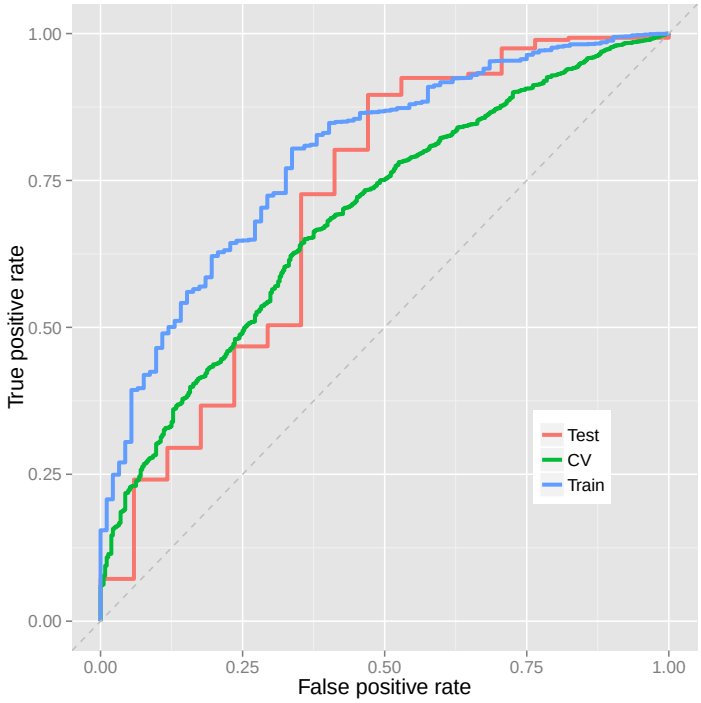


Figure 4.5: True positive rate versus false positive rate for logistic regression trained on high-risk samples using 40 non-HLA risk scores as features. Data are for training set (Train), 10-fold stratified cross-validation (CV), and test set (Test).

After removing all low risk subjects (HLA$\leq$ 2), the training set contained 2295 cases and 95 controls, the test set 278 cases and only 17 controls (see table 4.1). The same models as described in section 4.3.2 were trained. Model parameters were optimised by four repetitions of 5-fold, instead of 10-fold, stratified cross-validation. Each fold thus contained 19 instead of only 9

controls, which decreased the standard error of the AUC estimator. $\lambda = 0.0002$ was optimal for LASSO, $C = 100$ for LIN SVM, $C = 50$ and $\gamma = 0.1$ for RBF SVM, $K = 50$ and $L = 2$ for RF. Optimal parameters for LIN SVM and RBF SVM were sensitive to class weights, which were 0.04 for cases, and 0.96 for controls.

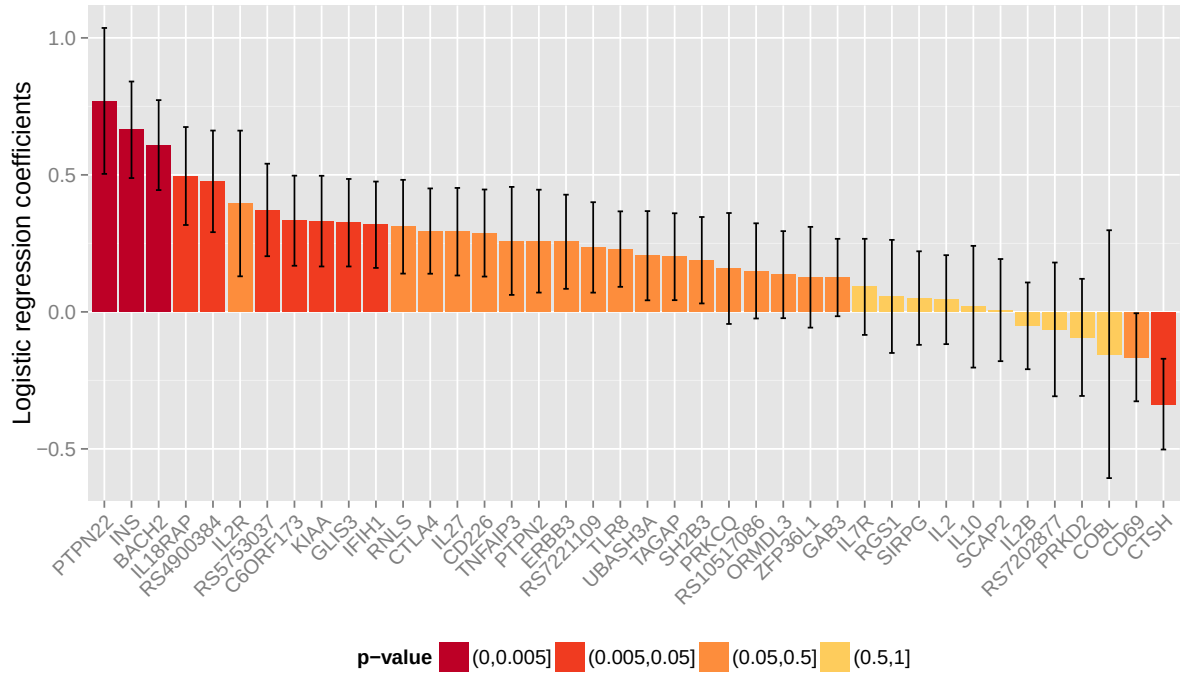


Figure 4.6: Logistic regression coefficients of 40 non-HLA risk scores estimated for high-risk samples. Error bars show 95% confidence interval. Colours refer to p-value of t-test.

All linear models achieved a similar prediction performance (see table 4.3). The test, cross-validation, and training AUC for logistic regression was 0.79, 0.69, and 0.71, respectively. This was 8%, 17%, and 13% lower compared with using all samples and the HLA risk score (see table 4.2). Figure 4.5 shows the corresponding ROC curves, which are jagged due to the small number of controls in the training and test set. Nonlinear models performed perfectly on the training set, but had a lower cross-validation AUC.

The standard error of the estimated logistic regression coefficients (see figure 4.6) increased due to the smaller training set. As for the complete training set (see figure 4.4), the non-HLA genes PTPN22 ($\beta = 0.77$) and INS ($\beta = 0.66$) had the highest effect size. However, the difference to the remaining genes was less pronounced.

4.3.4 Interactions

Interactions between multiple input variables occur if their effect on the output variable is not additive [64, 65]. Interactions arise in many biological settings [66, 67]. For instance, gene X_1 may only increase the risk for T1D in the presence of transcript X_2 if X_2 is an activator for X_1.

Linear models do not capture interactions. Let X_1 and X_2 be two input variables, and Y the output variable. A linear model can then be formulated as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{4.1}$$

where β_i are the regression coefficients. Differentiation Y with respect to X_1

$$\frac{\partial Y}{\partial X_1} = \beta_1 \tag{4.2}$$

yields the main effect size β_1 of X_1, which is constant and does not depend on X_2. Pairwise, or second-order interactions, can be modelled explicitly by adding product terms $X_1 X_2$ to the linear model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 \tag{4.3}$$

$$\frac{\partial Y}{\partial X_1} = \beta_1 + \beta_{12} X_2 \tag{4.4}$$

Now, X_1 has main effect size β_1 and interaction effect size β_{12}, which depends on X_2.

Nonlinear models can capture interactions without explicitly adding product terms. A decision tree can model an interaction between X_1 and X_2, by first splitting on X_1 and then on X_2. A SVM can model interactions using a nonlinear kernel K, such that $\phi(x)$ contains product terms of one or multiple variables (see section 3.1.3).

4.3.4.1 HLA interactions

I tested for pairwise interactions between HLA and all non-HLA risk scores X_i by training linear models with additional product terms HLA $* X_i$.

Logistic regression and LASSO were trained with all 5795 samples using 41 risk scores and 40 product terms as features. SVM with linear kernel was excluded from this analysis, since it scaled worse on the large training set and did not achieve a better performance in previous experiments (see table 4.2 and 4.3). 10-fold stratified cross-validation was used to optimise LASSO ($\lambda = 0.0022$) and to estimate the AUC.

	LR	LASSO
Train	0.87	0.87
CV	0.86	0.86
Test	0.83	0.84

Table 4.4: Area under the curve (AUC) for models trained with HLA, 40 non-HLA risk scores, and 40 HLA/non-HLA interaction terms. Training (Train), 10-fold stratified cross-validation (CV), and test (Test) AUC is shown. Models were logistic regression (LR) and LASSO.

The training, cross-validation, and test AUC was similar for logistic regression and LASSO (see table 4.4), and as high as is was for the model without product terms (see table 4.2).

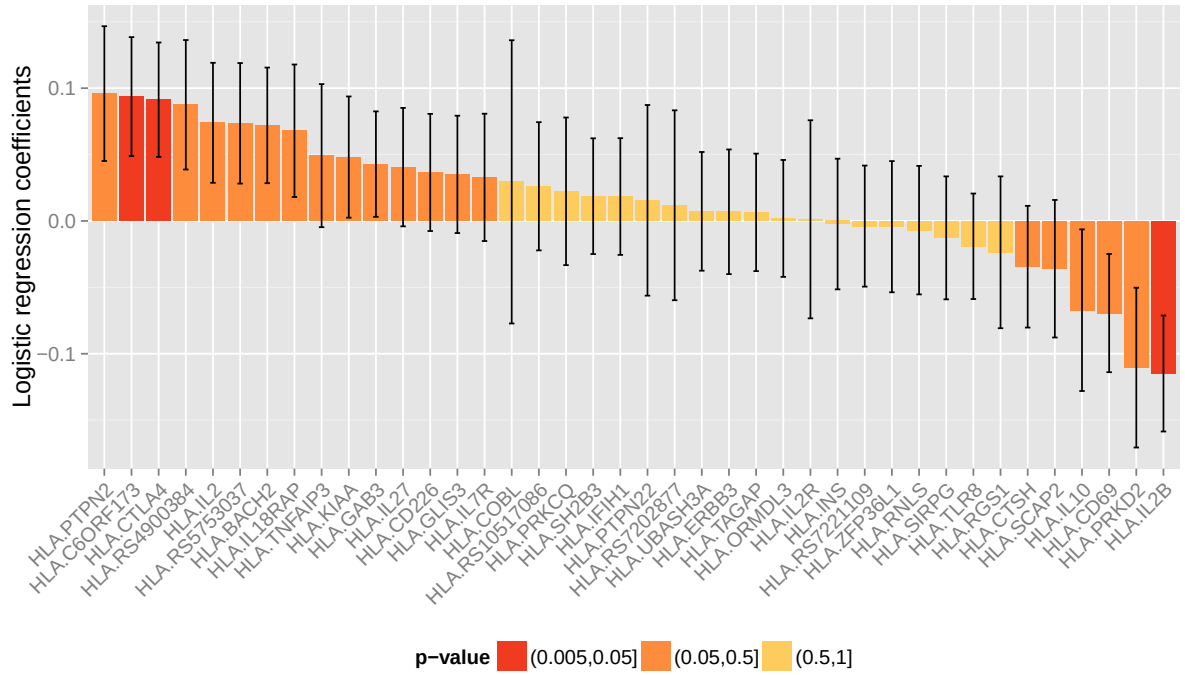


Figure 4.7: Estimated logistic regression coefficients of 40 HLA/non-HLA interaction terms. Error bars show 95% confidence interval. Colours refer to p-value of t-test.

Figure 4.7 shows the estimated effect size of all 40 product terms. Using 81 instead of 41 features increased the standard error. Weakly significant (p-value $\in (0.005, 0.05]$) interactions were found between HLA and IL2B ($\beta = -0.5$), HLA and C6ORF ($\beta = 0.094$), as well as HLA and CTLA4 ($\beta = 0.091$).

4.3.4.2 HLA/non-HLA interactions

I tested for pairwise interactions between all risk scores X_i and X_j, by training linear models with additional product terms $X_i X_j$.

Features were 41 risk scores and 860 product terms. Logistic regression and LASSO ($\lambda = 0.0052$) were used as linear models.

	LR	LASSO
Train	0.93	0.88
CV	0.78	0.86
Test	0.74	0.83

Table 4.5: Area under the curve (AUC) for models trained with HLA, 40 non-HLA risk scores, and 820 interaction terms. Training (Train), 10-fold stratified cross-validation (CV), and test (Test) AUC is shown. Models were logistic regression (LR), and LASSO.

The higher number of features (861 instead of 41) allowed logistic regression to overfit the training data: the training AUC increased by 6%, but the cross-validation and test AUC decreases by 5%, and 10%, respectively (see table 4.5 and 4.2). LASSO did not overfit, but adding product terms neither increased the CV AUC nor the test AUC. No significant interactions were found.

As described in section 4.3.2, nonlinear models did not perform better than linear models. Yet, strong interactions would have been captured by nonlinear models and should have resulted in a higher prediction performance.

4.3.5 Feature selection

I performed feature selection on the genomics dataset with the aim to estimate prediction performance depending on the number of features, and to rank genes according to their usefulness for predicting T1D.

FSCORE, LASSO-R, SVM RFE, MSVM RFE ($r = 1$, $k = 10$), RF-R ($K = 500$, $L = 13$), and VARBVS ($\pi = 0.1$, $\sigma_\theta^2 = 0.001$) were used to rank features on the complete training set. Each ranking was then evaluated as follows. Logistic regression models for the t top ranked features were trained, where $t = 1 \ldots n$, and n is the maximal number of features. AUC of all models was estimated by 10-fold stratified cross-validation on the training set. Logistic regression was used as predictor since it achieved the highest AUC in previous experiments (see table 4.2), had the lowest runtime, and did not overfit the training data.

Figure 4.8 shows the cross-validation AUC of all ranking methods. Dashed line denotes the AUC of the logistic regression model that was trained with all features (see section 4.3.2). The different rankings performed almost equally. The average AUC of RF-R was lower compared

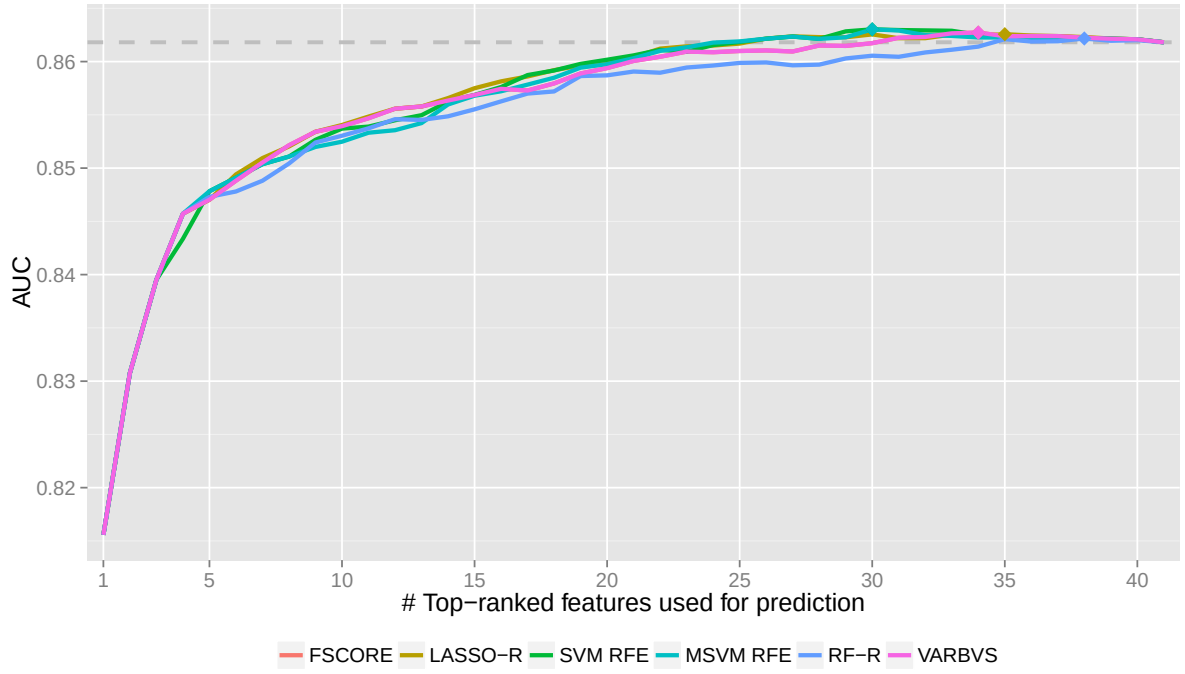


Figure 4.8: Cross-validation area under the curve (AUC) versus number of top ranked features used to train logistic regression model. Dashed line is AUC of full model. Diamonds mark maximal AUC.



Figure 4.9: Feature rankings clustered by Kendall tau rank correlation coefficient. Different colours correspond to different features.

with all other methods, but not significantly (standard error ≈ 0.01). The top 4 features (AUC 0.845) accounted for 98% of the maximal AUC (0.861). Only 24 out of 41 features were sufficient to reach the maximal AUC.

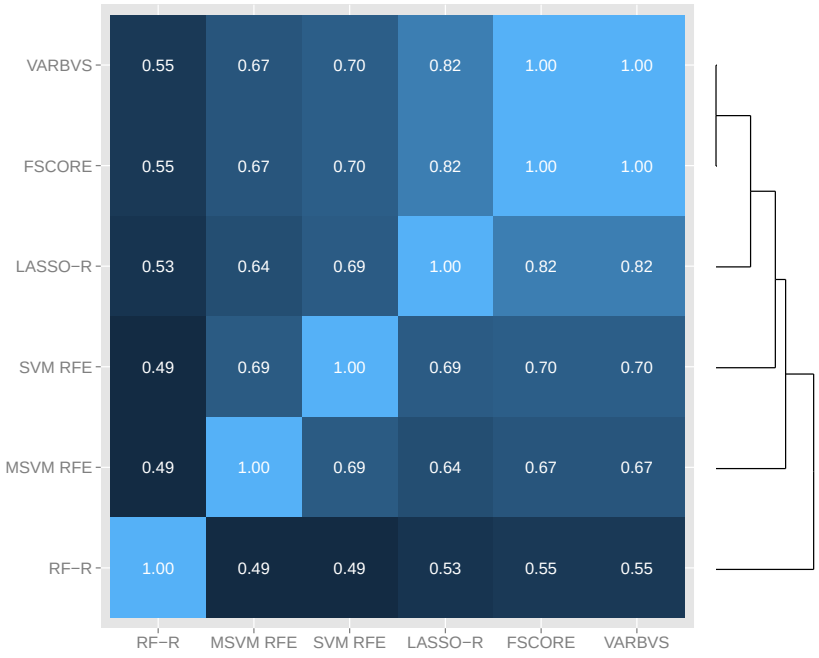


Figure 4.10: Kendall tau rank correlation coefficients and complete linkage clustering of feature rankings.

Figure 4.9 shows the individual rankings sorted by their similarity. All methods ranked HLA as the most useful feature for predicting T1D. INS and PTPN22 were the most useful non-HLA risk scores, which also increased the log odds of T1D the most (see figure 4.4). INS, PTPN22, and ERBB3 (BACH2 for SVM RFE) accounted for 65% of the total increase in AUC of all 40 non-HLA risk scores (see figure 4.8).

Figures 4.10 shows the Kendall τ rank correlation coefficients [68] and the corresponding complete linkage clustering. $\tau > 0.3$ for all rankings, which indicated a significant (p-value < 0.005) similarity. FSCORE and VARBVS ranked the features identically ($\tau = 1.0$). RF-R deviated most from all other rankings.

4.4 Discussion

This study aimed to a) improve predicting the risk for T1D based on HLA and 40 non-HLA genes, and b) to estimate their effect on the development of T1D. I trained different classification models to discriminate between subjects with T1D (cases) and healthy subjects (controls). I

measured classification performance by the area under the ROC curve (AUC), which I estimated by cross-validation and on an independent test set. I tested for gene-gene interactions and used feature selection to rank genes by their usefulness for prediction. The training set contained 4587 cases and 1208 controls from the Type 1 Diabetes Genetics Consortium (T1DGC). The test set contained 765 cases and 423 controls from the BABYDIAB and DiMelli study.

Using only HLA genes allowed for a cross-validation and test AUC of already 0.81 and 0.79, respectively. A logistic regression model trained with HLA and 40 non-HLA genes increased the cross-validation and test AUC to 0.86 and 0.84. For subjects with a high-risk HLA genotype, the cross-validation and test AUC was 0.69 and 0.71.

Weakly significant (p-value $\in (0.005, 0.05]$) HLA/non-HLA interactions were found between HLA and IL2B ($\beta = -0.5$), HLA and C6ORF ($\beta = 0.094$), as well as HLA and CTLA4 ($\beta = 0.091$). No significant interactions were found between non-HLA genes. Neither linear models with interaction terms nor nonlinear models did improve prediction performance. Hence, results suggested the absence of strong gene-gene interactions.

HLA, INS, PTPN22, and ERBB3, were ranked as most useful for predicting T1D and allowed for a cross-validation and test AUC of already 0.85 and 0.82. INS, PTPN22, and ERBB3 accounted for 65% of the total increase in AUC of all 40 non-HLA risk scores.

HLA, PTPN22, INS, IL2R, ERBB3 also had the strongest effect on T1D. The log odds increase per unit was 0.81 for HLA, 0.84 for PTPN22, 0.75 for INS, 0.45 for IL2R, and 0.4 for ERBB3. Since the maximum risk score was 5 for HLA and 2 for non-HLA genes, the maximum odds increase was 73.7 for HLA, 5.4 for PTPN22, 4.5 for INS, 2.5 for IL2R, and 2.3 for ERBB3.

This study has following strengths. Apart from HLA genes, it investigated the cumulative effect of 40 non-HLA genes and thus accounted for considerably more genetic risk factors than previous studies [57, 58]. It compared classification models of different complexity to capture possible gene-gene interactions and nonlinear effects. Feature selection was used to rank genes by their usefulness for predicting T1D, which can help to study causal genetic factors for T1D, and to select a cost-effective combination of genetic markers for prediction. Results apply to the general European population due to the large training set and the availability of an independent test, which was used to evaluate prediction performance.

Winkler et al. [58] used 12 non-HLA genes (ERBB3, PTPN2, IFIH1, PTPN22, KIAA0350, CD25, CTLA4, SH2B3, IL2, IL18RAP, IL10 and COBL) to discriminate between 47 BABYDIAB children with T1D and 1243 healthy children. They assigned a risk score (0, 1, or 2) to each gene and defined a child's risk for T1D as the unweighted sum of all 12 risk scores. The AUC was 0.59 for the complete data set and 0.66 for children with a high-risk HLA genotype. Better results could have been achieved by using more than 12 non-HLA genes, or by using a logistic regression model to learn gene-specific weights. Their findings only apply to the BABYDIAB

cohort due to the small number of cases (47) and the lack of an independent test set.

Valdes et al. [57] predicted T1D using HLA, INS, and CTLA4 genotypes. As in this study, they used samples from the T1DGC to train logistic regression models for discriminating between 1015 cases and 318 controls. They estimated classification performance on a holdout set with 318 cases and 318 controls from the T1DGC. The holdout AUC was 0.74 for a logistic regression model that only included HLA genes, which was lower compared with 0.81 of the baseline model from this study (see section 4.3.1). This can be explained by differences in the selection of T1DGC samples for training, or differences in the definition of the HLA risk score. Their model that included HLA, INS and CTLA4 increased the AUC from 0.74 to only 0.75, which suggested a minor effect of INS and CTLA4.

The results from this study confirm that HLA genes are the major genetic risk factors of T1D, followed by PTPN22 and INS [14, 20, 19]. Other studies also report a relative high effect for CTLA4 [69, 58], whereas ERBB3 was ranked as more important in this study (see figure 4.9).

Protein tyrosine phosphatase N22 (PTPN22) is expressed in lymphoid tissue and controls cell signalling. Mutations can increase its phosphatase activity, change the responsiveness of T and B cell receptors, and cause autoimmune diseases [70, 71].

INS encodes for proinsuline, which is expressed in pancreatic beta cells and further processed to insulin. Mutations can result in incorrect processing and non-functional insulin. Subsequent abnormal blood glucose levels can lead to diabetic symptoms [72, 73].

ERBB3 belongs to the receptor tyrosine kinase family. It is expressed on the surface of lymphocytes, where it controls cell differentiation and proliferation. Mutations can impair T-cell proliferation and cause autoimmune diseases [74, 75].

This study has following limitations. Results only apply to Europeans since all training and test samples were from the European Union. The dataset did not contain non-genetic confounders, such as age, breastfeeding, or Caesarian section, which can impact the effect of genes. Samples were assumed to be independent without accounting for possible genetic correlations in the population. Prediction performance of feature rankings was evaluated on samples that were used to rank features. This may have biased the cross-validation AUC (discussed in section 6).

Future studies should investigate a more comprehensive set of genetic and non-genetic risk factors for T1D. Dropping genotyping costs allow for cost-effective models that capture hundreds of genetic loci. Although most genes contribute only little to the overall risk for T1D, their cumulative risk can make T1D prediction more reliable. Environmental factors, such as Caesarian section, breastfeeding, or time to first gluten exposure, have been associated with T1D [9, 76], and can improve prediction. Data from proteomics and metabolomics experiments can be included to study T1D on different levels of organisation.

I developed a logistic regression model for predicting T1D, which uses HLA and non-HLA genotyping data. For HLA alone, the AUC was 0.78. For HLA and 40 non-HLA genes, the AUC was 0.84. A cost-effective model with HLA, PTPN22, INS and ERBB3 achieved an AUC of 0.82. I conclude that HLA genes are the major genetic risk factors and that most non-HLA genes contribute only little to the overall risk of T1D, but can improve T1D prediction.

5 Proteomics dataset

This chapter will outline the data analysis of the proteomics dataset.

5.1 Introduction

Autoantibodies against pancreatic islet beta cells indicate the development of T1D [12], but the time from their first appearance (seroconversion) to the onset of T1D varies between only few months and several years [77]. Predicting the pace of T1D progression and understanding the understanding the underlying factors is of great medical interest.

Studies tried to predict the pace of T1D progression solely based on the type and the number of islet autoantibodies [55, 56]. Achenbach et al. [9] found genetic and environmental factors to better discriminate between children that developed T1D within 3 years after seroconversion (rapid progressors) and children who remained non-diabetic for more than 10 years (slow progressors). However, they analysed only 12 genes, which do not account for the rich set of factors that may influence T1D progression.

In the second half of my thesis, I contributed to a study that aimed to detect potential biomarkers for discriminating between slow and rapid progressors, and to get a better understanding of factors that determine the pace of T1D progression. The study was a joint research project between the Institute of Diabetes Research and the Research Unit Protein Science at the Helmholtz-Zentrum Munich. An untargeted proteomics analysis of blood serum peptides was performed. It was hypothesised that peptides provide a more immediate view on the biochemical mechanisms of T1D progression than genes since gene expression is subject to complex regulatory control and environmental changes. Blood samples were taken from 15 slow progressors and 15 rapid progressors, and high-throughput mass-spectrometry [78] was used to measure the concentration of 4384 peptides. My task was to develop models for discriminating between slow and rapid progressors, and to identify markers for T1D progression. I trained and evaluated classification models, and employed feature selection to identify the most discriminative peptides. The dataset on which I carried out the analysis will henceforth be referred to as *proteomics dataset*.

5.2 Dataset

The Institute of Diabetes Research conducted the prospective BABYDIAB cohort study [61]. 1687 children from Germany, whose mother or father had T1D, were monitored from birth up to age 22. Recruitment started in 1989 and ended in 2001. Venous blood samples were taken after 9 months, 2, 5, 8, 11, 14, 17, and 20 years after birth, and tested for autoantibodies IAA, GADA, IA-2A, and ZnT8A. A child was defined as antibody positive if at least one autoantibody was found in two consecutive blood samples. Onset of diabetes was determined by the fasting plasma glucose level [79].

The proteomics dataset contained 15 BABYDIAB children that developed T1D within 3 years after the first occurrence of autoantibodies (rapid progressors), and 15 BABYDIAB samples that remained non-diabetic of at least 10 years (slow progressors). Each group contained 6 females and 9 males. Note that the genomics training set contained T1DGC samples, and the test set parents of BABYDIAB children, thus no BABYDIAB children (see section 4.2).

The Research Unit Protein Science determined the concentration of peptides in venous blood samples by liquid chromatography mass spectrometry. Here, peptides are first injected into a liquid chromatography tube, whose walls are coated with chemical particles. Peptides interact differently with these particles depending on their physicochemical properties, and thus leave the tube at different time points. In the mass spectrometer, peptides are ionised and separated by their charge in an electromagnetic field, such that their mass to charge ratio can be determined by a detector. To identify peptides, the experimentally determined mass to charge ratios are compared with theoretical mass to charge ratios of known peptides in a database. [80]

The Research Unit Protein Science used the enzyme trypsin to cleave proteins to smaller peptides, which were quantified with the LTQ-Orbitrap XL mass spectrometer. The Mascot software [81] was used to identify peptides, together with the Ensembl [82] and Mascot-integrated decoy database. The resulting proteomics dataset contained the concentration of 4384 peptides of 30 BABYDIAB samples, from which 15 were rapid progressors and 15 slow progressors.

5.3 Results

The following sections present the classification performance of models that were trained with all peptides and on a selected subset of peptides.

5.3.1 Classification without feature selection

In the proteomics dataset, the number of features ($p = 4384$) clearly exceeded the number of samples ($n = 30$), which is also referred to as *small n large p* problem. In contrast, the genomics dataset contained more samples than features (see section 4.2), and linear as well as nonlinear model achieved the highest prediction performance without feature selection. I therefore evaluated the same classification models on the proteomics dataset using all features for training.

	LASSO	LIN SVM	RBF SVM	RF
AUC	0.54	0.40	0.32	0.61
SD	0.18	0.21	0.19	0.19

Table 5.1: 5-fold stratified cross-validation AUC and standard deviation (SD) on the proteomics dataset without feature selection. Classifiers were LASSO, SVM with linear kernel (LIN SVM), SVM with RBF kernel (RBF SVM), and random forest (RF).

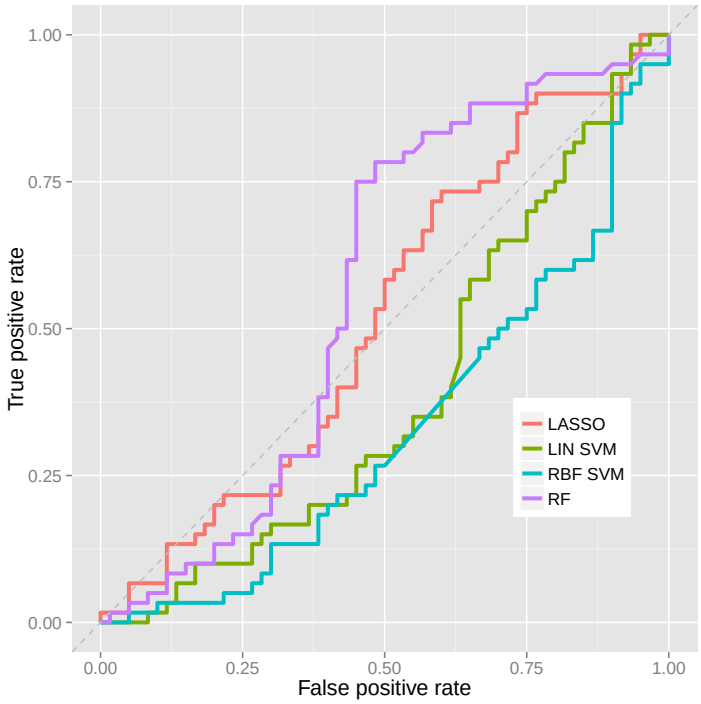


Figure 5.1: True positive rate versus false positive rate estimated by 5-fold stratified cross-validation on the proteomics data without feature selection. Models were LASSO, SVM with linear kernel (LIN SVM), SVM with RBF kernel (RBF SVM), and random forest (RF).

LASSO and SVM with linear kernel (LIN SVM) were used as linear models, SVM with RBF kernel (RBF SVM) and random forest (RF) as nonlinear models. Logistic regression without

regularization was excluded from this analysis since 30 samples were insufficient for training. In contrast to the genomics dataset, no independent test set was available for the proteomics dataset. 4 repetitions of 5-fold stratified cross-validation were used to estimate the AUC of classifiers. Thus, 24 samples were randomly selected for training and 6 for testing. Rapid progressors and slow progressors had equal class weights.

The cross-validation AUC of all classifiers did not significantly deviate from 0.5 (see table 5.1 and figure 5.1). Thus, the number of samples did not suffice to discover discriminative features, such that models did not perform better than a random predictor.

5.3.2 Classification with feature selection

Since using all features for parameter learning resulted in a poor prediction performance, I evaluated ranking feature selection methods to shrink the feature space to the most useful features for classification.

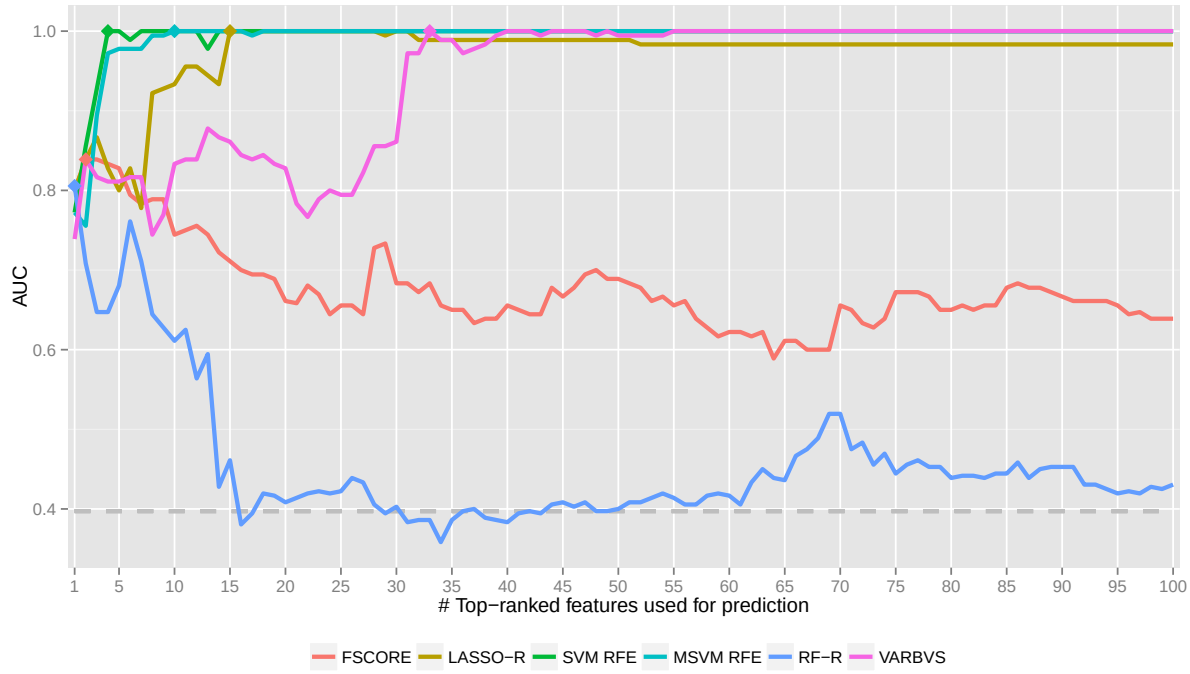


Figure 5.2: Cross-validation area under the curve (AUC) versus number of top ranked features used to train support vector machine. Dashed line is AUC of full model. Diamonds mark maximal AUC.

Features were ranked by FSCORE, SVM RFE, MSVM RFE ($r = 4$, $k = 5$), LASSO-R, RF-R ($K = 10$, $L = 20$), and VARBVS ($\pi = 0.1$, $\sigma_\theta^2 = 0.1$). Rankings were evaluated as described in

section 4.3.5. LIN SVM was used as predictor, since it is robust to overfitting and scales well for a small sample size. AUC was estimated by 4 repetitions of 5-fold stratified cross-validation.

	# Features	AUC	SD
FSCORE	2	0.84	0.14
LASSO-R	15	1.00	0.00
SVM RFE	4	1.00	0.00
MSVM RFE	10	1.00	0.00
RF-R	1	0.81	0.18
VARBVS	33	1.00	0.00

Table 5.2: Maximum 5-fold cross-validation AUC of feature ranking methods. Number of features, mean cross-validation AUC, and standard deviation (SD) is shown.

Figure 5.2 shows the cross-validation AUC for models trained with the top ranked features. Dashed line indicates the AUC (0.4) of LIN SVM trained with all features, to which all curves converged for $x = 4384$. For all rankings, diamonds mark the highest AUC for the smallest subset of features (see table 5.2). All rankings could find feature subsets with an AUC significantly above 0.4. Models that included only the top ranked feature had an AUC already greater than 0.73. Including more features ranked by FSCORE and RF-R decreased the AUC. Adding more feature ranked by SVM RFE or MSVM RFE monotonously increased the AUC to 1.0, which did not drop for models with up to 100 features.

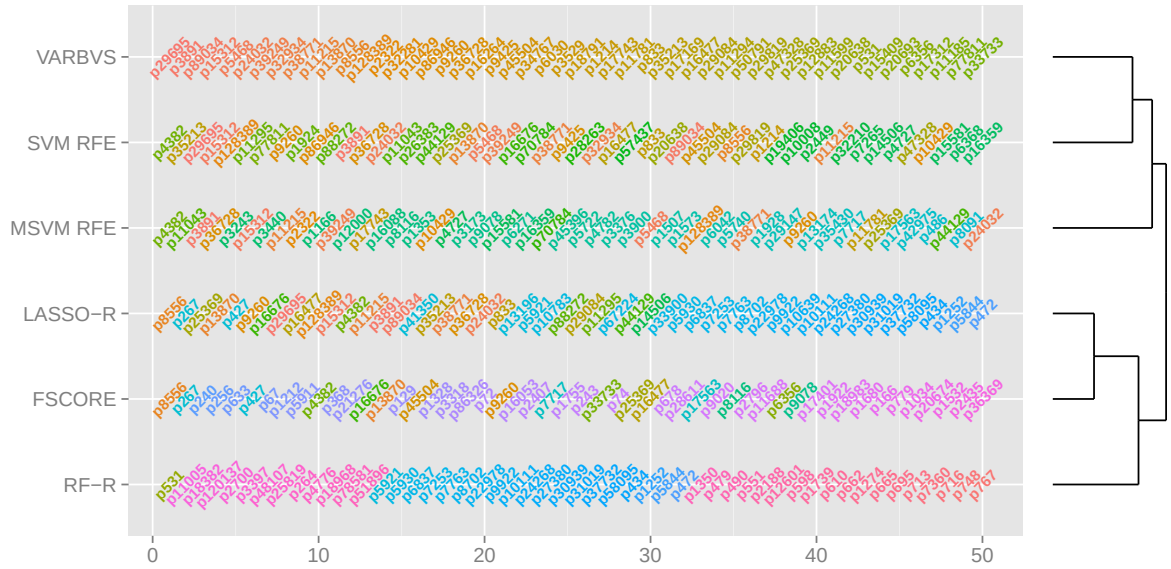


Figure 5.3: Feature rankings clustered by Kendall tau rank correlation coefficient. Different colours correspond to different features.

Figure 5.3 shows the top 50 features of each ranking, and figure 5.4 their corresponding τ rank correlation coefficients. The rankings were more diverse as compared with the genomics dataset

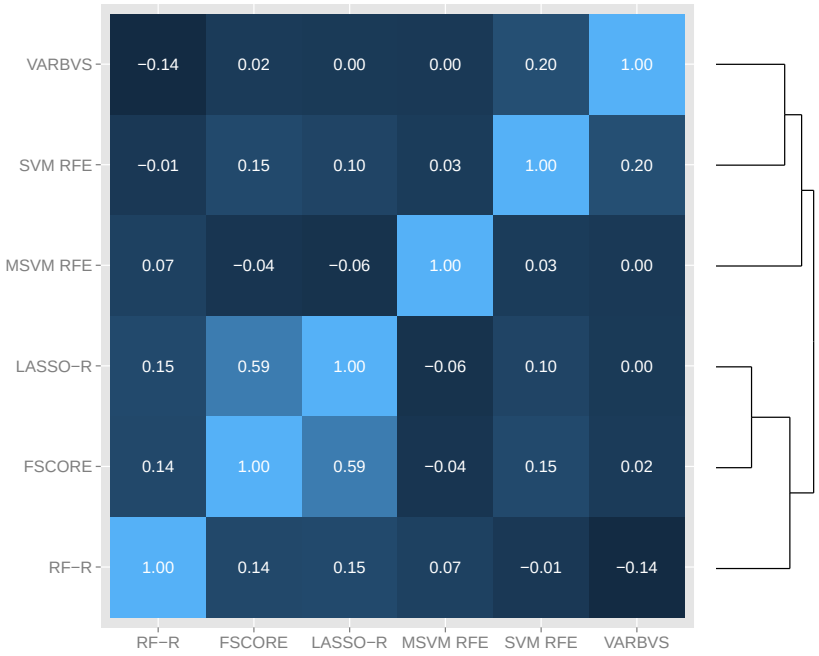


Figure 5.4: Kendall tau rank correlation coefficients and complete linkage clustering of feature rankings.

(see figure 4.10). FSCORE and LASSO-R computed the most similar rankings ($\tau = 0.59$) and agreed on the first two features (p8556, p267). MSVM RFE had the highest distance to the remaining methods. MSVM RFE and SVM RFE ranked p4382 as most useful, but were only weakly correlated ($\tau = 0.03$). Hence, the correlation between rankings did not correspond to the correlation of their AUC (see figure 5.2). This was due to the high number of correlated features in the dataset, which allowed for different rankings with a similar AUC.

5.3.3 Biochemical function of selected proteins

Table 5.3 lists peptides and their corresponding proteins that were ranked as important by SVM RFE and MSVM RFE. The rankings contain multiple proteins which either play a role in the cardiovascular system or the complement system. The following sections will review their function and medical implications.

5.3.3.1 Proteins of the cardiovascular system

Both SVM RFE and MSVM RFE ranked apolipoprotein A-IV (APOA4) as most useful, which was up regulated in rapid progressors (see figure 5.5). Apolipoproteins are lipid binding proteins. They are the major components of lipoprotein particles—including chylomicrons, LDL,

	SVM RFE			MSVM RFE		
Nr	Id	Symbol	Name	Id	Symbol	Name
1	p4382	APOA4	apolipoprotein	p4382	APOA4	apolipoprotein
2	p35213	FN1	fibronectin 1	p11043	HP	haptoglobin NA
3	p29695	MASP1	mannan-binding	p3891	SAA1	serum amyloid A1
4	p15312	SAA1	serum amyloid A1	p36728	C4BPB	complement
5	p128389	ATXN2	ataxin 2	p3243	FN1	fibronectin 1
6	p11295	ITIH1	inter-alpha-trypsin	p15312	SAA1	serum amyloid A1
7	p77811	SPTA1	spectrin, alpha,	p3440	FN1	fibronectin 1
8	p9260	FCN3	ficolin	p11215	FCN3	ficolin
9	p1924	ALB	albumin NA	p2322	SAA1	serum amyloid A1
10	p86946	CP	ceruloplasmin	p1166	C1QB	complement
11	p88272	ALB	albumin NA	p39249	SAA1	serum amyloid A1
12	p3891	SAA1	serum amyloid A1	p12000	FN1	fibronectin 1
13	p36728	C4BPB	complement	p17743	CFHR4	complement
14	p24032	PCYOX1	prenylcysteine	p16088	FN1	fibronectin 1
15	p11043	HP	haptoglobin NA	p8116	FN1	fibronectin 1
16	p26383	CNDP1	carnosine	p1353	APOA4	apolipoprotein
17	p44129	MRPS34	mitochondrial	p10429	SAA1	serum amyloid A1
18	p25369	FN1	fibronectin 1	p4727	HP	haptoglobin NA
19	p13870	ITIH1	inter-alpha-trypsin	p3173	FN1	fibronectin 1
20	p5468	SAA1	serum amyloid A1	p9078	FN1	fibronectin 1
21	p39249	SAA1	serum amyloid A1	p15581	A2M	alpha-2-macroglobulin
22	p16676	F5	coagulation	p6271	FN1	fibronectin 1
23	p70784	SLC4A1	solute carrier	p16359	CFH	complement
24	p38771	CP	ceruloplasmin	p70784	SLC4A1	solute carrier
25	p9425	BTD	biotinidase NA	p45396	APOA4	apolipoprotein
26	p28263	QSOX1	quiescin Q6	p5722	FN1	fibronectin 1
27	p32934	CFHR4	complement	p4782	FN1	fibronectin 1
28	p16477	F5	coagulation	p3376	FN1	fibronectin 1
29	p57437	ALB	albumin NA	p3900	APOH	apolipoprotein H
30	p833	APOC4	apolipoprotein	p5468	SAA1	serum amyloid A1
31	p20638	C4BPB	complement	p1507	APOA4	apolipoprotein
32	p89034	SPTA1	spectrin, alpha,	p1573	FN1	fibronectin 1
33	p45504	MASP1	mannan-binding	p128389	ATXN2	ataxin 2
34	p29084	GPR115	G	p6042	C3	complement
35	p8556	MASP1	mannan-binding	p5740	FN1	fibronectin 1
36	p29819	FGA	fibrinogen alpha	p38771	CP	ceruloplasmin
37	p1214	APOC4	apolipoprotein	p1928	APOA4	apolipoprotein
38	p19406	SAA2	serum amyloid A2	p29147	SLC4A1	solute carrier
39	p10008	VTN	vitronectin NA	p9260	FCN3	ficolin
40	p2449	SERPIND1	serpin peptidase	p13174	FN1	fibronectin 1

Table 5.3: Peptides ranked by SVM RFE and MSVM RFE. Gene symbol and description refers to the protein, to which the peptide was allocated.

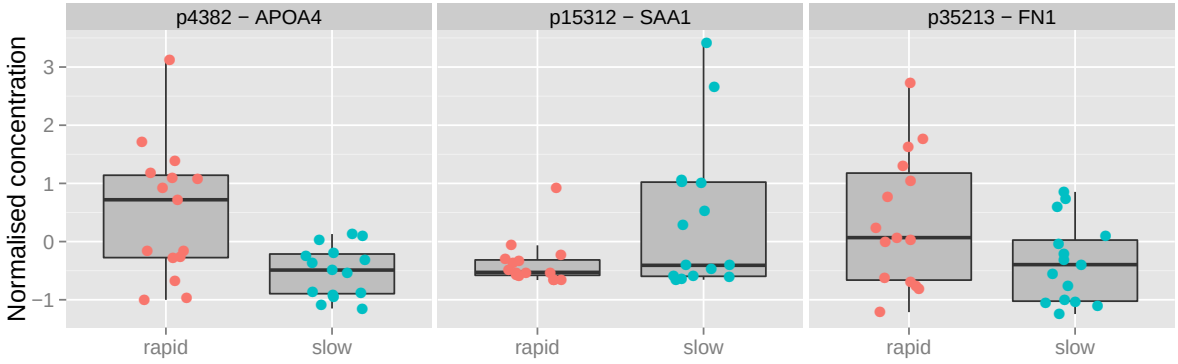


Figure 5.5: Mean normalised peptide concentrations of slow and rapid progressors. The allocated proteins (APOA4, SAA1, FN1) are part of the cardiovascular system.

or HDL—for transporting lipids in blood and lymph, and function as cell surface receptors as well as cofactors for enzymes. Mooradian et al. [83] observed a lower expression of APOA in mice with T2D, which may decrease HDL concentration and lead to cardiovascular disease, the leading cause of mortality in diabetes [84]. Mooradian et al. [83] found APOA and APOB as strong biomarkers for diabetic retinopathy: damages to the retina mainly caused by T1D or T2D after 10 years of onset.

The apolipoprotein serum amyloid A-1 (SAA1) is involved in HDL metabolism and cholesterol homeostasis. Specific isoforms are overexpressed in the acute phase of inflammation, recruit immune cells to inflammatory sites, and activate enzymes that degrade extracellular matrix. SAA is therefore a marker for the acute phase of inflammation, which may contribute to the metabolic syndrome (hypertension, dyslipidaemia, central obesity, high fasting glucose level), which is frequently observed in patients with T2D [85, 86].

Fibronectin-1 (FN1) binds components within the extracellular membrane and controls cell adhesion, growth, migration, as well as differentiation. Roy et al. [87] found that high glucose levels increase the expression of FN, which may account for the higher concentration of FN observed in diabetic tissue. FN resides in the extracellular matrix, a layer below the surface tissue (epithelium) of organs such as blood vessels. Vascular damage can therefore release FN from the extracellular matrix and increase its concentration in the blood. Kanters et al. [88] report a significantly higher concentration of FN in the blood of patients with T1D or T2D, compared with healthy subjects and subjects with other cardiovascular diseases. They suggested FN as biomarker for diabetes.

5.3.3.2 Proteins of the complement system

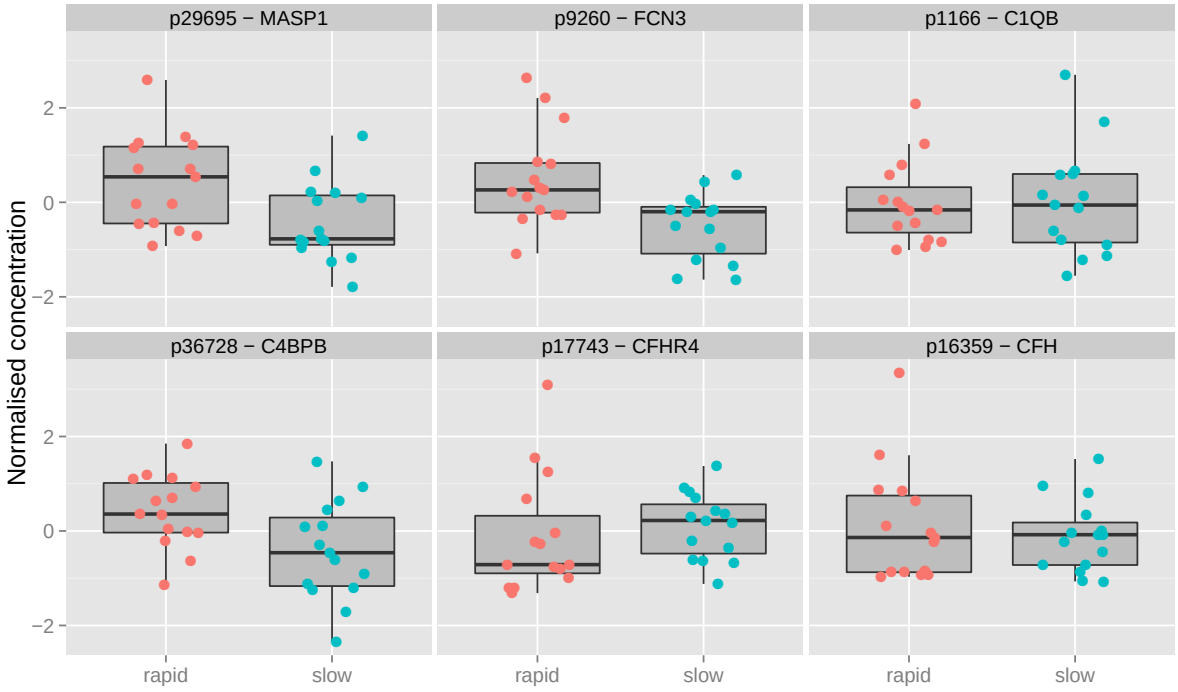


Figure 5.6: Mean normalised peptide concentrations of slow and rapid progressors. The allocated proteins (MASP1, FCN3, C1QB, C4BPB, CFHR4, CFH) are part of the complement system.

MASP1, FCN3, C4BP, C1QB, CFH, CFHR4, and C3 are members of the complement system, which is part of the innate (non-adaptive) immune system, and plays a critical role in the development of T1D and T2D [89, 90].

The complement system (see figure 5.7) 'complements' the function of the adaptive immune system by non-specific immune responses. It encompasses a set of plasma proteins, e.g. C3, and cleavage products, e.g. C3a and C3b. Three different pathways can activate the complement system. The classical pathway starts with the detection of an antigen/antibody complex. The lectin pathway functions similar to the classical pathway, but is activated by the recognition of mannose patterns on the surface of pathogenic cells. The alternative pathway requires spontaneous hydrolysis of C3. The final product of all three pathways is C3 convertase, which cleaves C3 to C3a and C3b. C3b attaches to the surface of pathogens, which are thereby recognised by immune cells and destroyed. C3a initiates cleavage reactions, which result in a) the recruitment of immune cells to the site of immunity and b) the formation of the membrane attack complex, which introduces pores into the membrane of pathogens and causes cell death. [89]

Mannose binding lectin serine protease 1 (MASP1) and ficolin 3 (FCN3) are enzymes of the lectin activation pathway. Mannose binding lectin (MBL) and ficolin proteins recognise mannose patterns that are present only on the surface of pathogens, and not on human cell. They activate

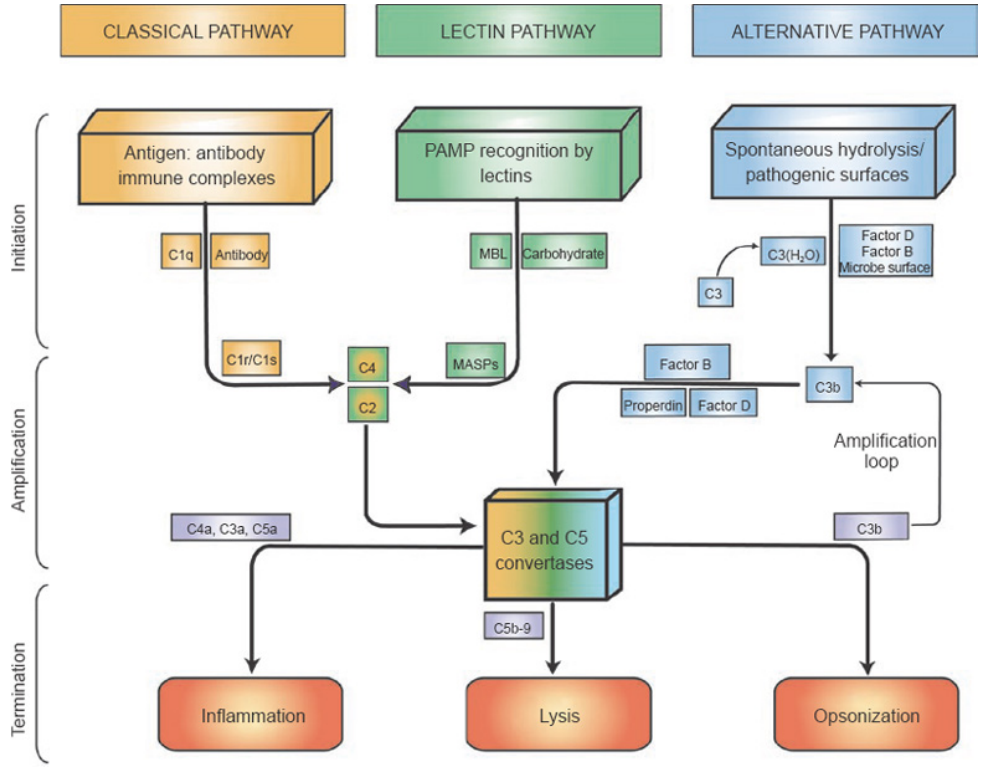


Figure 5.7: Complement system and the innate immune response. Three different pathways can initiate complement: the classical pathway, the lectin pathway, and the alternative pathway. Initiation pathways merge at the formation of C3 and C5 convertase. Cleavage of C3 and C5 leads to a cascade of further cleavage reactions. Complement activates inflammatory processes, marks pathogens for degradation by lymphocytes (opsonization), and expresses membrane attack complex for cell lysis. Figure taken from [1].

MASP1 and MASP2, which cleave C4 to C4a and C4b, and C2 to C2a and C2b. C4b and C2b finally bind to the cell surface and form the C3 convertase. Abnormal MBL concentrations are associated with infectious and autoimmune diseases [91]. Hansen [92] and Bouwman et al. [93] found significantly higher levels of MBL in patients with T1D compared with healthy subjects. High MBL levels also indicate diabetic nephropathy: damages to the kidney caused by elevated blood glucose levels [94, 95]. Hovind et al. [96] found that unusual MBL levels after the diagnosis of T1D are predictive for the later development of diabetic nephropathy.

Complement component 1q B and (C1QB) and complement component 4 binding protein B (C4BPB) belong to the classical pathway. Elevated concentrations can indicate a strong immune response [97].

Complement factor H (CFH) is a regulator of the alternative pathway and has been associated with diseases [98].

Complement factor H-related 4 (CFHR4) is one of five factor-H related proteins, which bind C3b after the initial cleavage of C3 and regulate the complement system. Abnormal concentrations can indicate immune diseases [99, 100, 101, 102].

5.4 Discussion

This study aimed to detect potential biomarkers for discriminating between slow and rapid progressors, and to get a better understanding of factors that determine the pace of T1D progression. It compared the concentration of 4384 blood serum peptides of 15 rapid progressors (developed T1D within 3 years after the first occurrence of autoantibodies) with 15 slow progressors (remained non-diabetic for at least 10 years). My task was to develop models for discriminating between slow and rapid progressors, and to identify markers for T1D progression. I found that only four peptides (allocated to APOA4, FN1, MASP1, and SAA1) can perfectly predict T1D progression. Further peptides were highly discriminative, and recent studies indicated a critical role of high-ranked peptides in the cardiovascular system or complement system.

This study differs from previous studies in the following three points. First, it investigated the role of peptides instead of genes since it was assumed that peptides better reflect the biochemical activity of cells. Second, it analysed the joint effect of 4384 peptides, and thus accounted for more potential risk factors as compared with previous studies. Third, established feature ranking methods were applied to discover small subsets of discriminative peptides in a high-dimensional dataset with few samples.

Many high-ranked peptides were disease associated. APOA4, FN1, and SAA1 are associated with cardiovascular diseases, which are often observed in patients with T1D or T2D. MASP1, FCN3, C1QB, C4BPB, CFHR4, and CFH are associated with infectious as well as autoimmune diseases. They belong to the complement system, which activates innate immune responses. These findings suggest a major role of the innate immune system in the progression of T1D. In contrast, Achenbach et al. [9] identified genes that are mainly involved in the regulation of lymphocytes of the adaptive immune system.

This study has following limitations. First, it analysed only 30 diabetic patients from Germany, which had a family history of T1D. The sample thus does not represent the general population, and findings may not apply to other ethnicities or patients without history of T1D. Prediction performance of feature rankings was evaluated on samples that were used to rank features. This may have biased the cross-validation AUC (discussed in section 6). The high standard error of the AUC estimator further limits generalisability of the results. Second, many peptides in the proteomics dataset were highly correlated. Only 276 distinct proteins covered all 4384 peptides, and peptides that were allocated to the same protein were often highly correlated.

Clustering peptides by their correlation coefficient and pre-selecting a single cluster member could have improved robustness of the results. Third, the study did not adjust for genetic and environmental confounders, such as Caesarian section or breastfeeding, which have been associated with T1D progression [9, 76].

The presented feature rankings were used to select a small set of peptides, which will be investigated in a follow-up study for a larger sample. Future studies may analyse blood samples from different time points, which can help to study changes in the concentration of proteins during the development of T1D. Including genetic risk factors as well as environmental factors which may account for changes in peptide concentrations can further improve to predict T1D progression.

In conclusion, I identified a small set of blood serum peptides that perfectly discriminates between slow or rapid progression to T1D after the occurrence of pancreatic islet autoantibodies.

6 Discussion and Conclusions

This thesis aimed to use machine learning methods for predicting T1D from high-throughput data. In the first half of this thesis, I studied a genomics dataset with 41 genetic risk factors of T1D positive and T1D negative subjects. In the second half of this thesis, I studied a proteomics dataset with 4384 blood serum peptides of 15 rapid T1D progressors and 15 slow T1D progressors. On both datasets, I trained logistic regression, LASSO, and linear SVM as linear classification models, as well as SVM with RBF kernel and random forest as nonlinear models. I evaluated classification performance by the AUC, which I estimated by k-fold stratified cross-validation and on an independent test set for the genomics dataset. I further applied feature selection to search for subsets of features with the best prediction performance, and to rank features by their usefulness for predicting T1D. I compared an univariate feature ranking method (FSCORE) with multivariate methods (SVM RFE, MSVM RFE, LASSO-R, RF-R, VARBVS), which I evaluated by the cross-validation AUC of models that were trained with features in the ranking list.

For the genomics dataset, logistic regression achieved the highest cross-validation and test AUC of 0.86 and 0.84, respectively. LASSO did not perform better, which indicated that 5795 samples were sufficient to fit parameters for 41 features without regularization. Likewise, nonlinear models did not improve classification, which indicated a linear dependency between features and the class variable, and the absence of feature interactions. The AUC of feature rankings increased monotonously with the number of features, which showed that all features were associated with the class variable, and that shrinking the feature space did not improve parameter estimation. Univariate and multivariate methods performed similarly, which indicated independence of features. Feature rankings were strongly correlated (τ rank correlation coefficient > 0.5) and were identical for the first three features, which accounted for already 98% of the maximal AUC.

In contrast to the genomics dataset, the number of features clearly exceeded the number of samples in the proteomics datasets, which posed new challenges for classification and feature selection.

The AUC of models trained with all peptides was not significantly greater than 0.5. Even SVM, which is commonly applied for high-dimensional problems such as text classification, did not perform better than a random predictor.

Feature selection identified perfectly discriminating feature subsets, as well as single highly pre-

dictive features. The univariate FSCORE method ranks features by their relevance for prediction when considered individually. The AUC was 0.8 for the top-ranked feature (p8556) and 0.84 for the model that also included the second feature (p267). Although subsequent features were strongly correlated with the class variable, they lowered the AUC. These results suggested that FSCORE failed to remove redundant features, which carry the same information about the class variable and thus do not improve prediction.

Multivariate methods rank features by their usefulness for prediction when considered in combination. The top-ranked feature of SVM RFE and MSVM RFE was p4382, which had an AUC of 0.77. This was lower compared with the top-ranked feature of FSCORE, but subsequent features monotonously increased the AUC to 1.0, which did not drop for models with up to 100 features. Although most high-ranked features were only weakly correlated with the class variable (see figure 5.5 and 5.6), they improved prediction in combination. Hence, SVM RFE and MSVM RFE succeeded in removing redundant features, such that high-ranked features contained orthogonal information about the class variable.

SVM RFE achieved the best prediction performance, followed by MSVM RFE, LASSO-R, VARBVS, and RF-R. VARBVS was the fastest method, followed by LASSO-R, RF-R, SVM RFE, and SVM MRFE.

LASSO-R thus provided a good trade-off between prediction performance and runtime, and may scale well for datasets with thousands of features. However, the number of features selected by LASSO is bounded by the sample size m. To obtain a smooth ranking of more than m features, ELASTIC-NET [43] or RANDOM-LASSO [44] can be used for computing the LASSO path.

RF-R had the least prediction performance. It estimates the importance of features on the out-of-bag samples of K ensemble trees (see section 3.3.5). $K = 10$ was optimal for the proteomics dataset and the expected number of out-of-bag samples per tree was about 10 for 30 samples. The total number of out-of-bag samples may thus have been insufficient for estimating the importance of features, and RF-R may perform better for higher K and on datasets with more samples.

This thesis exhibits methodological limitations. First, I used only SVM as classification model to evaluate feature rankings. As discussed in section 3.3.2, the SVM RFE selection criterion minimises the SVM cost function, such that the resulting rankings are optimised towards the SVM model. This may have biased the evaluation of SVM RFE and MSVM RFE.

Second, the AUC estimator for evaluating classification performance had a high standard error due to the small sample size, which limits generalisability of the results. Rankings therefore need to be curated manually, e.g. by GO terms [103] or the KEGG database [104]. Instead of evaluating ranking methods by their prediction performance, their false positive rate may be assessed on an artificial dataset with non-informative features [6].

Third, I estimated the AUC of feature rankings on samples that were also used to rank features. Feature rankings were computed on all samples, and their classification performance evaluated afterwards by k-fold stratified cross-validation. Although the classifier was trained and tested on distinct samples, it was tested on samples that were used to rank features. This may have resulted in a too optimistic estimate of the classification performance. Xiong et al. [105] and Guyon et al. [39] evaluated rankings in the same way and could identify small feature subsets with almost zero leave-one-out error. However, Ambroise and McLachlan [106] could later show that the estimated classification error was too optimistic, and proposed using disjoint datasets for ranking and evaluation to avoid selection bias. For small datasets, they suggested 0.642 bootstrap or k-fold cross-validation. In the latter case, features are ranked on k-1 folds, and their classification performance is estimated on the remaining fold.

Future studies may take ensemble feature ranking methods into account. Small changes in the training data can considerably change feature rankings, which is critical for applications such as biomarker discovery. Ensemble methods have been shown to increase robustness by combining multiple feature rankings [7]. Abeel et al. [107] generated different SVM RFE rankings for bootstrap samples from the training data, and combined them to a single ensemble ranking. In this way, they could improve robustness by 30% and classification performance by 15%.

Future studies may also investigate wrapper and embedded methods for feature selection if the main goal is not to rank features, but to optimise classification performance. Wrapper search heuristics, such as floating search, plus q-take-away-r, or beam search, explore the feature space more exhaustively and can thus achieve a higher classification performance. If the sample size is small, embedded methods may be more appropriate since they are less prone to overfitting. L0 norm regularization [108] is an alternative to L1 norm regularization (LASSO) for feature selection. Embedded methods for SVM are penalized SVM [109] and kernel penalized SVM [110], which can be more efficient than SVM RFE.

The feature selection methods discussed in this thesis identified only associations between features and the class variable. Instead, causal feature selection methods [111, 112] aim to identify causal relationships, which can improve prediction and knowledge discovery. Knowing that specific mutations in a gene cause T1D can be used to predict the risk for developing T1D. In contrast, an elevated concentration of a protein can be the consequence, i.e. a symptom, of T1D. Its concentration may also be influenced by the current nutrition, temperature, or measurement errors. Taking into account such confounding variables can make the protein a more reliable biomarker for T1D. Causal feature selection methods model causal relationships with Bayesian networks: directed acyclic graphs, which represent dependencies between random variables. They learn the structure of a Bayesian network, as well as the Markov Blanket of the target variable Y, i.e. all variables that are useful for predicting Y in the context of other variables [113, 114]. Prior knowledge about variables, for instance the sequence of proteins in a

biochemical pathway, can be incorporated easily into the learning process. The resulting network helps to discover causal relationships, to test biological hypotheses, and to infer new knowledge about all variables in the network.

In conclusion, I employed machine learning methods to analyse two datasets from T1D studies. On a genomics dataset, I trained a logistic regression model using 41 genetic risk factors of T1D positive and T1D negative subjects, which achieved an AUC of 0.84 on an independent test set. I used feature selection to identify a cost-effective model with only four genetic risk factors, which achieved a test AUC of 0.82. Most genes contributed only little to the overall risk of T1D, but improved prediction in combination. On a high-dimensional proteomics dataset with only few samples, SVM RFE achieved the highest prediction performance among different feature selection methods. It identified a set of only four peptides that perfectly discriminated between slow T1D progressors and rapid T1D progressors. Further peptides could be curated manually with literature and will be investigated in a follow-up study. Machine learning provides powerful models for identifying pattern in high-throughput data, which can help to better understand complex human diseases, to discover disease markers, and to make medicine more predictive and personalised.

Bibliography

[1] Jason R Dunkelberger and Wen-Chao Song. Complement and its role in innate and adaptive immune responses. *Cell research*, 20(1):34–50, January 2010. ISSN 1748-7838. doi: 10.1038/cr.2009.139. URL http://www.ncbi.nlm.nih.gov/pubmed/20010915. (document), 5.7

[2] Hiroaki Kitano. Systems biology: a brief overview. *Science (New York, N.Y.)*, 295(5560): 1662–4, March 2002. ISSN 1095-9203. doi: 10.1126/science.1069492. URL http://www.ncbi.nlm.nih.gov/pubmed/11872829. 1

[3] Leroy Hood, James R Heath, Michael E Phelps, and Biaoyang Lin. Systems biology and new technologies enable predictive and preventative medicine. *Science (New York, N.Y.)*, 306(5696):640–3, October 2004. ISSN 1095-9203. doi: 10.1126/science.1104635. URL http://www.ncbi.nlm.nih.gov/pubmed/15499008. 1

[4] P. Larranaga. Machine learning in bioinformatics. *Briefings in Bioinformatics*, 7(1): 86–112, February 2006. ISSN 1467-5463. doi: 10.1093/bib/bbk007. URL http://bib.oxfordjournals.org/cgi/doi/10.1093/bib/bbk007. 1

[5] Adi L Tarca, Vincent J Carey, Xue-wen Chen, Roberto Romero, and Sorin Drăghici. Machine learning and its applications to biology. *PLoS computational biology*, 3(6):e116, June 2007. ISSN 1553-7358. doi: 10.1371/journal.pcbi.0030116. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1904382&tool=pmcentrez&rendertype=abstract. 1

[6] Isabelle Guyon. An Introduction to Variable and Feature Selection 1 Introduction. *Journal of Machine Learning Research*, 3:1157–1182, 2003. 1, 3.3, 6

[7] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics (Oxford, England)*, 23(19):2507–17, October 2007. ISSN 1367-4811. doi: 10.1093/bioinformatics/btm344. URL http://www.ncbi.nlm.nih.gov/pubmed/17720704. 1, 6

[8] Isabelle Guyon, Constantin Aliferis, Greg Cooper, Andre Elisseeff, Jean-Philippe Pellet, Alexander Statnikov, and Peter Spirtes. *Causation and Prediction Challenge Challenges*, volume 2. 2008. ISBN 9780971977723. 1

[9] P Achenbach, M Hummel, L Thümer, H Boerschmann, D Höfelmann, and a G Ziegler. Characteristics of rapid vs slow progression to type 1 diabetes in multiple islet autoantibody-positive children. *Diabetologia*, 56(7):1615–22, July 2013. ISSN 1432-0428. doi: 10.1007/s00125-013-2896-y. URL http://www.ncbi.nlm.nih.gov/pubmed/23539116. 1, 4.4, 5.1, 5.4

[10] Jeffrey A Bluestone, Kevan Herold, and George Eisenbarth. Genetics, pathogenesis and clinical interventions in type 1 diabetes. *Nature*, 464(7293):1293–1300, April 2010. ISSN 0028-0836. doi: 10.1038/nature08933. URL http://dx.doi.org/10.1038/nature08933http://www.nature.com/doifinder/10.1038/nature08933. 1, 2

[11] Tom L van Belle, Ken T Coppieters, and Matthias G von Herrath. Type 1 diabetes: etiology, immunology, and therapeutic strategies. *Physiological reviews*, 91(1):79–118, January 2011. ISSN 1522-1210. doi: 10.1152/physrev.00003.2010. URL http://www.ncbi.nlm.nih.gov/pubmed/21248163. 1, 2

[12] Michael Hummel, Ezio Bonifacio, Sandra Schmid, Markus Walter, Annette Knopff, and Anette-G Ziegler. Brief communication: early appearance of islet autoantibodies predicts childhood type 1 diabetes in offspring of diabetic parents. *Annals of internal medicine*, 140(11):882–6, June 2004. ISSN 1539-3704. URL http://www.ncbi.nlm.nih.gov/pubmed/15172902. 1, 5.1

[13] David W Cooke and Leslie Plotnick. Type 1 diabetes mellitus in pediatrics. *Pediatrics in review / American Academy of Pediatrics*, 29(11):374–84; quiz 385, November 2008. ISSN 1526-3347. doi: 10.1542/pir.29-11-374. URL http://www.ncbi.nlm.nih.gov/pubmed/18977856. 2

[14] Anette-G Ziegler and Gerald T Nepom. Prediction and pathogenesis in type 1 diabetes. *Immunity*, 32(4):468–78, April 2010. ISSN 1097-4180. doi: 10.1016/j.immuni.2010.03.018. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2861716&tool=pmcentrez&rendertype=abstract. 2, 4.4

[15] S. Lindley, C. M. Dayan, A. Bishop, B. O. Roep, M. Peakman, and T. I.M. Tree. Defective Suppressor Function in CD4+CD25+ T-Cells From Patients With Type 1 Diabetes. *Diabetes*, 54(1):92–99, January 2005. ISSN 0012-1797. doi: 10.2337/diabetes.54.1.92. URL http://diabetes.diabetesjournals.org/cgi/doi/10.2337/diabetes.54.1.92. 2

[16] Gabrielle G M Pinkse, Odette H M Tysma, Cees A M Bergen, Michel G D Kester, Ferry Ossendorp, Peter A van Veelen, Bart Keymeulen, Danny Pipeleers, Jan W Drijfhout, and Bart O Roep. Autoreactive CD8 T cells associated with beta cell destruction in type 1 diabetes. *Proceedings of the National Academy of Sciences of the United States of America*, 102(51):18425–30, December 2005. ISSN 0027-8424. doi:

10.1073/pnas.0508621102. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1317949&tool=pmcentrez&rendertype=abstract. 2

[17] E Allison Green, Leonid Gorelik, Catrin M McGregor, Elise H Tran, and Richard A Flavell. CD4+CD25+ T regulatory cells control anti-islet CD8+ T cells through TGF-beta-TGF-beta receptor interactions in type 1 diabetes. *Proceedings of the National Academy of Sciences of the United States of America*, 100(19):10878–83, September 2003. ISSN 0027-8424. doi: 10.1073/pnas.1834400100. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=196896&tool=pmcentrez&rendertype=abstract. 2

[18] Anna M Bulek, David K Cole, Ania Skowera, Garry Dolton, Stephanie Gras, Florian Madura, Anna Fuller, John J Miles, Emma Gostick, David A Price, Jan W Drijfhout, Robin R Knight, Guo C Huang, Nikolai Lissin, Peter E Molloy, Linda Wooldridge, Bent K Jakobsen, Jamie Rossjohn, Mark Peakman, Pierre J Rizkallah, and Andrew K Sewell. Structural basis for the killing of human beta cells by CD8(+) T cells in type 1 diabetes. *Nature immunology*, 13(3):283–9, March 2012. ISSN 1529-2916. doi: 10.1038/ni.2206. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3378510&tool=pmcentrez&rendertype=abstract. 2

[19] Patrick Concannon, Stephen S Rich, and Gerald T Nepom. Genetics of type 1A diabetes. *The New England journal of medicine*, 360(16):1646–54, April 2009. ISSN 1533-4406. doi: 10.1056/NEJMra0808284. URL http://www.ncbi.nlm.nih.gov/pubmed/19369670. 2, 4.4

[20] Jeffrey C Barrett, David G Clayton, Patrick Concannon, Beena Akolkar, Jason D Cooper, Henry A Erlich, Cécile Julier, Grant Morahan, Jø rn Nerup, Concepcion Nierras, Vincent Plagnol, Flemming Pociot, Helen Schuilenburg, Deborah J Smyth, Helen Stevens, John A Todd, Neil M Walker, and Stephen S Rich. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature genetics*, 41(6):703–7, June 2009. ISSN 1546-1718. doi: 10.1038/ng.381. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2889014&tool=pmcentrez&rendertype=abstract. 2, 4.4

[21] Marc S. Horwitz, Linda M. Bradley, Judith Harbertson, Troy Krahl, Jae Lee, and Nora Sarvennick. Diabetes induced by Coxsackie virus: Initiation by bystander damage and not molecular mimicry. *Nature Medicine*, 4(7):781–785, July 1998. ISSN 1078-8956. doi: 10.1038/nm0798-781. URL http://www.nature.com/doifinder/10.1038/nm0798-781. 2

[22] Li Wen, Ruth E Ley, Pavel Yu Volchkov, Peter B Stranges, Lia Avanesyan, Austin C Stonebraker, Changyun Hu, F Susan Wong, Gregory L Szot, Jeffrey A Bluestone, Jeffrey I Gordon, and Alexander V Chervonsky. Innate immunity and intestinal micro-

biota in the development of Type 1 diabetes. *Nature*, 455(7216):1109–13, October 2008. ISSN 1476-4687. doi: 10.1038/nature07336. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2574766&tool=pmcentrez&rendertype=abstract. 2

[23] J Karjalainen, J M Martin, M Knip, J Ilonen, B H Robinson, E Savilahti, H K Akerblom, and H M Dosch. A bovine albumin peptide as a possible trigger of insulin-dependent diabetes mellitus. *The New England journal of medicine*, 327(5):302–7, July 1992. ISSN 0028-4793. doi: 10.1056/NEJM199207303270502. URL http://www.ncbi.nlm.nih.gov/pubmed/1377788. 2

[24] Anette-G Ziegler, Sandra Schmid, Doris Huber, Michael Hummel, and Ezio Bonifacio. Early infant feeding and risk of developing type 1 diabetes-associated autoantibodies. *JAMA : the journal of the American Medical Association*, 290(13):1721–8, October 2003. ISSN 1538-3598. doi: 10.1001/jama.290.13.1721. URL http://www.ncbi.nlm.nih.gov/pubmed/14519706. 2

[25] E Hyppönen, E Läärä, A Reunanen, M R Järvelin, and S M Virtanen. Intake of vitamin D and risk of type 1 diabetes: a birth-cohort study. *Lancet*, 358(9292):1500–3, November 2001. ISSN 0140-6736. doi: 10.1016/S0140-6736(01)06580-1. URL http://www.ncbi.nlm.nih.gov/pubmed/11705562. 2

[26] M Füchtenbusch, W Rabl, B Grassl, W Bachmann, E Standl, and A G Ziegler. Delay of type I diabetes in high risk, first degree relatives by parenteral antigen administration: the Schwabing Insulin Prophylaxis Pilot Trial. *Diabetologia*, 41(5):536–41, May 1998. ISSN 0012-186X. doi: 10.1007/s001250050943. URL http://www.ncbi.nlm.nih.gov/pubmed/9628270. 2

[27] Tihamer Orban and Janos Tibor Kis. Prevention of Type 1 Diabetes Mellitus using a Novel Vaccine. *Therapeutic advances in endocrinology and metabolism*, 2 (1):9–16, February 2011. ISSN 2042-0196. doi: 10.1177/2042018810390546. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3474625&tool=pmcentrez&rendertype=abstract. 2

[28] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 3.1.2, 3.3.4

[29] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20 (3):273–297, September 1995. ISSN 0885-6125. doi: 10.1007/BF00994018. URL http://link.springer.com/10.1007/BF00994018. 3.1.3

[30] L Breiman, J H Friedman, R A Olshen, and C J Stone. *Classification and Regression Trees*, volume 19. 1984. ISBN 0412048418. 3.1.4

[31] J R Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986. ISSN 1062936X. doi: 10.1080/10629360600933723. 3.1.4

[32] Laurent Hyafil and Ronald L. Rivest. Constructing optimal binary decision trees is NP-complete. *Information Processing Letters*, 5(1):15–17, May 1976. ISSN 00200190. doi: 10.1016/0020-0190(76)90095-8. URL http://linkinghub.elsevier.com/retrieve/pii/0020019076900958. 3.1.4

[33] Leo Breiman. Random Forests. *Machine Learning*, 45:5–32, 2001. doi: 10.1023/A:1010933404324. 3.1.4, 3.2.1, 3.3.5

[34] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, June 2006. ISSN 01678655. doi: 10.1016/j.patrec.2005.10.010. URL http://linkinghub.elsevier.com/retrieve/pii/S016786550500303X. 3.2.1, 3.2.1

[35] John Platt and Others. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999. 3.2.1

[36] Ron Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *IJCAI*, 14, 1995. 3.2.2

[37] Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, December 1997. ISSN 00043702. doi: 10.1016/S0004-3702(97)00043-X. URL http://dx.doi.org/10.1016/S0004-3702(97)00043-X. 3.3

[38] Yi-wei Chen and Chih-jen Lin. Combining SVMs with Various Feature Selection Strategies. *Strategies*, 324:1–10, 2006. ISSN 14349922. doi: 10.1007/978-3-540-35488-8_13. 3.3.1

[39] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46(1-3):389–422, January 2002. ISSN 1573-0565. doi: 10.1023/A:1012487302797. URL http://link.springer.com/article/10.1023/A:1012487302797. 3.3.2, 3.3.2, 6

[40] O Frank, B Brors, A Fabarius, L Li, M Haak, S Merk, U Schwindel, C Zheng, M C Müller, N Gretz, R Hehlmann, A Hochhaus, and W Seifarth. Gene expression signature of primary imatinib-resistant chronic myeloid leukemia patients. *Leukemia*, 20(8):1400–7, August 2006. ISSN 0887-6924. doi: 10.1038/sj.leu.2404270. URL http://www.ncbi.nlm.nih.gov/pubmed/16728981. 3.3.2

[41] Kai-Bo Duan, Jagath C Rajapakse, Haiying Wang, and Francisco Azuaje. Multiple SVM-RFE for gene selection in cancer classification with expression data. *IEEE transactions on nanobioscience*, 4(3):228–34, September 2005. ISSN

1536-1241. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2789385&tool=pmcentrez&rendertype=abstract. 3.3.2, 3.3.3

[42] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal Of Statistical Software*, 33:1–22, 2010. ISSN 15487660. doi: 10.1359/JBMR.0301229. 3.3.4

[43] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, April 2005. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2005.00503.x. URL http://doi.wiley.com/10.1111/j.1467-9868.2005.00503.x. 3.3.4, 6

[44] Sijian Wang, Bin Nan, Saharon Rosset, and Ji Zhu. RANDOM LASSO. *The annals of applied statistics*, 5(1):468–485, March 2011. ISSN 1932-6157. doi: 10.1214/10-AOAS377. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3445423&tool=pmcentrez&rendertype=abstract. 3.3.4, 6

[45] Peter Carbonetto and Matthew Stephens. Scalable Variational Inference for Bayesian Variable Selection in Regression, and its Accuracy in Genetic Association Studies. *Bayesian Analysis*, 7(1):73–108, 2012. ISSN 1931-6690. doi: 10.1214/12-BA703. URL http://ba.stat.cmu.edu/journal/2012/vol07/issue01/carbonetto.pdf. 3.3.6, 3.3.6

[46] Peter J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995. ISSN 0006-3444. doi: 10.1093/biomet/82.4.711. URL http://biomet.oxfordjournals.org/cgi/doi/10.1093/biomet/82.4.711. 3.3.6

[47] Siddhartha Chib and Edward Greenberg. Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49(4):327–335, November 1995. ISSN 0003-1305. doi: 10.1080/00031305.1995.10476177. URL http://www.tandfonline.com/doi/abs/10.1080/00031305.1995.10476177. 3.3.6

[48] David J Lunn, John C Whittaker, and Nicky Best. A Bayesian toolkit for genetic association studies. *Genetic epidemiology*, 30(3):231–47, April 2006. ISSN 0741-0395. doi: 10.1002/gepi.20140. URL http://www.ncbi.nlm.nih.gov/pubmed/16544290. 3.3.6

[49] David J Nott and Peter J Green. Bayesian Variable Selection and the Swendsen-Wang Algorithm. *Journal of Computational and Graphical Statistics*, 13(1):141–157, March 2004. ISSN 1061-8600. doi: 10.1198/1061860042958. URL http://www.tandfonline.com/doi/abs/10.1198/1061860042958. 3.3.6

[50] Merlise a. Clyde, Joyee Ghosh, and Michael L. Littman. Bayesian Adaptive Sampling for Variable Selection and Model Averaging. *Journal of Computational and Graphical*

Statistics, 20(1):80–101, January 2011. ISSN 1061-8600. doi: 10.1198/jcgs.2010.09049. URL http://www.tandfonline.com/doi/abs/10.1198/jcgs.2010.09049. 3.3.6

[51] Jennifer A Hoeting, David Madigan, Adrian E Raftery, and Chris T Volinsky. Bayesian Model Averaging : A Tutorial. *Statistical Science*, 14(4):382–417, 1999. 3.3.6

[52] Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6 (2):461–464, March 1978. ISSN 0090-5364. doi: 10.1214/aos/1176344136. URL http://projecteuclid.org/euclid.aos/1176344136. 3.3.6

[53] David Madigan and Adrian E. Raftery. Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. *Journal of the American Statistical Association*, 89(428):1535–1546, December 1994. ISSN 0162-1459. doi: 10.1080/01621459.1994.10476894. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1994.10476894. 3.3.6

[54] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 233:183–233, 1999. ISSN 08856125. doi: 10.1023/A:1007665907178. 3.3.6

[55] P J Bingley, M R Christie, E Bonifacio, R Bonfanti, M Shattock, M T Fonte, G F Bottazzo, and E A Gale. Combined analysis of autoantibodies improves prediction of IDDM in islet cell antibody-positive relatives. *Diabetes*, 43(11):1304–10, November 1994. ISSN 0012-1797. URL http://www.ncbi.nlm.nih.gov/pubmed/7926304. 4.1, 5.1

[56] C F Verge, R Gianani, E Kawasaki, L Yu, M Pietropaolo, R A Jackson, H P Chase, and G S Eisenbarth. Prediction of type I diabetes in first-degree relatives using a combination of insulin, GAD, and ICA512bdc/IA-2 autoantibodies. *Diabetes*, 45(7):926–33, July 1996. ISSN 0012-1797. URL http://www.ncbi.nlm.nih.gov/pubmed/8666144. 4.1, 5.1

[57] Ana M Valdes, Michael D Varney, Henry a Erlich, and Janelle a Noble. Receiver Operating Characteristic Analysis of HLA, CTLA4, and Insulin Genotypes for Type 1 Diabetes. *Diabetes care*, pages 1–4, April 2013. ISSN 1935-5548. doi: 10.2337/dc12-2284. URL http://www.ncbi.nlm.nih.gov/pubmed/23628620. 4.1, 4.2.3, 4.4

[58] C Winkler, J Krumsiek, J Lempainen, P Achenbach, H Grallert, E Giannopoulou, M Bunk, F J Theis, E Bonifacio, and a G Ziegler. A strategy for combining minor genetic susceptibility genes to improve prediction of disease in type 1 diabetes. *Genes and immunity*, 13(7):549–55, October 2012. ISSN 1476-5470. doi: 10.1038/gene.2012.36. URL http://www.ncbi.nlm.nih.gov/pubmed/22932816. 4.1, 4.4

[59] David G Clayton. Prediction and interaction in complex disease genetics: experience in type 1 diabetes. *PLoS genetics*, 5(7):e1000540, July 2009. ISSN 1553-7404. doi: 10.1371/journal.pgen.1000540. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2703795&tool=pmcentrez&rendertype=abstract. 4.1

[60] Stephen S Rich, Patrick Concannon, Henry Erlich, Cecile Julier, Grant Morahan, Jorn Nerup, Flemming Pociot, and John A Todd. The Type 1 Diabetes Genetics Consortium. *Annals of the New York Academy of Sciences*, 1079:1–8, October 2006. ISSN 0077-8923. doi: 10.1196/annals.1375.001. URL http://www.ncbi.nlm.nih.gov/pubmed/17130525. 4.2.1

[61] a G Ziegler, M Hummel, M Schenker, and E Bonifacio. Autoantibody appearance and risk for development of childhood diabetes in offspring of parents with type 1 diabetes: the 2-year analysis of the German BABYDIAB Study. *Diabetes*, 48(3):460–8, March 1999. ISSN 0012-1797. URL http://www.ncbi.nlm.nih.gov/pubmed/10078544. 4.2.2, 5.2

[62] Leonore Thümer, Kerstin Adler, Ezio Bonifacio, Frank Hofmann, Manfred Keller, Christine Milz, Axel Munte, and Anette-Gabriele Ziegler. German new onset diabetes in the young incident cohort study: DiMelli study design and first-year results. *The review of diabetic studies : RDS*, 7(3):202–8, January 2010. ISSN 1614-0575. doi: 10.1900/RDS.2010.7.202. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3061610&tool=pmcentrez&rendertype=abstract. 4.2.2

[63] Henry Erlich, Ana Maria Valdes, Janelle Noble, Joyce A Carlson, Mike Varney, Pat Concannon, Josyf C Mychaleckyj, John A Todd, Persia Bonella, Anna Lisa Fear, Eva Lavant, Anthony Louey, Priscilla Moonsamy, Type Diabetes, and Genetics Consortium. Analysis of the Type 1 Diabetes Genetics Consortium Families. 57(April), 2008. doi: 10.2337/db07-1331.Additional. 4.2.3

[64] D. R. Cox. Interaction. *International Statistical Review / Revue Internationale de Statistique*, 52(1):1, April 1984. ISSN 03067734. doi: 10.2307/1403235. URL http://www.jstor.org/stable/1403235?origin=crossref. 4.3.4

[65] Amy Berrington de González and D. R. Cox. Interpretation of interaction: A review. *The Annals of Applied Statistics*, 1(2):371–385, December 2007. ISSN 1932-6157. doi: 10.1214/07-AOAS124. URL http://projecteuclid.org/euclid.aoas/1196438023. 4.3.4

[66] Lu Qi, Marilyn C Cornelis, Cuilin Zhang, Rob M van Dam, and Frank B Hu. Genetic predisposition, Western dietary pattern, and the risk of type 2 diabetes in men. *The American journal of clinical nutrition*, 89(5):1453–8, May 2009. ISSN 1938-3207. doi: 10.3945/ajcn.2008.27249. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2676999&tool=pmcentrez&rendertype=abstract. 4.3.4

[67] Leonid Padyukov, Camilla Silva, Patrik Stolt, Lars Alfredsson, and Lars Klareskog. A gene-environment interaction between smoking and shared epitope genes in HLA-DR provides a high risk of seropositive rheumatoid arthritis. *Arthritis and rheumatism*, 50(10):3085–92, October 2004. ISSN 0004-3591. doi: 10.1002/art.20553. URL http://www.ncbi.nlm.nih.gov/pubmed/15476204. 4.3.4

[68] Ben Carterette. On rank correlation and the distance between rankings. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '09*, page 436, 2009. doi: 10.1145/1571941.1572017. URL http://portal.acm.org/citation.cfm?doid=1571941.1572017. 4.3.5

[69] Hironori Ueda, Joanna M M Howson, Laura Esposito, Joanne Heward, Hywel Snook, Giselle Chamberlain, Daniel B Rainbow, Kara M D Hunter, Annabel N Smith, Gianfranco Di Genova, Mathias H Herr, Ingrid Dahlman, Felicity Payne, Deborah Smyth, Christopher Lowe, Rebecca C J Twells, Sarah Howlett, Barry Healy, Sarah Nutland, Helen E Rance, Vin Everett, Luc J Smink, Alex C Lam, Heather J Cordell, Neil M Walker, Cristina Bordin, John Hulme, Costantino Motzo, Francesco Cucca, J Fred Hess, Michael L Metzker, Jane Rogers, Simon Gregory, Amit Allahabadia, Ratnasingam Nithiyananthan, Eva Tuomilehto-Wolf, Jaakko Tuomilehto, Polly Bingley, Kathleen M Gillespie, Dag E Undlien, Kjersti S Rø nningen, Cristian Guja, Constantin Ionescu-Tîrgovişte, David A Savage, A Peter Maxwell, Dennis J Carson, Chris C Patterson, Jayne A Franklyn, David G Clayton, Laurence B Peterson, Linda S Wicker, John A Todd, and Stephen C L Gough. Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. *Nature*, 423(6939):506–11, May 2003. ISSN 0028-0836. doi: 10.1038/nature01621. URL http://www.ncbi.nlm.nih.gov/pubmed/12724780. 4.4

[70] Li-Tzu Yeh, Shi-Chuen Miaw, Ming-Hong Lin, Feng-Cheng Chou, Shing-Jia Shieh, Yi-Ping Chuang, Shih-Hua Lin, Deh-Ming Chang, and Huey-Kang Sytwu. Different modulation of ptpn22 in effector and regulatory T cells leads to attenuation of autoimmune diabetes in transgenic nonobese diabetic mice. *Journal of immunology (Baltimore, Md. : 1950)*, 191(2):594–607, July 2013. ISSN 1550-6606. doi: 10.4049/jimmunol.1203380. URL http://www.ncbi.nlm.nih.gov/pubmed/23752610. 4.4

[71] Xuezhi Dai, Richard G James, Tania Habib, Swati Singh, Shaun Jackson, Socheath Khim, Randall T Moon, Denny Liggitt, Alejandro Wolf-Yadlin, Jane H Buckner, and David J Rawlings. A disease-associated PTPN22 variant promotes systemic autoimmunity in murine models. *The Journal of clinical investigation*, 123(5):2024–36, May 2013. ISSN 1558-8238. doi: 10.1172/JCI66963. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3638909&tool=pmcentrez&rendertype=abstract. 4.4

[72] Norio Kanatsuna, Jalal Taneera, Fariba Vaziri-Sani, Nils Wierup, Helena Elding Larsson, Ahmed Delli, Hanna Skärstrand, Alexander Balhuizen, Hedvig Bennet, Donald F Steiner, Carina Törn, Malin Fex, and Ake Lernmark. Autoimmunity against INS-IGF2 expressed in human pancreatic islets. *The Journal of biological chemistry*, August 2013. ISSN 1083-351X. doi: 10.1074/jbc.M113.478222. URL http://www.ncbi.nlm.nih.gov/pubmed/23935095. 4.4

[73] Simone Renner, Christina Braun-Reichhart, Andreas Blutke, Nadja Herbach, Daniela Emrich, Elisabeth Streckel, Annegret Wünsch, Barbara Kessler, Mayuko Kurome, Andrea Bähr, Nikolai Klymiuk, Stefan Krebs, Oliver Puk, Hiroshi Nagashima, Jochen Graw, Helmut Blum, Ruediger Wanke, and Eckhard Wolf. Permanent neonatal diabetes in INS(C94Y) transgenic pigs. *Diabetes*, 62(5):1505–11, May 2013. ISSN 1939-327X. doi: 10.2337/db12-1065. URL http://www.ncbi.nlm.nih.gov/pubmed/23274907. 4.4

[74] Hongjie Wang, Yulan Jin, M V Prasad Linga Reddy, Robert Podolsky, Siyang Liu, Ping Yang, Bruce Bode, John Chip Reed, R Dennis Steed, Stephen W Anderson, Leigh Steed, Diane Hopkins, Yihua Huang, and Jin-Xiong She. Genetically dependent ERBB3 expression modulates antigen presenting cell function and type 1 diabetes risk. *PloS one*, 5(7):e11789, January 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0011789. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2909911&tool=pmcentrez&rendertype=abstract. 4.4

[75] Eleftherios Vairaktaris, Lambros Goutzanis, Stavros Vassiliou, Sofia Spyridonidou, Emeka Nkenke, Georgios Papageorgiou, Pashalis Strantzias, Andreas Lazaris, Christos Yapijakis, and Efstratios Patsouris. Enhancement of erbB2 and erbB3 expression during oral oncogenesis in diabetic rats. *Journal of cancer research and clinical oncology*, 134(3):337–44, March 2008. ISSN 1432-1335. doi: 10.1007/s00432-007-0289-7. URL http://www.ncbi.nlm.nih.gov/pubmed/17704947. 4.4

[76] Brittni Frederiksen, Miranda Kroehl, Molly M Lamb, Jennifer Seifert, Katherine Barriga, George S Eisenbarth, Marian Rewers, and Jill M Norris. Infant Exposures and Development of Type 1 Diabetes Mellitus: The Diabetes Autoimmunity Study in the Young (DAISY). *JAMA pediatrics*, July 2013. ISSN 2168-6211. doi: 10.1001/jamapediatrics.2013.317. URL http://www.ncbi.nlm.nih.gov/pubmed/23836309. 4.4, 5.4

[77] Andrea K Steck, Kelly Johnson, Katherine J Barriga, Dongmei Miao, Liping Yu, John C Hutton, George S Eisenbarth, and Marian J Rewers. Age of islet autoantibody appearance and mean levels of insulin, but not GAD or IA-2 autoantibodies, predict age of diagnosis of type 1 diabetes: diabetes autoimmunity study in the young. *Diabetes care*, 34(6):1397–9, June 2011. ISSN 1935-5548. doi: 10.2337/

dc10-2088. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3114355&tool=pmcentrez&rendertype=abstract. 5.1

[78] Christine von Toerne, Melanie Kahle, Alexander Schäfer, Ruben Ispiryan, Marcel Blindert, Martin Hrabe De Angelis, Susanne Neschen, Marius Ueffing, and Stefanie M Hauck. Apoe, Mbl2, and Psp plasma protein levels correlate with diabetic phenotype in NZO mice–an optimized rapid workflow for SRM-based quantification. *Journal of proteome research*, 12 (3):1331–43, March 2013. ISSN 1535-3907. doi: 10.1021/pr3009836. URL http://www.ncbi.nlm.nih.gov/pubmed/23350727. 5.1

[79] G Puavilai, S Chanprasertyotin, and A Sriphrapradaeng. Diagnostic criteria for diabetes mellitus and other categories of glucose intolerance: 1997 criteria by the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus (ADA), 1998 WHO consultation criteria, and 1985 WHO criteria. World Healt. *Diabetes research and clinical practice*, 44 (1):21–6, April 1999. ISSN 0168-8227. URL http://www.ncbi.nlm.nih.gov/pubmed/10414936. 5.2

[80] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422 (6928):198–207, March 2003. ISSN 0028-0836. doi: 10.1038/nature01511. URL http://www.ncbi.nlm.nih.gov/pubmed/12634793. 5.2

[81] Markus Brosch, Lu Yu, Tim Hubbard, and Jyoti Choudhary. Accurate and sensitive peptide identification with Mascot Percolator. *Journal of proteome research*, 8(6):3176–81, June 2009. ISSN 1535-3893. doi: 10.1021/pr800982s. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2734080&tool=pmcentrez&rendertype=abstract. 5.2

[82] Paul Flicek, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, Laurent Gil, Leo Gordon, Maurice Hendrix, Thibaut Hourlier, Nathan Johnson, Andreas K Kähäri, Damian Keefe, Stephen Keenan, Rhoda Kinsella, Monika Komorowska, Gautier Koscielny, Eugene Kulesha, Pontus Larsson, Ian Longden, William McLaren, Matthieu Muffato, Bert Overduin, Miguel Pignatelli, Bethan Pritchard, Harpreet Singh Riat, Graham R S Ritchie, Magali Ruffier, Michael Schuster, Daniel Sobral, Y Amy Tang, Kieron Taylor, Stephen Trevanion, Jana Vandrovcova, Simon White, Mark Wilson, Steven P Wilder, Bronwen L Aken, Ewan Birney, Fiona Cunningham, Ian Dunham, Richard Durbin, Xosé M Fernández-Suarez, Jennifer Harrow, Javier Herrero, Tim J P Hubbard, Anne Parker, Glenn Proctor, Giulietta Spudich, Jan Vogel, Andy Yates, Amonida Zadissa, and Stephen M J Searle. Ensembl 2012. *Nucleic acids research*, 40(Database issue):D84–90, January 2012. ISSN 1362-4962. doi: 10.1093/nar/gkr991. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245178&tool=pmcentrez&rendertype=abstract. 5.2

[83] Arshag D Mooradian, Michael J Haas, and N. C.W. Wong. Transcriptional Control of Apolipoprotein A-I Gene Expression in Diabetes. *Diabetes*, 53(3):513–520, March 2004. ISSN 0012-1797. doi: 10.2337/diabetes.53.3.513. URL http://diabetes.diabetesjournals.org/cgi/doi/10.2337/diabetes.53.3.513. 5.3.3.1

[84] Arshag D Mooradian. Cardiovascular disease in type 2 diabetes mellitus: current management guidelines. *Archives of internal medicine*, 163(1):33–40, January 2003. ISSN 0003-9926. URL http://www.ncbi.nlm.nih.gov/pubmed/12523914. 5.3.3.1

[85] J C Pickup and M A Crook. Is type II diabetes mellitus a disease of the innate immune system? *Diabetologia*, 41(10):1241–8, October 1998. ISSN 0012-186X. doi: 10.1007/s001250051058. URL http://www.ncbi.nlm.nih.gov/pubmed/9794114. 5.3.3.1

[86] Yoshitaka Kumon, Tadashi Suehiro, Takashi Itahara, Yukio Ikeda, and Kozo Hashimoto. Serum amyloid A protein in patients with non-insulin-dependent diabetes mellitus. *Clinical Biochemistry*, 27(6):469–473, December 1994. ISSN 00099120. doi: 10.1016/0009-9120(94)00044-V. URL http://linkinghub.elsevier.com/retrieve/pii/000991209400044V. 5.3.3.1

[87] S Roy, R Sala, E Cagliero, and M Lorenzi. Overexpression of fibronectin induced by diabetes or high glucose: phenomenon with a memory. *Proceedings of the National Academy of Sciences*, 87(1):404–408, January 1990. URL http://www.pnas.org/content/87/1/404.abstract. 5.3.3.1

[88] S. D.J.M. Kanters, J.-D. Banga, A. Algra, R. C.J.M. Frijns, J. J. Beutler, and R. Fijnheer. Plasma Levels of Cellular Fibronectin in Diabetes. *Diabetes Care*, 24(2):323–327, February 2001. ISSN 0149-5992. doi: 10.2337/diacare.24.2.323. URL http://care.diabetesjournals.org/cgi/doi/10.2337/diacare.24.2.323. 5.3.3.1

[89] Teizo Fujita. Evolution of the lectin-complement pathway and its role in innate immunity. *Nature reviews. Immunology*, 2(5):346–53, May 2002. ISSN 1474-1733. doi: 10.1038/nri800. URL http://www.ncbi.nlm.nih.gov/pubmed/12033740. 5.3.3.2, 5.3.3.2

[90] Jakob Ostergaard, Troels Krarup Hansen, Steffen Thiel, and Allan Flyvbjerg. Complement activation and diabetic vascular complications. *Clinica chimica acta; international journal of clinical chemistry*, 361(1-2):10–9, November 2005. ISSN 0009-8981. doi: 10.1016/j.cccn.2005.04.028. URL http://dx.doi.org/10.1016/j.cccn.2005.04.028. 5.3.3.2

[91] Lee H Bouwman, Bart O Roep, and Anja Roos. Mannose-binding lectin: clinical implications for infection, transplantation, and autoimmunity. *Human immunology*, 67(4-5):247–56. ISSN 0198-8859. doi: 10.1016/j.humimm.2006.02.030. URL http://dx.doi.org/10.1016/j.humimm.2006.02.030http://www.ncbi.nlm.nih.gov/pubmed/16720204. 5.3.3.2

[92] T. K. Hansen. Elevated Levels of Mannan-Binding Lectin in Patients with Type 1 Diabetes. *Journal of Clinical Endocrinology & Metabolism*, 88(10):4857–4861, October 2003. ISSN 0021-972X. doi: 10.1210/jc.2003-030742. URL http://jcem.endojournals.org/content/88/10/4857.short. 5.3.3.2

[93] L. H. Bouwman, P. Eerligh, O. T. Terpstra, M. R. Daha, P. de Knijff, B. E.P.B. Ballieux, G. J. Bruining, A. R. van der Slik, A. Roos, and B. O. Roep. Elevated Levels of Mannose-Binding Lectin at Clinical Manifestation of Type 1 Diabetes in Juveniles. *Diabetes*, 54 (10):3002–3006, October 2005. ISSN 0012-1797. doi: 10.2337/diabetes.54.10.3002. URL http://diabetes.diabetesjournals.org/content/54/10/3002.short. 5.3.3.2

[94] M Saraheimo, C Forsblom, T K Hansen, A-M Teppo, J Fagerudd, K Pettersson-Fernholm, S Thiel, L Tarnow, P Ebeling, A Flyvbjerg, and P-H Groop. Increased levels of mannan-binding lectin in type 1 diabetic patients with incipient and overt nephropathy. *Diabetologia*, 48(1):198–202, January 2005. ISSN 0012-186X. doi: 10.1007/s00125-004-1594-1. URL http://www.ncbi.nlm.nih.gov/pubmed/15616805. 5.3.3.2

[95] T K Hansen, C Forsblom, M Saraheimo, L Thorn, J Wadén, P Hø yem, J Ø stergaard, A Flyvbjerg, and P-H Groop. Association between mannose-binding lectin, high-sensitivity C-reactive protein and the progression of diabetic nephropathy in type 1 diabetes. *Diabetologia*, 53(7):1517–24, July 2010. ISSN 1432-0428. doi: 10.1007/s00125-010-1742-8. URL http://www.ncbi.nlm.nih.gov/pubmed/20396997. 5.3.3.2

[96] P. Hovind, T. K. Hansen, L. Tarnow, S. Thiel, R. Steffensen, A. Flyvbjerg, and H.-H. Parving. Mannose-Binding Lectin as a Predictor of Microalbuminuria in Type 1 Diabetes: An Inception Cohort Study. *Diabetes*, 54(5):1523–1527, May 2005. ISSN 0012-1797. doi: 10.2337/diabetes.54.5.1523. URL http://diabetes.diabetesjournals.org/content/54/5/1523.short. 5.3.3.2

[97] Jinhui Zhang, Wendy Wright, David A Bernlohr, Samuel W Cushman, and Xiaoli Chen. Alterations of the classic pathway of complement in adipose tissue of obesity and insulin resistance. *American journal of physiology. Endocrinology and metabolism*, 292(5):E1433–40, May 2007. ISSN 0193-1849. doi: 10.1152/ajpendo.00664.2006. URL http://www.ncbi.nlm.nih.gov/pubmed/17244723. 5.3.3.2

[98] Santiago Rodríguez de Córdoba, Jorge Esparza-Gordillo, Elena Goicoechea de Jorge, Margarita Lopez-Trascasa, and Pilar Sánchez-Corral. The human complement factor H: functional roles, genetic variations and disease associations. *Molecular immunology*, 41(4):355–67, June 2004. ISSN 0161-5890. doi: 10.1016/j.molimm.2004.02.005. URL http://www.ncbi.nlm.nih.gov/pubmed/15163532. 5.3.3.2

[99] Christine Skerka, Qian Chen, Veronique Fremeaux-Bacchi, and Lubka T Roumenina. Complement factor H related proteins (CFHRs). *Molecular immunology*, 56(3):170–80, December 2013. ISSN 1872-9142. doi: 10.1016/j.molimm.2013.06.001. URL http://www.ncbi.nlm.nih.gov/pubmed/23830046. 5.3.3.2

[100] Christine Skerka and Peter F Zipfel. Complement factor H related proteins in immune diseases. *Vaccine*, 26 Suppl 8:I9–14, December 2008. ISSN 0264-410X. URL http://www.ncbi.nlm.nih.gov/pubmed/19388158. 5.3.3.2

[101] Peter F Zipfel, Teresia Hallström, Sven Hammerschmidt, and Christine Skerka. The complement fitness factor H: role in human diseases and for immune escape of pathogens, like pneumococci. *Vaccine*, 26 Suppl 8:I67–74, December 2008. ISSN 0264-410X. URL http://www.ncbi.nlm.nih.gov/pubmed/19388168. 5.3.3.2

[102] Mihály Józsi and Peter F Zipfel. Factor H family proteins and human diseases. *Trends in immunology*, 29(8):380–7, August 2008. ISSN 1471-4906. doi: 10.1016/j.it.2008.04.008. URL http://www.ncbi.nlm.nih.gov/pubmed/18602340. 5.3.3.2

[103] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–9, May 2000. ISSN 1061-4036. doi: 10.1038/75556. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3037419&tool=pmcentrez&rendertype=abstract. 6

[104] Minoru Kanehisa, Susumu Goto, Masahiro Hattori, Kiyoko F Aoki-Kinoshita, Masumi Itoh, Shuichi Kawashima, Toshiaki Katayama, Michihiro Araki, and Mika Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic acids research*, 34(Database issue):D354–7, January 2006. ISSN 1362-4962. doi: 10.1093/nar/gkj102. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1347464&tool=pmcentrez&rendertype=abstract. 6

[105] M Xiong, W Li, J Zhao, L Jin, and E Boerwinkle. Feature (gene) selection in gene expression-based tumor classification. *Molecular genetics and metabolism*, 73(3):239–47, July 2001. ISSN 1096-7192. doi: 10.1006/mgme.2001.3193. URL http://www.ncbi.nlm.nih.gov/pubmed/11461191. 6

[106] Christophe Ambroise and Geoffrey J McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences of the United States of America*, 99(10):6562–6, May 2002. ISSN 0027-8424. doi: 10.1073/pnas.102102699. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=124442&tool=pmcentrez&rendertype=abstract. 6

[107] Thomas Abeel, Thibault Helleputte, Yves Van de Peer, Pierre Dupont, and Yvan Saeys. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics (Oxford, England)*, 26(3):392–8, February 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp630. URL http://www.ncbi.nlm.nih.gov/pubmed/19942583. 6

[108] Jason Weston, Andre Elisseeff, Bernhard Scholkopf, and Mike Tipping. Use of the Zero-Norm with Linear Models and Kernel Methods. *Journal of Machine Learning Research*, 3:1439–1461, 2003. ISSN 15324435. doi: 10.1162/153244303322753751. 6

[109] Natalia Becker, Wiebke Werft, Grischa Toedt, Peter Lichter, and Axel Benner. penalizedSVM: a R-package for feature selection SVM classification. *Bioinformatics (Oxford, England)*, 25(13):1711–2, July 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp286. URL http://bioinformatics.oxfordjournals.org/content/25/13/1711.long. 6

[110] Sebastián Maldonado, Richard Weber, and Jayanta Basak. Simultaneous feature selection and classification using kernel-penalized support vector machines. *Information Sciences*, 181(1):115–128, January 2011. ISSN 00200255. doi: 10.1016/j.ins.2010.08.047. URL http://linkinghub.elsevier.com/retrieve/pii/S0020025510004287. 6

[111] Isabelle Guyon. Causal feature selection. pages 1–40, 2007. 6

[112] Constantin F Aliferis. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I : Algorithms and Empirical Evaluation. *Journal of Machine Learning Research*, 11:171–234, 2010. ISSN 15324435. 6

[113] Alexander Statnikov, Laura E Brown, and Constantin F Aliferis. Generating Realistic Large Bayesian Networks by Tiling. In *FLAIRS Conference*, 2006. 6

[114] Alexander Statnikov, Vanderbilt Edu, and Constantin F Aliferis. TIED : An Artificially Simulated Dataset with Multiple Markov Boundaries. *Journal of Machine Learning Research*, pages 249–256, 2009. 6