



**POLITECNICO**  
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE  
E DELL'INFORMAZIONE

# Causal proteomic predictors of diabetes in South-Asia: a methodological and applied study using Stability Selection and Mendelian Randomization

TESI DI LAUREA MAGISTRALE IN  
MATHEMATICAL ENGINEERING - INGEGNERIA MATEMATICA

Author: **Paolo Triflio**

Student ID: 244886

Advisor: Prof. Francesca Ieva

Co-advisors: Solene Cadiou, Nicole Fontana

Academic Year: 2025-26



# Abstract

Type 2 diabetes (T2D) is a complex, multifactorial disease that affects hundreds of millions of people worldwide. Despite extensive research, identifying the biological factors that causally drive its onset and progression remains a key challenge for improving prevention, diagnosis, and treatment. This thesis addresses two interconnected research questions: a methodological one, evaluating the performance of Stability Selection (SS) in identifying potential causal predictors among blood proteins, and a clinical one, aimed at uncovering specific proteins that may play a causal role in type 2 diabetes development. To this aim, we used data coming from the BELIEVE study, which included genetic data, health data and 7244 blood protein measurements for around 10000 individuals. Methodologically, SS demonstrated strong robustness and consistency as a feature selection approach: across diverse input configurations, it consistently identified a stable set of 18 unique proteins. Compared with classical LASSO regression, SS provided greater model stability while maintaining competitive predictive power, confirming its suitability for high-dimensional omics data. Moreover, a comparison with the literature suggested that 8 out of the 18 predictors might have a causal role in the development of T2D. To assess causality, SS-identified proteins were further examined using Mendelian Randomization (MR) in both one sample and two sample designs. However, due to the small sample size, MR did not provide usable results. The effectiveness of SS for causal inference therefore needs to be confirmed through MR in larger studies, although the existing literature provides a promising starting point. Overall, this thesis highlights the complementary strengths of SS and MR and supports their combined use.

**Keywords:** Stability Selection, causal inference, Mendelian Randomization, type 2 diabetes, proteins



## Abstract in lingua italiana

Il diabete di tipo 2 (T2D) è una malattia complessa e multifattoriale che colpisce centinaia di milioni di persone in tutto il mondo. Nonostante le numerose ricerche, l'identificazione dei fattori biologici che determinano in modo causale l'insorgenza e la progressione della malattia rimane una sfida fondamentale per migliorare la prevenzione, la diagnosi e il trattamento. Questa tesi affronta due domande di ricerca tra loro interconnesse: una di natura metodologica, volta a valutare le prestazioni della Stability Selection (SS) nell'identificazione di potenziali predittori causali tra le proteine del sangue, e una di natura clinica, mirata a individuare specifiche proteine che possano avere un ruolo causale nello sviluppo del diabete di tipo 2. A tale scopo, sono stati utilizzati i dati dello studio BELIEVE, che includeva informazioni genetiche, dati clinici e 7244 misurazioni di proteine ematiche relative a circa 10000 individui. Dal punto di vista metodologico, la SS si è dimostrata un metodo di selezione delle variabili robusto e coerente: sottoposta a input diversi, ha stabilmente identificato un insieme di 18 proteine uniche. Rispetto alla regressione LASSO classica, la SS ha dunque fornito un modello più stabile mantenendo un potere predittivo competitivo, confermandone così l'idoneità per l'analisi di dati omici ad alta dimensionalità. Inoltre, il confronto con la letteratura ha suggerito che 8 delle 18 proteine individuate potrebbero avere un ruolo causale nello sviluppo del T2D. Per valutarne la causalità, le proteine identificate da SS sono state successivamente analizzate tramite Mendelian Randomization (MR), sia con un approccio one sample che two sample. Tuttavia, a causa della bassa numerosità campionaria, le analisi MR non hanno prodotto risultati utilizzabili. L'efficacia della SS per l'inferenza causale deve quindi essere confermata attraverso analisi MR condotte su campioni più ampi, sebbene la letteratura esistente fornisca un punto di partenza promettente. Nel complesso, questa tesi mette in evidenza le forze complementari di SS e MR e ne sostiene l'uso combinato.

**Parole chiave:** Stability Selection, inferenza causale, Mendelian Randomization, diabete di tipo 2, proteine



# Contents

<b>Abstract</b>	<b>i</b>
<b>Abstract in lingua italiana</b>	<b>iii</b>
<b>Contents</b>	<b>v</b>
<b>Introduction</b>	<b>1</b>
<b>1 Methods</b>	<b>5</b>
1.1 The Stability Selection algorithm . . . . .	5
1.1.1 From LASSO to Stability Selection . . . . .	5
1.1.2 The idea behind Stability Selection . . . . .	6
1.1.3 The stability score . . . . .	7
1.1.4 Predictive power comparison . . . . .	9
1.2 Causal inference . . . . .	10
1.2.1 Causal inference in a nutshell . . . . .	10
1.2.2 On the relationship between stability and causality . . . . .	12
1.2.3 Causal inference with observational data . . . . .	12
1.3 Mendelian Randomization (MR) . . . . .	14
1.3.1 MR's assumptions . . . . .	16
1.3.2 Estimating causal effects using MR . . . . .	17
<b>2 Data engineering</b>	<b>21</b>
2.1 Stability Selection: pre-processing . . . . .	22
2.2 Stability Selection: dimensional reduction and confounders adjustment . .	27
2.2.1 Dimensionality reduction on proteomics data . . . . .	27
2.2.2 Adjusting for confounders . . . . .	28
2.3 Mendelian Randomization: pre-processing . . . . .	29
2.3.1 Outcome data creation: regenie regression . . . . .	29

2.3.2	Exposure data: instrument selection . . . . .	31
<b>3</b>	<b>Results</b>	<b>33</b>
3.1	Preliminary LASSO Analysis . . . . .	33
3.2	Stability Selection . . . . .	36
3.2.1	Consistency analysis . . . . .	36
3.2.2	Reliability analysis . . . . .	46
3.3	Mendelian Randomization . . . . .	51
3.3.1	MR on the SS proteins . . . . .	52
3.3.2	MR on all the available proteins . . . . .	55
<b>4</b>	<b>Discussion and conclusion</b>	<b>61</b>
<b>A</b>	<b>Appendix A</b>	<b>65</b>
A.1	Prevalent diabetes data engineering . . . . .	65
A.2	Stability Selection on prevalent diabetes . . . . .	68
	<b>Bibliography</b>	<b>73</b>
	<b>List of Figures</b>	<b>83</b>
	<b>List of Tables</b>	<b>85</b>
	<b>Ringraziamenti</b>	<b>87</b>

# Introduction

Type 2 diabetes is a chronic metabolic disorder characterized by high levels of blood glucose (sugar) resulting from the body's ineffective use of insulin. Unlike type 1 diabetes, which is caused by the immune system attacking insulin-producing cells in the pancreas, type 2 diabetes is often associated with insulin resistance and a gradual decline in insulin production. It typically develops in adults over the age of 45, but it is increasingly being diagnosed in younger individuals due to rising rates of obesity, sedentary lifestyles, and poor dietary habits [1]. If left unmanaged, type 2 diabetes can lead to serious health complications, including heart disease, kidney failure, nerve damage, and vision problems [2]. Early detection and proper management through lifestyle changes, medication, and regular monitoring are essential to prevent or delay the progression of the disease.

Globally, type 2 diabetes has become a critical public health issue. According to the International Diabetes Federation [3], around 589 million adults aged 20-79 (approximately one in nine adults) are living with diabetes worldwide, and many of these cases are type 2 diabetes. Moreover, the International Diabetes Federation has estimated that by 2050 the number of people with diabetes will rise significantly, reaching over 850 million.

The prevalence is rising especially fast in low- and middle-income countries [4]. In particular, in Bangladesh the trend is significantly concerning: the country is currently ranked eighth worldwide in terms of diabetes prevalence [5]. Recent estimates indicate a 13.2% prevalence of diabetes among adults (corresponding to about 13.9 million people [6]), though a large number of cases also go undiagnosed, so the real burden may be higher [7]. Moreover, more than 14% of people with diabetes in Bangladesh face catastrophic health expenditure [8].

In this context, it is critically important not only to identify early predictors of type 2 diabetes, but also to uncover causal factors that could serve as therapeutic targets to prevent disease onset. Circulating blood proteins are especially attractive candidates: they reflect integrative physiology across tissues, are accessible and measurable in plasma, and many already serve as drug targets [9]. Recent large-scale proteomic and proteogenomic studies have already moved in this direction, identifying a significant number of plasma proteins with evidence of a causal relationship to type 2 diabetes [10, 11].

To find causal predictors, classical statistical inference is not enough, because its methods only look at associations between variables. On the other hand, statistical causal inference precisely aims at understanding causal links between variables, i.e. establishing whether or not some variables of interest (in our case, blood proteins) have a direct causal effect on some outcomes, in this case type 2 diabetes.

This thesis has a methodological aim and a clinical aim. The methodological aim is to evaluate the Stability Selection (SS) algorithm [12, 13] as a tool for identifying potential causal predictors. Indeed, although Stability Selection is not a formal causal inference method, it is still expected to find causal predictors, outperforming classical regression methods like LASSO. Indeed, while in standard LASSO a variable is selected according to its predictive power, in Stability Selection the selection procedure is the result of many underlying selection procedures on subsamples, such that only the variables that have been selected above a certain proportion are considered stable. Therefore, since stable predictors are repeatedly selected across various model configurations, they may more likely represent underlying causal relationships rather than random noise or simple associations. To assess Stability Selection ability to select causal predictors, the proteins identified through Stability Selection will subsequently be studied via a literature review, refining their plausibility as causal candidates. We will also compare them against results from Mendelian Randomization (MR) [14], a well-known causal inference method that uses genetic data as proxies for the exposure: the method checks for the effect of the genetic instrument on both outcome and exposure, and then leverage these effects to compute the causal link between exposure (the blood proteins) and outcome (type 2 diabetes).

The second, clinical aim, of the thesis is to evaluate causal effects of all 7244 available plasma proteins on type 2 diabetes. Using the results of the Stability Selection study as well as the ones from Mendelian Randomization and the literature review, we want to point out potential proteomic causal predictor of diabetes onset. The data used in this thesis come from the BELIEVE cohort study [15], a South-Asian cohort in which we have genetic and clinical data and extensive proteomics measurements for around 10000 individuals.

The structure of the thesis is therefore as follows: Chapter 1 introduces the theoretical background on the main statistical methods that will be implemented in the thesis (LASSO, Stability Selection, causal inference and Mendelian Randomization). Chapter 2 presents the BELIEVE study and the data sources, continues with the pre-processing procedure and all the necessary steps to prepare the data for analyses. Chapter 3 states all the results obtained by running the models, detailed with explanation and model-to-model comparison. This includes several variations of the Stability Selection model to

assess its consistency, a thorough comparison with LASSO also in terms of predictive power, and the Mendelian Randomization analysis of the set of stable predictors found by Stability Selection. Moreover, MR will be applied also to all the available proteins. A literature review will be also performed for every possible causal predictor selected by any of the 2 selection methods applied.

Finally, Chapter 4 contains the overall conclusions of our work, as long as some possible future developments.



# 1 | Methods

In this chapter, the two main methodologies implemented in this work are presented. Section 1.1 introduces Stability Selection, beginning with its foundation, the LASSO method, and proceeding through its most recent extensions. Section 1.3 discusses Mendelian Randomization, which is preceded by an overview of the causal inference framework (Section 1.2).

## 1.1. The Stability Selection algorithm

Stability selection (SS) is a variable selection method first introduced by Meinshausen and Bühlmann [12] in 2010. It can be seen as an extension of any variable selection technique that depends on a regularization parameter  $\lambda$ , such as a LASSO regression model. However, rather than finding the most predictive variables, it focuses on the search of a stable set of predictors.

### 1.1.1. From LASSO to Stability Selection

Before presenting Stability Selection, we will present the LASSO framework and its limitations. Let  $Y$  be the response vector of length  $n$ ,  $\beta$  the vector of  $p$  coefficients and  $X$  the matrix of the predictors. The LASSO regression model [16] aims at finding the best  $\beta$  as

$$\hat{\beta} = \arg \min_{\beta} \sum_{i=1}^n (y_i - x_i^T \beta)^2 + \lambda \sum_{k=1}^p |\beta_k| \quad (1.1)$$

where  $\lambda$  is a penalization parameter that can be tuned through cross-validation. The penalty for the absolute value of the coefficients allows LASSO to shrink coefficients of non predictive variables to exactly zero, therefore functioning as a feature selection technique.

However, a variable can be flagged as a predictor even if its association with the response variable is only due to chance. This is because one major limitation of LASSO is that it

tends to select different sets of variables under small perturbations of the data, especially in high-dimensional settings or when variables are highly correlated [12]. Such instability can compromise reproducibility and interpretability, as a variable selected in one instance of the model may be omitted in another, therefore questioning its true relevance and effect as a predictor. Stability Selection, on the contrary, aims at selecting only a stable set of predictors, obtained by perturbing the data multiple times (e.g. by subsampling) and then keeping only the variables that have been selected the most. Although this may yield a smaller set of features, the resulting selection is more reliable, since the probability of including variables that were previously chosen by chance is reduced.

### 1.1.2. The idea behind Stability Selection

More formally, Stability Selection starts from the concept of regularization path to then define the new concept of stability path.

A regularization path is defined as the value of the coefficient  $\beta$  for each variable  $k$  and for each value of the regularization parameter  $\lambda$  ( $\{\hat{\beta}_k^\lambda, \lambda \in \Lambda, k = 1, \dots, p\}$ ).

Stability paths, on the other hand, are defined as the probability of each variable of being selected after randomly subsampling the data. In notation, let  $I$  be a random subset of  $\{1, \dots, n\}$  of size  $\lfloor n/2 \rfloor$ <sup>1</sup>, drawn without replacement, and  $\hat{S}^\lambda(I)$  be the set of variables selected by the base LASSO procedure with penalization parameter  $\lambda$ . Then, for every set  $K \subseteq \{1, \dots, p\}$  of variables, the probability<sup>2</sup> of being in the selected set  $\hat{S}^\lambda(I)$  is

$$\hat{\Pi}_K^\lambda = \mathbb{P}\{K \subseteq \hat{S}^\lambda(I)\} \quad (1.2)$$

For every variable  $k = 1, \dots, p$ , its stability path is therefore just the collection of the selection probabilities  $\hat{\Pi}_k^\lambda, \lambda \in \Lambda$ .

Stability Selection consists in identifying the variables that exhibit sufficiently high selection probabilities along their stability paths. Specifically, for a chosen threshold  $\pi \in (0, 1)$ , the data are perturbed repeatedly (say,  $N$  times), and variables with a selection probability exceeding  $\pi$  are deemed stable. The subset of stable variables is therefore defined as:

$$\hat{S}^{stable} = \{k : \max_{\lambda \in \Lambda} \hat{\Pi}_k^\lambda \geq \pi\} \quad (1.3)$$

---

<sup>1</sup> $\lfloor n/2 \rfloor$  was chosen because it represents most closely the bootstrap procedure.

<sup>2</sup>The probability is with respect to the random subsampling.

### 1.1.3. The stability score

The Stability Selection algorithm, as originally proposed, involves two key parameters: the regularization parameter  $\lambda$  and the selection probability threshold  $\pi$ . While  $\lambda$  controls model complexity (with smaller values leading to more complex models),  $\pi$  determines the stringency of the selection criterion, which becomes more restrictive as  $\pi$  increases. Together, these two parameters balance the trade-off between false positives and statistical power, making their appropriate calibration essential. However, in the original framework, optimization was typically performed with respect to one parameter while keeping the other fixed at an arbitrary value. The absence of an analytical framework to jointly derive their optimal settings can therefore be problematic, as manual tuning may yield suboptimal or unstable results.

To solve this calibration problem, Bodinier *et al.* [13] proposed an index called stability score, that allowed for a mathematical derivation of the optimal value of both  $\lambda$  and  $\pi$ . Let  $H_\lambda(k) \in \{0, \dots, N\}$  be the selection count for each of the  $p$  features  $k$ . A feature is then defined as

- stably selected if  $H_\lambda(k) \geq N\pi$
- stably excluded if  $H_\lambda(k) \leq N(1 - \pi)$
- unstably selected if  $(1 - \pi)N < H_\lambda(k) < N\pi$

Under the most unstable feature selection procedure, all the features would have the same probability of being selected  $\gamma_\lambda = q_\lambda/p$ , with  $q_\lambda = \lfloor 1/N \sum_{k=1}^p H_\lambda(k) + 1/2 \rfloor$  being the average number of selected features across  $N$  models fitted with penalty  $\lambda$  on the different subsamples of the data. Under the further assumption of independence of the different subsamples, it can be shown that  $H_\lambda(k)$  follows a binomial distribution

$$H_\lambda(k) \sim B(N, \gamma_\lambda) \quad (1.4)$$

By further considering the  $p$  selection counts as independent, the likelihood of the specific classification described above can be derived as:

$$L_{\lambda, \pi} = \prod_{j=1}^N \left[ (1 - F_{N, \gamma_\lambda}(N\pi - 1))^{\mathbb{1}_{\{H_\lambda(k) \geq N\pi\}}} \times \right. \\ \left. (F_{N, \gamma_\lambda}(N\pi - 1) - F_{N, \gamma_\lambda}(N(1 - \pi)))^{\mathbb{1}_{\{(1-\pi)N < H_\lambda(k) < N\pi\}}} \times \right. \\ \left. F_{N, \gamma_\lambda}(N(1 - \pi))^{\mathbb{1}_{\{H_\lambda(k) \leq N(1-\pi)\}}} \right] \quad (1.5)$$

where  $F_{N,\gamma_\lambda}$  is the cumulative probability function of the binomial distribution with parameters  $N$  and  $\gamma_\lambda$ .

Finally, the stability score is defined as the negative log-likelihood under the hypothesis of equi-probability of selection,  $S_{\lambda,\pi} = -\log(L_{\lambda,\pi})$ .

This score measures how unlikely a model can arise from the null hypothesis, so the higher the score, the more stable the set of selected features. The advantage of such an index is the possibility of a multi-way optimization problem because the optimal couple  $(\lambda, \pi)$  can be simultaneously found as:

$$(\lambda^*, \pi^*) = \arg \max_{\lambda, \pi} S_{\lambda, \pi} \quad (1.6)$$

To show the result of the equation 1.6, the usual reference plot is called "calibration plot" (see Figure 1 for an example). In this plot, the  $\lambda$  and  $\pi$  axes show the refinement grid of parameters chosen to find the optimal parameters (the dashed line intersecting at the optimal pair  $(\lambda^*, \pi^*)$ ), the  $q$  axis shows the number of predictors chosen for each combination of parameters, and finally the colors represent the stability score, with highest values corresponding to the darkest and optimal region.

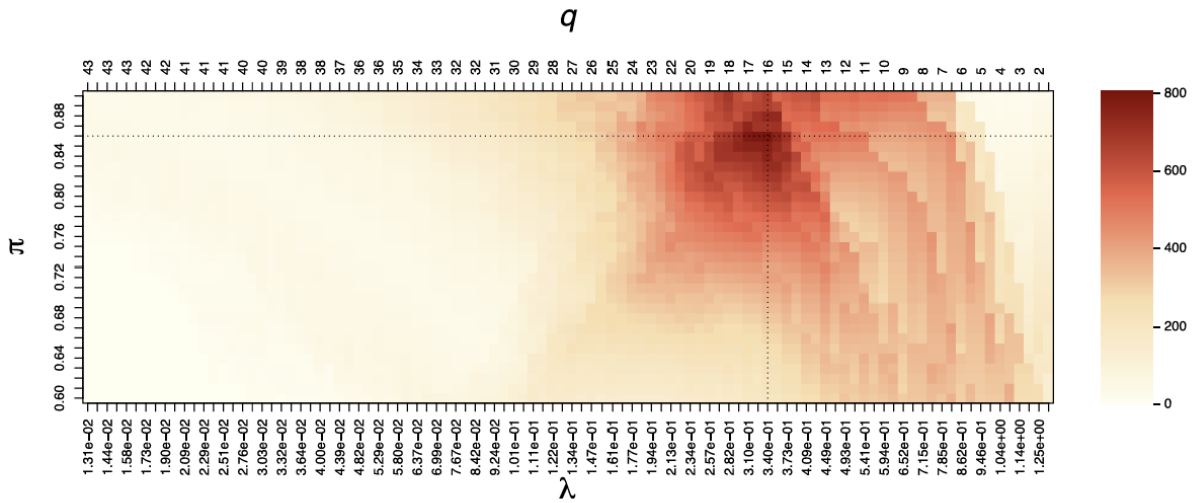


Figure 1: Example of a calibration plot for a generic Stability Selection model (source: Bodinier *et al.* [13]).

### 1.1.4. Predictive power comparison

In this thesis, Stability Selection and LASSO will be compared in terms of both number of selected predictors and predictive power. For the latter, the best and most known metrics are the  $R^2$  and the  $R_{adj}^2$  to account for the number of covariates. They are defined as:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad (1.7)$$

$$R_{adj}^2 = 1 - \frac{\frac{SS_{res}}{n-k-1}}{\frac{SS_{tot}}{n-1}} \quad (1.8)$$

where  $SS$  is the sum of squares (residual or total),  $n$  is the number of observations and  $k$  is the number of covariates.

However, since our response variable is binary, a different metric will be used. While several pseudo- $R^2$  have been proposed, the chosen metric is McFadden's  $R^2$  [17] since it's the only one that also has an adjusted version [18]. As in standard linear regression, we want to account for the number of covariates, because otherwise a more complex model would result more powerful than a simpler one, even if includes non-predictive variables. The formulas are the following:

$$R_{McFadden}^2 = 1 - \frac{l_{tot}}{l_{null}} \quad (1.9)$$

$$R_{adj,McFadden}^2 = 1 - \frac{l_{tot} - k}{l_{null}} \quad (1.10)$$

where  $l_{tot}$  and  $l_{null}$  are the log-likelihoods of the full model and the null model (i.e, the model with just the intercept) respectively, and  $k$  is the number of covariates. Note that unlike the classic  $R^2$ , the McFadden's  $R^2$  rarely goes above 0.4, and already indicates a very good fitting when between 0.2 and 0.4 [19].

## 1.2. Causal inference

Causal inference is the branch of statistics and data science that focuses on answering causal rather than associative questions. This paragraph will present the basic concepts about causal inference and also an insight on why Stability Selection, although formally not a causal method, is expected to find causal predictors.

### 1.2.1. Causal inference in a nutshell

Predictive models aim to identify variables that improve the accuracy of outcome prediction, regardless of the underlying mechanisms [20]. Their primary goal is to capture statistical associations and patterns in the data that help in forecasting future observations. Techniques such as LASSO or other forms of regularized regression are particularly effective in this context, as they select variables that optimize predictive performance while controlling for overfitting. Causal inference, on the other hand, investigates whether a change in one variable would cause a change in another, independently of correlations induced by confounders or shared mechanisms. While a variable may be highly predictive of an outcome, it is not necessarily causal: it may instead be a downstream consequence of the true causal factor, or merely correlated due to confounding.

Establishing causality therefore requires a framework that goes beyond simple association and that explicitly accounts for the underlying data-generating process, often through the use of assumptions on the causal relationships [20]. These assumptions are commonly represented through graphical models known as causal diagrams, such the Directed Acyclic Graphs (DAGs), as introduced and formalized by Hernán *et al.* [21]. DAGs provide a visual and mathematical framework to encode assumptions about causal directions, confounding, and conditional independence, and form the foundation for modern causal reasoning in epidemiology and related fields. In general, a causal DAG is a graph where



Figure 2: A causal DAG in its simplest form (source: Hernán *et al.* [21]).

each node is a variable and the arrows (the directed edges) represent causal relationships. By nature a causal DAG has to be acyclic because no variable can cause itself, either directly or through another variable. Figure 2 shows a causal DAG in its simplest form,

where variable E causes variable D (as an example, E could be an antiretroviral treatment and D could be AIDS). We will call the event of interest "outcome" and the possible cause(s) we are studying "exposure(s)".

Causal DAGs can also be useful to explain why correlation does not imply causation: even if two events are correlated, it's not always the case that one is causing the other. In Figure 2, the association between E and D is purely a causal one. In contrast, the following examples show when it's not the case.

- Common cause (confounder): in Figure 3, E and D are associated, but E does not cause D due to the presence of a common cause, L. For instance, consider D: lung cancer, E: carrying matches in the pocket and L: being a smoker. Clearly carrying matches (E) alone is not sufficient to cause lung cancer (D): the association arises because both are influenced by smoking.
- Common effect (collider): in Figure 4, E and D are associated, but E does not cause D due to the presence of a common effect. For instance, consider the events E: being stressed, D: being genetically predisposed to depression and C: being diagnosed with depression. In this case, both E and D can cause C but clearly being stressed can not influence a genetic predisposition.



Figure 3: A causal DAG with a common cause (source: Hernàn *et al.* [21]).

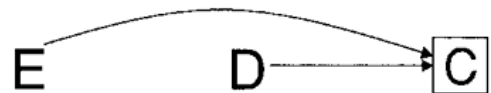


Figure 4: A causal DAG with a common effect (source: Hernàn *et al.* [21]).

As previously stated, the event “L” in Figure 3 represents a "confounder". In our context, all covariates introduced in Chapter 2 will be treated as potential confounders. Accounting for them is essential, as failing to do so could lead to spurious associations. For instance, if a protein is highly correlated with sex and Stability Selection identifies it as significant, it would be unclear whether the observed effect truly reflects the protein's causal influence on the outcome (type 2 diabetes) or merely captures the indirect association induced by sex.

### 1.2.2. On the relationship between stability and causality

Although Stability Selection is not formally a causal inference method, there are theoretical and empirical reasons to expect an intrinsic connection between stability and causality. In general, truly causal relationships tend to persist across different model specifications, subsamples, or sources of random perturbation, while spurious associations often disappear or change direction when the data or modeling assumptions are slightly modified. This notion of invariance is central to causal inference theory (Pearl [22]), which postulates that a causal mechanism should remain stable under interventions or environmental changes. In this sense, model stability can be seen as an empirical proxy for causal robustness. Cadiou and Slama [23] further demonstrated that variable-selection algorithms displaying instability (such as LASSO) are less likely to identify genuine predictors, since unstable associations typically reflect random noise or confounding effects rather than true causal links. Consequently, predictors repeatedly selected across resampling iterations — as is the case in Stability Selection — are more credible as potentially causal candidates, even though SS itself does not provide estimates of causal effects.

### 1.2.3. Causal inference with observational data

The gold standard for causal inference in healthcare are the Randomized Control Trials (RCTs) [24] because, through randomization, they facilitate the elimination of confounding and allow for unbiased estimation of causal effects. They are typically employed to evaluate the efficacy of a specific treatment (or exposure) on a given outcome, particularly in medical and clinical research. As illustrated in Figure 5, the study population is randomly divided into two groups: the experimental group, which receives the new treatment (A in Figure 5), and the control group, which receives a placebo or the standard treatment (B in Figure 5). The outcomes in both groups are then measured and compared; if no significant difference is observed, the new treatment is deemed ineffective, that is, the exposure of interest has no effect on the outcome.

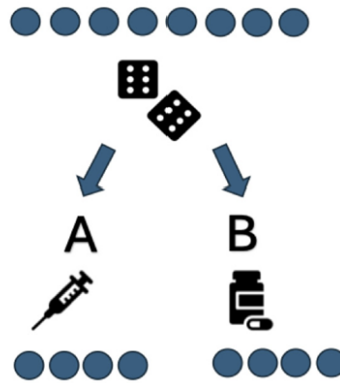


Figure 5: The Randomized Control Trial scheme (source: Braga *et al.* [24]).

In practice, RCTs have many limitations: they require a lot of time, they are very costly, and in some scenarios they can even become unethical [24] (for example, if the aim of the study changes from curing a disease to assessing its causes). Back to the example used in Figure 3, if we wanted to test whether or not smoking is causal of lung cancer, we can not force an entire group of our study to become smokers possibly damaging their health status permanently. In other cases, the design of randomized controlled trials (RCTs) is intrinsically unfeasible. This is particularly evident in the context of omics research, where exposures are molecular measurements rather than external interventions, and therefore can not be randomized. In our case, for instance, the exposure consists of protein concentration levels: it is clearly impossible to randomize such quantities, as one cannot “assign” individuals to have higher or lower levels of a given protein.

For this reason, and thanks to the increasing availability of large cohorts and Electronic Health Records (EHRs) that capture extensive data for numerous patients, an observational cohort study represents a viable alternative approach. By nature, observational cohort studies are longitudinal [25, 26]. Measurements of the variables can be collected either at the beginning of the study (baseline time,  $t$ ) or after a given follow-up period, at a later time point  $t^* > t$ . When the outcome of interest is measured concurrently with the exposure, we refer to it as a *prevalent outcome*. Conversely, when the outcome is measured after the exposure, it is referred to as an *incident outcome*. This temporal separation allows for a more meaningful investigation of potential causal relationships. Coming back to the lung cancer example, an observational cohort study would mean observing two different groups over time, one made of smokers and one made of non-smokers. Then, after the outcome (lung cancer) has been registered for both groups, make a comparison and deduce possible causal links.

However, despite their practicality and ethical feasibility, observational studies also present

several methodological limitations. The most critical issue is confounding, as the lack of randomization makes it difficult to distinguish whether the observed association between exposure and outcome is causal or driven by other unmeasured factors [27]. Even with advanced statistical adjustments, residual confounding can rarely be fully ruled out. Furthermore, observational designs are susceptible to selection bias, since the individuals who participate in the study may differ systematically from those who do not, thus limiting internal validity. Information bias may also arise when exposure or outcome data are inaccurately measured or self-reported. Finally, while observational studies allow for the investigation of real-world data across large populations, their findings may have limited generalizability due to differences in study populations or settings [28, 29].

### 1.3. Mendelian Randomization (MR)

The main method to do causal inference in an observational setting is known as the Mendelian Randomization (MR) algorithm [14, 30]. In a Mendelian Randomization study, three main components are involved: the outcome, one or more non genetic modifiable exposures, and a set of genetic variants. The aim is to estimate the causal effect of the exposure(s) on the outcome. To achieve this, genetic variants are used as proxies for the exposure: they serve as naturally randomized variables associated with the exposure of interest. The method evaluates the effect of these genetic variants on both the exposure and the outcome, and then combines this information to infer the causal effect of the exposure on the outcome. Formally, genetic variants in MR act as Instrumental Variables (IVs): they are not of direct interest themselves but function as instruments that enable the estimation of the aforementioned causal relationship. The name Mendelian Randomization naturally comes from the underlying principle, Mendel's second law [31],

*"that the behavior of each pair of differentiating characters in hybrid union is independent of the other differences between the two original plants"[...]*

In simple terms this means that the inheritance of one trait is independent of (i.e. randomized with respect to) the inheritance of other traits.

The MR algorithm is often referred to as a "natural" version of the RCT (see Figure 6).

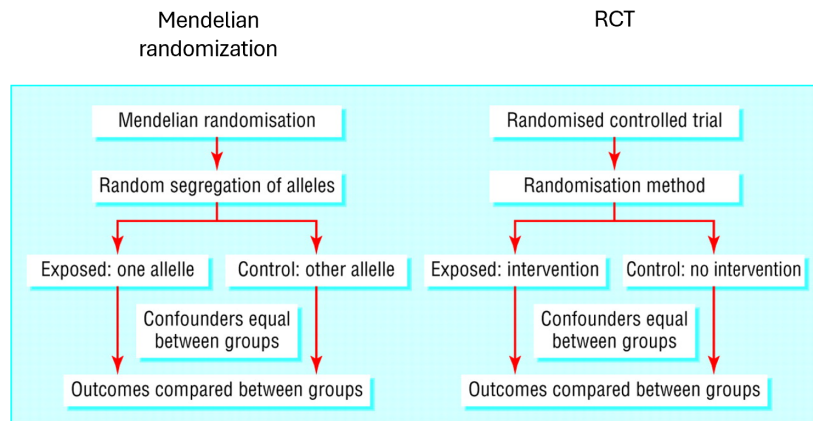


Figure 6: Comparison between Mendelian Randomization and Randomized Control Trials.

We can indeed make the following comparisons:

- in MR, the randomization comes from the random distribution of genetic information in an individual, while in the RCT randomization must be done by the researchers artificially;
- unlike RCTs, MR studies do not require predefined inclusion or exclusion criteria, since genotypes are fixed at conception and independent of lifestyle habits or socio-economic conditions. This makes MR applicable to a wide range of individuals, provided that their genetic and phenotypic data are available. As a consequence, MR can often be implemented using existing biobank or cohort data, without the need for active recruitment or intervention. However, this also implies that researchers have less control over participant characteristics and data quality compared to an RCT.
- MR analyses are often population-specific, since genetic data can vary substantially between different population groups. Therefore, findings from MR analyses are typically most robust within the population from which the data originate and may not directly generalize to other populations. In contrast, RCTs can be explicitly designed to include participants from diverse backgrounds, therefore being more generalizable.

### 1.3.1. MR's assumptions

MR relies on the use of IVs to estimate causal effects. For a variable  $Z$  to serve as a valid IV, it must satisfy three key assumptions:

1.  $Z$  is associated with the exposure of interest  $X$ ;
2.  $Z$  is independent on the possible confounding factors  $U$  that confound the exposure  $X$  and the outcome  $Y$ ;
3.  $Z$  is independent of  $Y$  given  $X$  and  $U$ .

A causal DAG representing these three hypothesis can be seen in Figure 7.

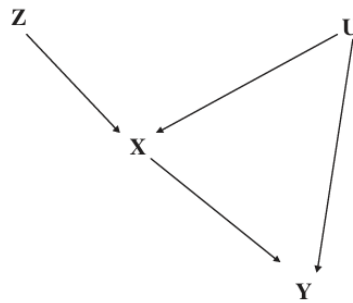


Figure 7: Causal DAG representing the definition of the IV  $Z$  (source: Lawlor *et al.* [14]).

In MR terms, these hypothesis translate into:

1. The genotype  $Z$  is associated with the non genetic exposure  $X$ ;
2. The genotype  $Z$  is independent on possible confounding factors  $U$  that affect both the exposure  $X$  and the outcome  $Y$ ;
3. The genotype  $Z$  is related to the outcome  $Y$  only via the association with exposure  $X$ .

These three hypothesis are enough when dealing with null hypothesis testing, i.e testing whether or not a given exposure has an effect on the outcome. However, results are usually not conveyed through point estimates only. In most cases, the estimated effects are returned with their respective confidence intervals. To do so, a fourth hypothesis is needed:

4. All the associations in the DAG of Figure 7 are linear and unaffected by statistical interactions.

This last assumption clearly becomes a problem when the outcome variable is binary and the best way to summarize the results becomes using odds ratio, which are exponential functions. However, as shown by Wu and Wang [32], this hypothesis can be circumvented, by thinking of the binary variables as noisy representations of underlying continuous measurements. Figure 8 shows their reasoning in the most general case where both the outcome  $X$  and the exposure  $Y$  are binary.

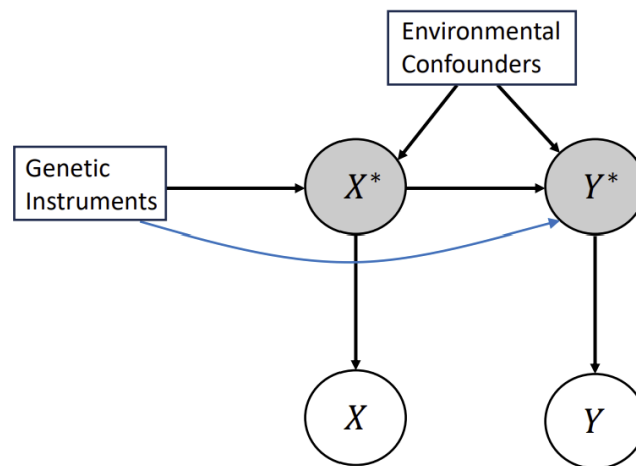


Figure 8: Underlying scheme of MR with binary exposure  $X$  and outcome  $Y$  (source: Wu and Wang [32]).

Their main idea is that the binary variables are simply dichotomizations of the continuous (unmeasured) traits, i.e  $X = 1$  if  $X^* > c_X$  and  $Y = 1$  if  $Y^* > c_Y$ , where  $c_X$  and  $c_Y$  are appropriate thresholds. This is indeed the usual situation with most diseases, for example diabetes ( $Y$ ) is diagnosed if blood glucose levels ( $Y^*$ ) surpass certain thresholds. With this new underlying scheme, the causal relationships are always between the continuous exposure  $X^*$  and outcome  $Y^*$ , therefore requiring only assumptions 1-3 plus the extra assumption that the continuous underlying traits must follow a Gaussian distribution (or any known distribution after standardization).

### 1.3.2. Estimating causal effects using MR

In this work, we focus on single-SNP MR, where a single instrumental variable is used for each exposure. In our case, these instruments are genetic variants known as Single-Nucleotide Polymorphisms (SNPs).

For single-SNP MR, the estimate of the causal effect is done via the Wald estimator

$$\hat{\beta}_{IV} = \frac{\hat{\beta}_{ZY}}{\hat{\beta}_{ZX}} \quad (1.11)$$

where  $\hat{\beta}_{ZY}$  is the regression coefficient for the outcome Y on the SNP Z, and  $\hat{\beta}_{ZX}$  is the regression coefficient for the exposure X on the SNP Z.<sup>3</sup>

Two versions of MR are possible: one sample and two sample. The main differences in terms of potential biases, statistical power, and applicability are summarized in Table 1.

	One sample MR	Two sample MR
Data sources	Same study	Two different studies
Population	Homogeneous by nature, since the study is the same	Can be heterogeneous and could introduce bias
Statistical power	Limited if small dataset	Usually higher if external studies are broader
Overfitting risk	High because all estimates come from the same individuals ("bias towards the estimate")	Low if the datasets are independent ("bias towards the null")

Table 1: Comparison between one sample and two sample MR.

In essence, the key distinction lies in the data source: in one sample MR, both the SNP–exposure and SNP–outcome associations are estimated within the same cohort, whereas in two sample MR they are derived from two independent datasets. As a consequence, one sample MR is more prone to overfitting, which can lead to an overestimation of the causal effect, often biased toward the observational association or, in some cases, even resulting in an inversion of the effect direction ("bias towards the effect") [33]. For this reason, results from one sample MR should be interpreted with caution, particularly in studies with small sample sizes or limited statistical power. Two sample MR is also susceptible to bias, as using an external study may involve a broader population, poten-

<sup>3</sup>Note that in case of binary variables, logistic regressions are being performed, so the results are in the log-odds scale.

tially introducing population bias in the genetic data. However, it is expected to be more conservative and have a smaller number of false positives ("bias towards the null"). In this thesis, since we will work with a limited number of patients (1270) we will employ both one sample and two sample MR.

Independently of the method, it's important to select SNPs that are highly associated with the exposure. The higher the association, the smaller the bias in the final estimate. A good metric to evaluate such exposures is the F-statistics [34, 35], that can be computed as follows:

$$F = \frac{R^2(n-1-k)}{(1-R^2)k} \quad (1.12)$$

where the  $R^2$  is the proportion of variability of the exposure explained by the SNPs,  $n$  is the sample size and  $k$  is the number of SNPs used.

When the  $R^2$  is not known, it is also possible to use an equivalent formula [35, 36]:

$$F = \left( \frac{\hat{\beta}_{ZX}}{SE_{ZX}} \right)^2 \quad (1.13)$$

where  $\hat{\beta}_{ZX}$  and  $SE_{ZX}$  are, respectively, the estimated effect and standard error of the SNP on the exposure, obtained from GWAS summary statistics.

Regardless of its computation method, the interpretation remain the same. According to the Staiger-Stock's rule [37]:

- $F < 10$  suggest a weak instrumental variable.
- $F \geq 10$  defines a strong association between the IV and the exposure. <sup>4</sup>

Moreover, especially in two-sample MR, where exposure and outcome data come from different studies, any analysis must be preceded by an essential step: *harmonization*. Harmonization is a crucial process in MR, as it ensures that the exposure and outcome datasets are compatible and can be analyzed jointly. The goal is to align the data on the same reference alleles, ensuring that the effect of each SNP on both the exposure and the outcome is consistently measured. This step prevents mismatches that could otherwise lead to biased estimates or incorrect conclusions.

Listed below are some of the key aspects of harmonization.

- **Aligning effect alleles:** in MR, each SNP has two alleles: the *effect allele* and the *other allele* (sometimes called the non-effect or reference allele). The effect allele is

---

<sup>4</sup>This rule of thumb is based on the observation that an F-statistics greater than 11 ensures that relative bias on the final estimate will be smaller than 10% at least 95% of the time, regardless of the number of IVs used in the analysis.

the one for which the effect size ( $\beta$ ) and standard error are reported in the summary statistics, while the other allele represents the alternative version of the genetic variant. The harmonization process ensures that the effect allele in the exposure dataset corresponds to the same allele in the outcome dataset. If the alleles are not correctly aligned (i.e., the effect allele in the exposure does not match the one in the outcome), the direction of the effect will be reversed, which can result in incorrect causal estimates. For example, if a SNP increases the level of a protein (exposure) but appears to decrease the risk of diabetes (outcome) simply because of an allele mismatch, this could lead to a false interpretation of the causal relationship.

- Handling palindromic SNPs: palindromic SNPs are those for which the alleles are complementary pairs (A/T or G/C). Because they appear identical on both DNA strands, their alignment can be ambiguous. These SNPs are problematic since their allelic orientation is uncertain, and if not handled properly, they can result in incorrect harmonization. During the harmonization process, palindromic SNPs are typically removed or flagged for exclusion to avoid ambiguity.
- Incompatible alleles: if the alleles between the exposure and outcome datasets are completely incompatible (for example, one dataset reports A/G, while the other reports C/T), these SNPs are excluded from the analysis since they cannot be reliably matched.
- Missing or unavailable SNPs: harmonization often also involves the removal of SNPs with missing data or those not present in both datasets.

## 2 | Data engineering

This Chapter describes the datasets, contextualizes them within the original studies from which they were obtained, and explains the pre-processing workflows applied prior to analysis. Most of the data used in this thesis came from the BELIEVE study. [15]

The BELIEVE (BangladEsh Longitudinal Investigation of Emerging Vascular and non-vascular Events) cohort study is a large-scale research initiative aimed at quantifying the burden of non-communicable diseases (NCDs) in urban, rural, and urban slum areas of Bangladesh.

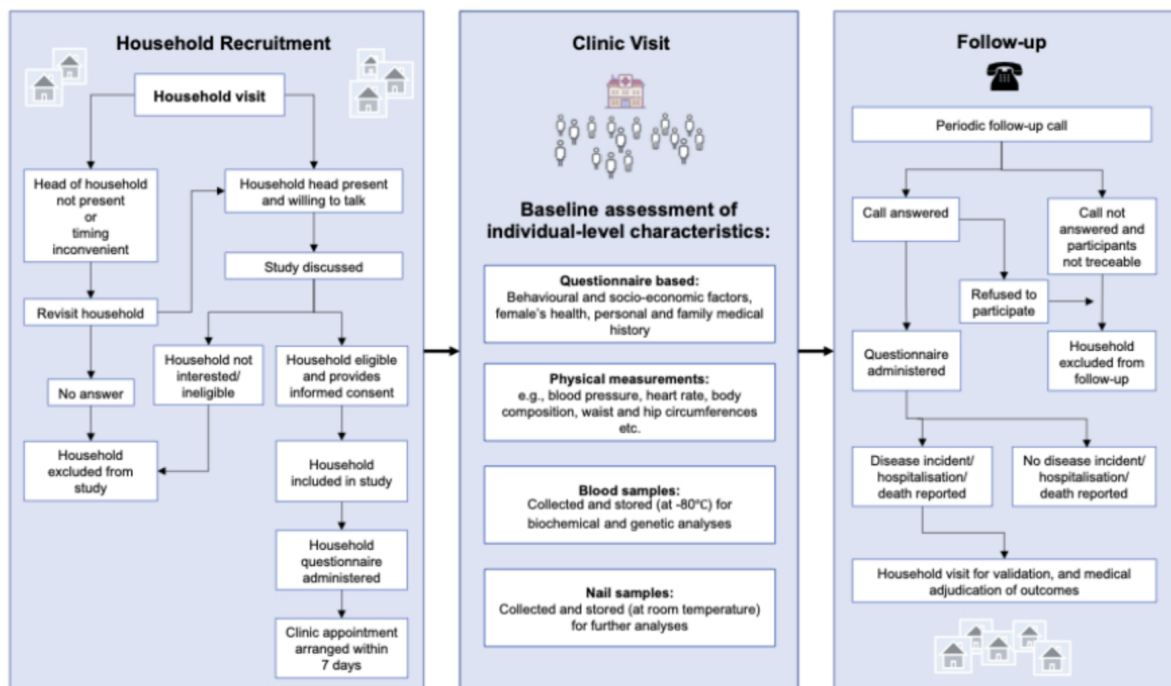


Figure 9: Patient recruitment procedure for the BELIEVE study (source: the BELIEVE study <https://www.believestudy-bangladesh.org/> ) [15].

As shown in Figure 9, the data collection procedure was divided into three steps:

1. household visit, where individuals were asked for voluntary participation to the

study;

2. clinical visit, where individual health data were recorded. These included physical measurements, questionnaire-based answers (like socio-economic status or family medical history). A blood sampling was also performed, therefore containing our 7244 blood proteins measurement.
3. follow-up period, for available patients.

By design, BELIEVE is a longitudinal study, which means that most variables including the protein measurements were taken at baseline time  $t$ . Regarding the outcome, type 2 diabetes, the timeline can be summarized as follows:

- At baseline time point  $t$ , protein levels were measured, and each patient's diabetes status was recorded. If a patient was already diabetic at this time, we refer to this as prevalent diabetes.
- At a later time point  $t^* > t$ , diabetes status was assessed again. New cases identified at this stage are classified as incident diabetes.

In practice, patients who were non-diabetic at baseline ( $t$ ) but developed diabetes by  $t^*$  can be further analyzed, and causal predictors can be searched among the blood proteins.

## 2.1. Stability Selection: pre-processing

The dataset we used was obtained combining several information, which consisted of 7244 different protein measurements for 9934 unique individuals, and a very large number of covariates. From these, we only kept the ones expected to act as counfounders. The retained variables before the preliminary analysis were therefore the following:

- 7244 protein measurements;
- two categorical variables representing whether or not a patient has prevalent (*hx-diab2*) or incident (*ep1\_diab2*) type 2 diabetes;
- *age*, a numerical variable describing the age of a patient, in decimal notation;
- *sex*, a categorical variable for the sex (1 male, 2 female);
- *BMI*, a numerical variable for the patient's BMI (Body Mass Index);
- *DateBloodDraw* and *DateLastMeal*, which are respectively the day of the most recent blood draw and of the last meal (before the blood draw);

- *TimeBloodDraw* and *TimeLastMeal*, which have the same meaning as above, only adding hours, minutes and seconds for a bigger precision;
- *smokstat*, a categorical variable with three levels (Never, Ex/Current, Current), representing a patient’s smoking status;
- *KidneyDisease*, a categorical variable stating whether or not a patient suffers from kidney disease;
- *diadstat*, a categorical variable representing whether or not a patient takes any antidiabetic drug.

First of all, we needed to ensure that the outcome status of the patients were correct and usable for causal inference. Since the protein measurements were taken at time  $t$ , they can be causal of diabetes only for patients that are healthy at time  $t$ . Considering Table 2, we therefore restricted ourselves with only the top row patients: the 212 patients that never developed diabetes and the 1179 patients that were healthy at time  $t$  and ill at time  $t^*$ . Note that patients with  $hxdiab2 = 1$  and  $ep1\_diab2 = 0$  likely represent data entry errors or misinterpretations of the survey questions. Logically, a patient can either have never developed diabetes (both variables equal to zero), have developed diabetes during the study ( $hxdiab2 = 0$  and  $ep1\_diab2 = 1$ ), or already have had diabetes at baseline ( $hxdiab2 = 1$ ), in which case the incident diabetes indicator should be one ( $ep1\_diab2 = 1$ ) since diabetes is a chronic disease that can not disappear during the study. However, this issue does not affect our analysis, as patients who were already diabetic at baseline were excluded, given that no causal interpretation can be made for them. In Appendix A we will instead use all patients and test Stability Selection to find simple associations.

	$ep1\_diab2 = 0$	$ep1\_diab2 = 1$	Total
$Hxdiab2 = 0$	212	1179	1391
$Hxdiab2 = 1$	240	170	310
Total	452	1249	1701

Table 2: Frequency table of patients.

Since a large part of the data was censored for the outcome  $ep1\_diab2$ , the final number of patients was significantly smaller, with only 1391 out of the 9934 patients available for the preliminary analysis. Regarding the covariates, we combined the information stored into the time and date of both blood draw and last meal into a single, more biologically

relevant variable, which we named *fasting\_time*:

- first, the time difference in days was created (and converted into seconds), as  $DateBloodDraw - DateLastMeal$ . This first difference was denoted as *fasting\_time\_1* and ended up being either 0 days (0 seconds) or 1 day (86400 seconds).
- then, the time difference in seconds (*fasting\_time\_2*) was computed, as  $TimeBloodDraw - TimeLastMeal$ . Note that this variable resulted in having some negative values. However, this is completely reasonable because we are not considering the difference in days yet.
- finally, the variable *fasting\_time* was created as  $fasting\_time\_1 + fasting\_time\_2$ .

The variables *TimeBloodDraw*, *TimeLastMeal*, *DateBloodDraw* and *DateLastMeal* were of no interest after the creation of *fasting\_time* and were therefore not directly used as covariates.

Next, we moved to outlier identification. The only possible outliers were in the numeric variables, so we decided to cross-check them using a bagplot for *age* and *BMI*. As it can be seen in Figure 10, no extreme outliers were found: the ones outside the bagplot were just far from the standard but clinically possible.

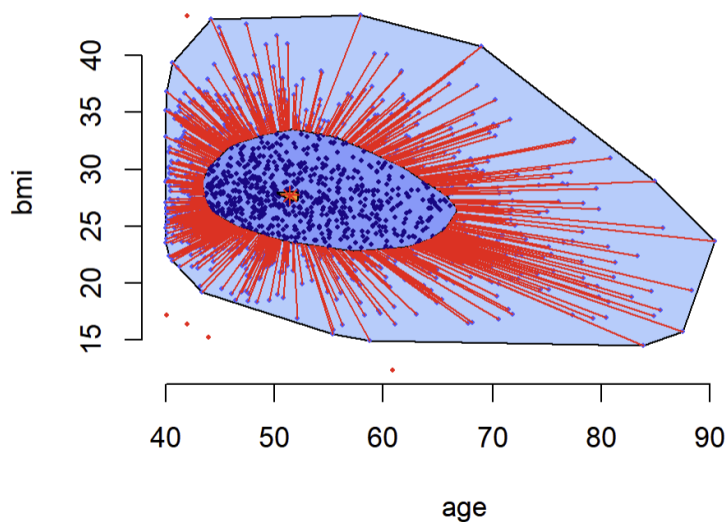


Figure 10: Bagplot of age vs BMI.

Regarding missing values, we only found 2 of them for the *BMI* variable, and decided to impute them by the median value.

We also cleaned the categorical covariates, with a focus on *diadstat* and *smokstat*. We were forced to remove the variable *diadstat* because its distribution (see Figure 11) was extremely unbalanced, meaning that the variable would only add noise if kept in the final analysis.

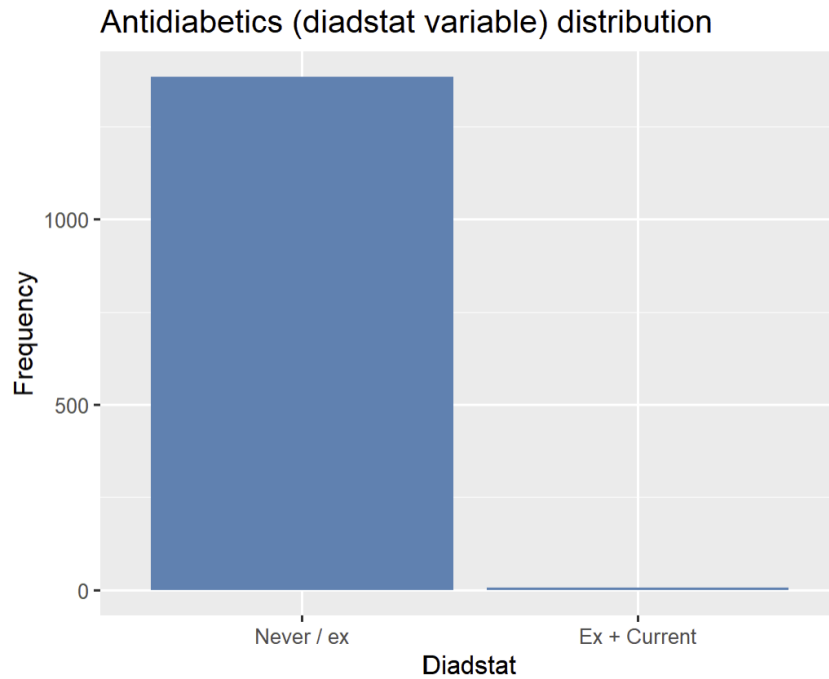


Figure 11: Distribution of the *diadstat* variable.

Regarding *smokstat*, the naming convention used for the categories was quite unclear (see Table 3).

Never	Ex/Current	Current
995	180	216

Table 3: *smokstat* categories distribution.

To assess if the variable would be useful for the analysis, we tested if categories were different. To do this, an ANOVA test was performed with *BMI* as response variable (since we know that the smoking status has indeed an effect on *BMI* [38]). The ANOVA test was of the form

$$H_0 : C_1 = C_2 = C_3,$$

$$H_1 : \exists j : C_i \neq C_j \text{ for some } i = 1, 2, 3.$$

where  $C_i$  is category  $i$ . We got a p-value of 0, so we rejected  $H_0$ . This however only

confirmed that the categories were not *all* equal: looking at the boxplot (Figure 12), we still had doubts on the categories "Ex / Current" and "Current".

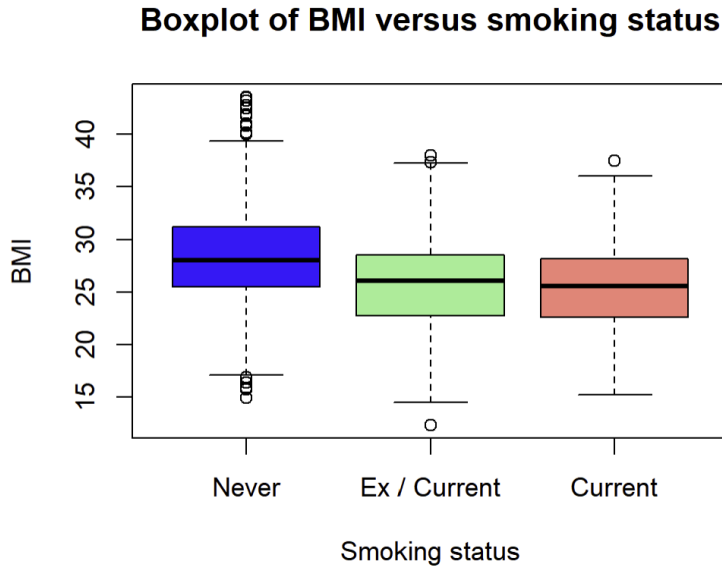


Figure 12: Boxplot of *BMI* vs *smokstat*.

To test the pairwise difference we therefore opted for the Tukey method, that compares the means two categories at a time. In detail, for all pairs  $i, j = 1, 2, 3$ , it tests the hypothesis:

$$H_0 : C_i = C_j$$

$$H_1 : C_i \neq C_j$$

What we got was a p-value of 0 for both "Ex / Current" vs "Never" and "Current" vs "Never", and a p-value of 0.7748 for "Ex / Current" vs "Current", further confirming our doubt. The categories "Ex / Current" and "Current" were not statistically different, and were therefore combined into the new class "At least once".

Finally, every protein was transformed using an inverse-normal transformation, and after that all proteins were scaled. The final dataset, which we will refer to as  $D$ , now contained 1391 patients and 7251 variables.

## 2.2. Stability Selection: dimensional reduction and confounders adjustment

In this section, we explain how the dataset was reduced to make the application of Stability Selection more manageable, and how we adjusted for confounders.

### 2.2.1. Dimensionality reduction on proteomics data

All the models that will be tested for the Stability Selection and LASSO part will be using two different versions of the dataset. The first is using the full dataset  $D$  as created in Chapter 2.1, while the second form is a reduced version of the dataset,  $\tilde{D}$ , designed to decrease the number of proteins by discarding those that show no significant association with incident diabetes, even at a surface level. As noted by Bodinier *et al.* [13], reducing dataset dimensionality facilitates the application of Stability Selection, and the most conservative strategy is to exclude proteins that are not correlated with the outcome at all.

To achieve this, we implemented a Protein Wide Association Study (PWAS) by fitting  $P = 7244$  generalized linear models, one for each protein. For  $i = 1, \dots, P$ , the  $i$ -th model is:

$$Y = ep1\_diab2 \sim BMI + age + sex + KidneyDisease + smokstat + fasting\_time + X_i \quad (2.1)$$

where  $X_i$  is the expression of the  $i$ -th protein. By doing this, every protein is tested individually, one at a time. At the end, we selected only the most significant proteins as those whose p-value is smaller than a Bonferroni-correction threshold ( $threshold = 0.05/P$ , as first proposed by Bonferroni himself [39]). This procedure led to the preselection of  $\tilde{P} = 447$  proteins and can be summarized by the Manhattan plot seen in Figure 13. In the plot, the red line is  $-\log_{10}(threshold)$ , the circles are the proteins. All proteins above the threshold are selected (this is because the condition  $p < threshold$  is equivalent to  $-\log_{10}(p) > -\log_{10}(threshold)$ ).

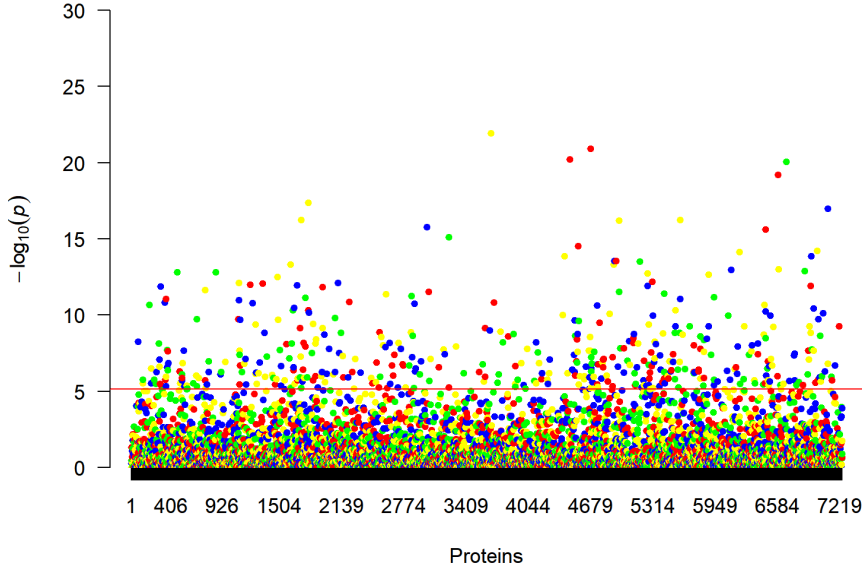


Figure 13: Manhattan plot.

### 2.2.2. Adjusting for confounders

In a standard linear modelling framework, potential confounders are typically included as covariates in the regression model, allowing their effects to be estimated directly. However, in the implementation of Stability Selection adopted in this work, there is no straightforward way to include unpenalized variables, that is, covariates that should not be subject to regularization. To still account for confounding effects, we first regressed out all confounders from the proteins, and then applied Stability Selection using the residual matrix as predictor.

Specifically, the residual matrix  $R$  is an  $N \times P$  matrix, where  $N$  is the number of individuals and  $P$  the number of proteins. Each column  $i$  of  $R$  contains the residuals  $\varepsilon_i$  from the linear model:

$$\begin{aligned}
 X_i = & \beta_0^{(i)} + \beta_1^{(i)} BMI + \beta_2^{(i)} age + \beta_3^{(i)} sex \\
 & + \beta_4^{(i)} KidneyDisease + \beta_5^{(i)} smokstat + \beta_6^{(i)} fasting\_time + \varepsilon_i,
 \end{aligned}
 \tag{2.2}$$

where  $X_i$  is the vector of measurements of the  $i$ -th protein. To be consistent, a reduced residual matrix  $\tilde{R}$  was also created. The idea behind remains the same, the only difference being that  $\tilde{R}$  comes from the reduced dataset  $\tilde{D}$ , i.e.  $\tilde{D}$  is an  $N \times \tilde{P}$  matrix.

## 2.3. Mendelian Randomization: pre-processing

The aim of MR is to assess the causal effect of an exposure of interest (in our case, the protein measurements) on an outcome (incident diabetes) using the genetic variables (the SNPs) as instrumental variables. In order to perform MR, we need two different datasets:

- the *outcome\_data*, that contains the SNP-outcome associations. Specifically, these associations are obtained as results of the GWAS (Genome-Wide Association Studies) of the genetic SNPs on the incident diabetes outcome, *ep1\_diab2*, adjusted on covariates. Two different outcome datasets were used, one for one sample MR and one for two sample MR.
- the *exposure\_data*, that contains the SNP-exposure associations, again obtained as results of the GWAS of the genetic SNPs on each of the protein measurements.

The final dataset that will be used for the MR analyses will be obtained via the harmonization procedure between *outcome\_data* and *exposure\_data*.

### 2.3.1. Outcome data creation: regenie regression

The *outcome\_data* must contain, for each SNP, its estimated association ( $\hat{\beta}$ ) with the outcome, as long as its estimated standard effect and p-value. To obtain those starting from our current data, we performed the regenie regression, a two-step regression method designed by Mbatchou *et al.* [40] to efficiently deal with the estimation of SNP-outcome associations in a large-dimensionality contest. In the first step, the objective is to adjust for confounders (which include both the covariates and the individuals), obtaining a general estimate of the global (genome-wide) genetic effect. To this aim, not all SNPs are used (since it would be computationally heavy), but only a smaller subset of already preselected high-quality SNPs. These SNPs are divided into  $b$  blocks, and then for each block, the following Ridge regression model is performed:

$$y^{(b)} = X^{(b)}\beta^{(b)} + Z\gamma + \varepsilon \quad (2.3)$$

where  $y^{(b)}$  is the logit of the outcome (in our case, type 2 diabetes),  $X^{(b)}$  is the matrix of the SNPs, and  $Z$  is the matrix of the covariates. This first model is known as null model because it includes all the SNPs simultaneously as long as all the covariates.

From these models, the residuals  $\hat{y}$  are obtained to serve as inputs for the second step, whose aim is to compute the SNP-outcome associations as the variability of the outcome that is not explained by the null model. First, the residuals of the null model are computed

as:

$$r = y - \hat{y} \quad (2.4)$$

Then, a score test is performed for each SNP, to infer if a specific SNP has a significant association on the outcome. In formulas, for each SNP  $j$  the quantity  $U_j$  is computed as:

$$U_j = \sum_{i=1}^N G_{ij} r_i \quad (2.5)$$

where  $G_{ij}$  is the genotype of patient  $i$  for SNP  $j$ .

Finally, the test statistics is computed as:

$$Z_j = \frac{U_j}{\sqrt{\text{Var}(U_j)}} \quad (2.6)$$

where

$$\text{Var}(U_j) = G_j^T V G_j \quad (2.7)$$

with  $V$  being the estimated variance of the residuals of the first step.

For each SNP, its corresponding hypothesis test is therefore:

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

Under  $H_0$ ,  $Z_j \sim N(0, 1)$ . Estimates, standard errors and p-values are returned for each test.

To replicate this procedure, the *nf-regenie* pipeline was used: it is based on the *nf-gwas* pipeline, as described by Schönherr *et al.* [41].

The pipeline required three different input files: the genetic data for the patients (which we had access to from a previous internal study), the patients' diabetes status and the covariates to be adjusted in step 1. Not all patients could be matched with their genetic data, resulting in a reduction from 1391 to 1270 patients. The covariates file included all the covariates described in Chapter 2.1 with the preliminary analyses already performed, as well as the first ten genetic principal components to account for population structure, following the approach used by Price *et al.* [42].

For two sample MR, none of the above procedures were necessary, since the outcome data was obtained directly from an external existing GWAS from the GWAS Catalog [43], specifically from Loh *et al.* [44, 45], which included 50533 South-Asian individuals.

Although this study offers a substantially larger sample size (which is still not that big when compared to the standards in the genetic field), it represents a broader South-Asian population rather than exclusively Bangladeshi participants. This ancestry difference may introduce bias due to population stratification. Nevertheless, in the absence of more comparable datasets, these data provided the best available option.

### 2.3.2. Exposure data: instrument selection

The exposure dataset, which was the same for both one sample and two sample MR, contained the associations between each SNP and each protein (again, including the estimated effect  $\hat{\beta}$ , the standard error  $SE$ , and the p-value). This dataset was already available from a pre-existing study. The included SNPs were both *cis* and *trans*: *cis*-acting SNPs are genetic variants located close to the gene encoding the protein, whereas *trans*-acting SNPs are positioned further away, potentially affecting gene expression through indirect mechanisms [46], making them less suitable as instrumental variables. Formally, a SNP is defined as *cis* if it lies within  $\pm 500$  kb (kilobases, i.e., 500000 nucleotides) of the corresponding gene (see for example Xhang *et al.* [47]). Since we adopted a single-SNP MR approach, we included only one SNP per protein. To ensure the strongest possible instrument, we retained the most significant *cis*-SNP for each protein.

The workflow differed when considering only SS proteins or all available proteins.

- for the Stability Selection proteins, to avoid losing proteins because their selected SNP was not in the outcome data, SNPs were manually selected to ensure they were present in both datasets. If the top SNP for a given protein was missing, the second most significant was considered. This procedure continued up until either a SNP was found in both datasets or no significant (or strong) SNP could be used<sup>5</sup>. Because the conventional *cis*-definition threshold was too strict, it was extended to  $\pm 1500$  kb to ensure that all SS proteins had at least one *cis*-SNP. In particular, in one-sample MR, the furthest significant SNP was the sixth (only for one protein, as for the others the third was sufficient). Despite this, all F-statistics exceeded 10, indicating a sufficiently strong association to proceed. In two-sample MR, however, due to the use of a different outcome dataset, retrieving all proteins would have required moving to the seventh SNP. Since this procedure would have resulted in weak instruments (F-statistics below 10), the same exposure data as in the one-sample MR was retained, though at the cost of losing some proteins.

---

<sup>5</sup>Note that this procedure was possible because for the SS proteins we had access to additional SNP-exposure associations.

- for all other proteins, only the most significant *cis*-SNP for each protein was retained (when available). Considering that a large number of proteins had only *trans*-SNPs, the final number of proteins went from 7244 to 1542. By further checking how many of those proteins had their SNP in the outcome data, the number of available proteins reduced to 1477 for one sample MR and 855 for two sample MR.

# 3 | Results

In this chapter we present the results of the various analysis that were performed. We first applied LASSO (Section 3.1), to assess its (un)reliability for finding stable sets of predictors. Then we moved to Stability Selection: in Section 3.2.1 we tested its consistency (i.e, how a different input set affected the output set), while in Section 3.2.2 we compared its performance with that of standard LASSO. In the same Section we also performed a literature review on the stable set identified by SS, to evaluate whether the algorithm could be considered appropriate for causal inference.

Finally, the Mendelian Randomization analysis is discussed in Section 3.3. The Section contains both the MR analysis on the SS selected proteins (Section 3.3.1) and on all available proteins (Section 3.3.2. Furthermore, causal candidates identified by MR also underwent a literature review.

All the analyses performed in this thesis were done using the *R* statistical software [48]. In particular, the libraries *sharp* [49] and *TwoSampleMR* [50] were used to implement SS and MR respectively. The corresponding code is available on the GitHub repository <https://github.com/Trifilio02/Stability-selection-thesis>.

## 3.1. Preliminary LASSO Analysis

The objective of this preliminary analysis was to evaluate the consistency of logistic LASSO when applied to the prediction of incident diabetes. Specifically, we aimed to assess whether LASSO could identify a stable set of protein predictors across both different input sets and different random seeds, since we knew that even small perturbations in the data or in the random splits used for cross-validation may lead to different sets of predictors. Since Stability Selection was specifically designed to overcome this issue by integrating resampling and selection frequency, it is important to quantify how unstable LASSO actually is in this context. Furthermore, because Stability Selection is computationally more expensive, if both approaches yielded similar results, there would be no reason to adopt the more complex one.

Two LASSO models (i.e., two different input configurations) were analyzed: the first,

$LASSO_{full}$ , used the full residual matrix  $R$ , while the second,  $LASSO_{reduced}$ , used the reduced residual matrix  $\tilde{R}$ . The purpose of evaluating both models was to determine whether the preselection step influences the final set of selected predictors. To discriminate the role of chance (which in the case of LASSO comes from the random data splitting in the cross-validation step) and to ensure a robust validation of the results, each model was run 100 times, each with a distinct random seed.

Figure 14 illustrates the resulting distributions of the number of selected proteins under 100 different runs. This number, which we will call  $n_{proteins}$ , can be interpreted as a random variable: for  $LASSO_{full}$ , its mean is 19.45 with a variance of 4.85, while for  $LASSO_{reduced}$  the mean increases to 23.54 with a variance of 10.87. We can thus state that the preselection tends to select more proteins, likely because removing correlated features allows weaker but genuine associations to emerge more clearly. However, this does not eliminate the instability, because the number of selected proteins is still a random variable. Overall, we demonstrated the instability of LASSO, showing the dependence of its output on both randomness and input settings. These results support the move toward Stability Selection to obtain more stable predictors.

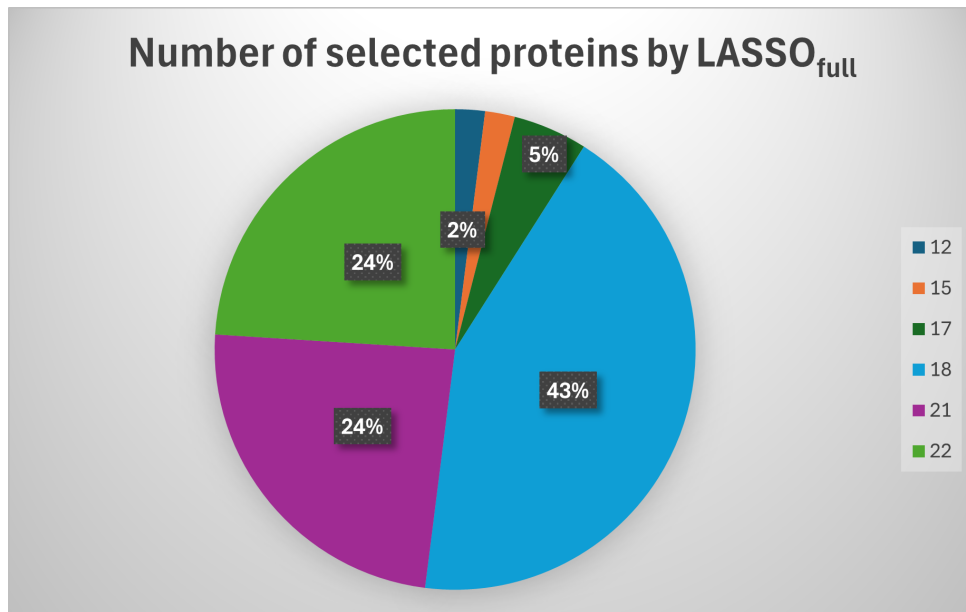
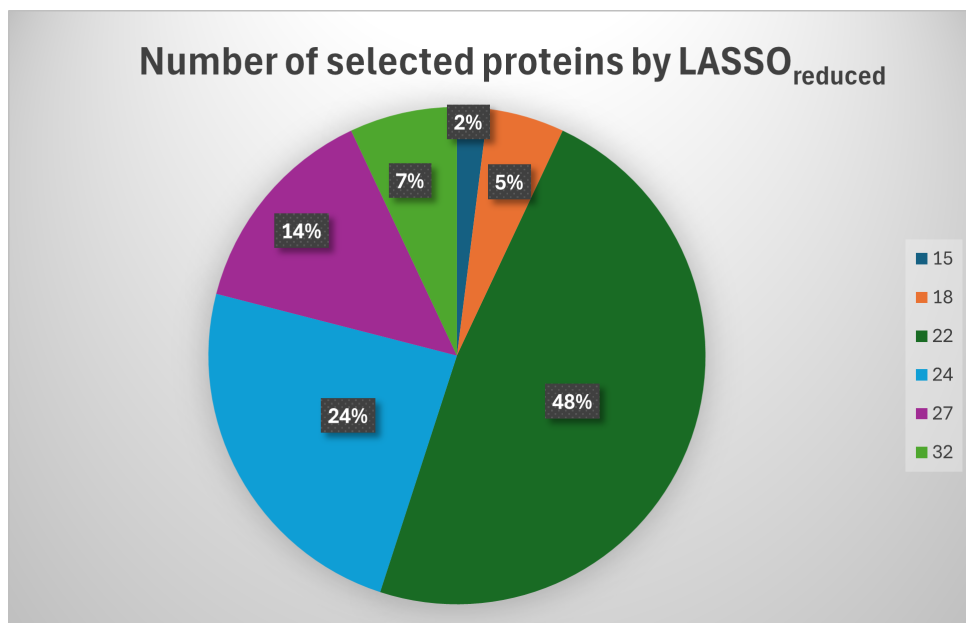
(a) Pie chart for  $LASSO_{full}$ .(b) Pie chart for  $LASSO_{reduced}$ .

Figure 14: Pie charts for the selected number of proteins throughout the 100 runs for each of the two LASSO models.

## 3.2. Stability Selection

In this Section, we present the results of the Stability Selection analysis. Our aim was to assess the consistency of the model, specifically examining how variations in the input set influenced the final set of stable predictors. We then compared these results with LASSO, considering both the number and identity of selected predictors, as well as the overall predictive performance.

### 3.2.1. Consistency analysis

To assess the consistency of the model to different input sets, we started with a baseline model ( $M_{baseline}$ ), and then perturbed it in different ways to see if the final result remains consistent. In total, four different models were implemented, all with *ep1\_diab2* as response variable.

1. The first model is the baseline model  $M_{baseline}$ , with the reduced residual matrix  $\tilde{R}$  as predictor matrix. The number of predictors is therefore 447.
2. The second model is a perturbation of the first. We will call it  $M_{perturbed}$ : its input set consists of the  $k$  proteins selected by  $M_{baseline}$  plus some random proteins (from the full protein list) such that the total number of predictors is still 447. Its purpose is to test whether or not random "noise" (potentially, proteins that were even discarded from the preselection) have an influence on the final set. What we expected is that, if the  $k$  proteins from  $M_{baseline}$  are indeed the stable set we are looking for, they should *all* have been selected by  $M_{perturbed}$ .
3. The third model ( $M_{random}$ ) is a completely random model: its input set consists of 447 random proteins from the full protein list. Here, the number of proteins selected by  $M_{baseline}$  that randomly appear in the input set ( $k^*$ ) can vary from 0 to  $k$ . What we expected is that the stable set returned by  $M_{random}$  contained all those  $k^*$  proteins, provided that  $k^* > 0$ .
4. The final model is  $M_{full}$ , and uses all 7244 proteins, i.e the full residual matrix  $R$ . By doing this, we could test Stability Selection's performances on a large number of variables and when all the information was given to the model.

To discriminate the role of chance, to have a statistically valid analysis and to make a fair comparison with LASSO, all models were run 100 times with 100 different seeds. Since SS is supposed to give the same results independently of chance when run repeatedly on the same dataset, we expected the fixed input set models ( $M_{baseline}$  and  $M_{full}$ ) to

have no dependence at all on the seed. Conversely,  $M_{random}$  was expected by design to produce highly variable results at each run. The crucial aspect, however, was to evaluate how many of the proteins identified in  $M_{baseline}$  could still be recovered when included in a noisy input. This reflects the model's ability to detect stable predictors even in the presence of substantial random noise.  $M_{perturbed}$ , although seed-dependent in its construction (variable input set), was expected to behave always the same, because the stable proteins were always part of the input set. The key elements of each model are summarized in Table 4.

	Input Set ( $I$ )	Expected results	Expected chance dependence
$M_{baseline}$	All 447 rows of $\tilde{R}$	A certain number $k$ of (initially unknown) stable proteins	No
$M_{perturbed}$	$k$ predictors selected by $M_{baseline}$ + 447 - $k$ random rows of $R$	Re-selection of the same $k$ proteins	No
$M_{random}$	447 random rows of $R$	Selection of $k^* \leq k$ proteins	Yes
$M_{full}$	All 7244 rows of $R$	Re-selection of the same $k$ proteins	No

Table 4: Summary of the models' main features.

For each model, we saved the number of proteins flagged as stable, the common proteins with  $M_{baseline}$  and the  $\lambda^*$  and  $\pi^*$  values obtained as solution to the optimization problem in the equation 1.6, to keep track of the optimization procedure between different runs. Figures 15-18 show the calibration plots for the four models for one of the 100 runs. Note that the  $q$  detected by the plots will sometimes differ from the actual number of selected proteins by the models. However, this is an expected behaviour, as explained by the documentation of the *sharp* library used to produce such plots: "simulation studies suggest that the peak corresponding to the largest number of selected features tend to

give better selection performances. This is not necessarily the highest peak (which is automatically retained by the functions in this package)" [49].

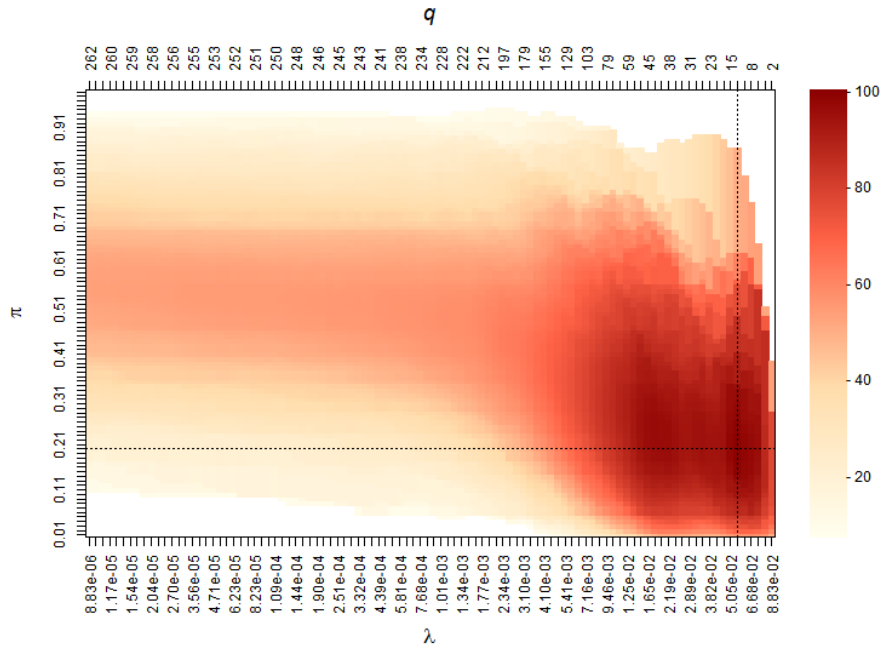


Figure 15: Calibration plot for  $M_{baseline}$ .

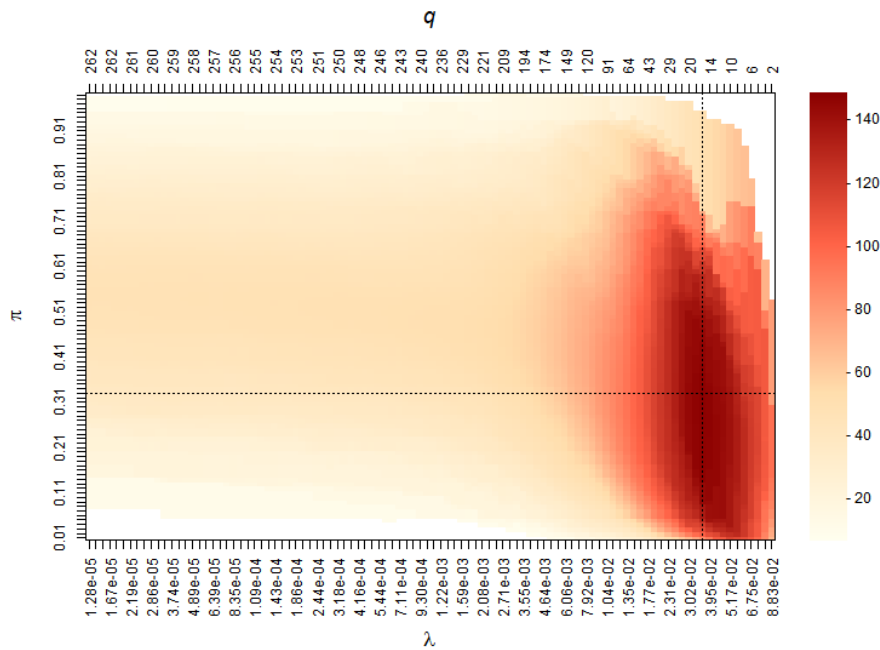


Figure 16: Calibration plot for  $M_{perturbed}$ .

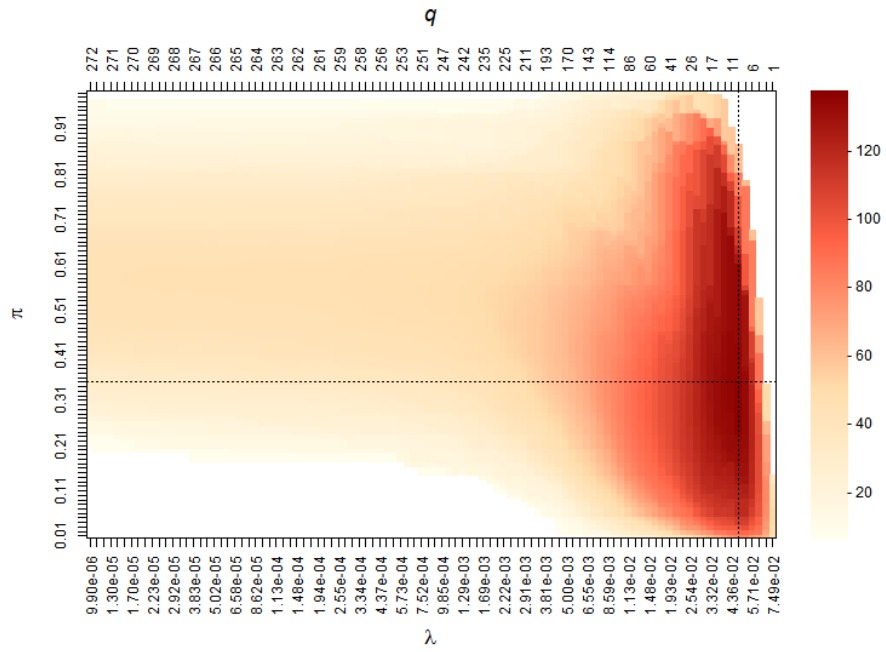


Figure 17: Calibration plot for  $M_{random}$ .

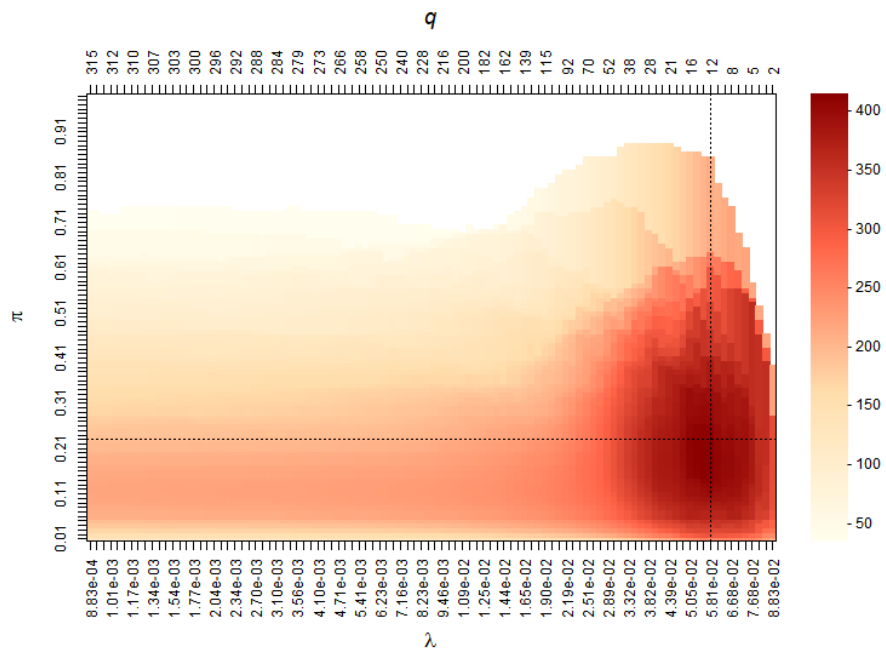


Figure 18: Calibration plot for  $M_{full}$ .

The statistical validation gave the following results:

- both fixed input set models,  $M_{baseline}$  and  $M_{full}$ , proved to be seed-independent, each producing the same output across its own 100 runs. Importantly, both of the optimization parameters  $\lambda^*$  and  $\pi^*$  (which were not set *a priori* and could therefore have been subjected to variation) were *not* affected by the seed selection, remaining consistent across all the 100 runs for their respective models. The main difference between the models came in the number of selected proteins, being 20 for  $M_{baseline}$  and 17 for  $M_{full}$ . This reflects the inner behaviour of LASSO <sup>6</sup>: the more correlated variables are added, the more LASSO may pick one of a correlated group interchangeably across subsamples. Therefore, selection probabilities tend to decrease. As a consequence, three proteins that were already very close to the threshold in  $M_{baseline}$  ended up being discarded in  $M_{full}$ .
- Regarding  $M_{random}$ , the model confirmed its high sensitivity to the random seed. Across 100 runs, the mean number of selected proteins was 5.66, considerably lower than the 20 observed in  $M_{baseline}$ , and the variance was extremely high at 26.61. This behavior is expected: since the number of true predictive factors is random in each run, the number of selected predictors is supposed to behave as a random variable with very high variance. Unsurprisingly, because the input itself is random, the optimal threshold also varies across runs, as illustrated in Figure 19. The model explores almost the entire parameter space when tuning this parameter. The mean of  $\pi^*$  for  $M_{random}$  was 0.5586 with a standard deviation of 0.2047, significantly higher than for the fixed-input set models, which had lower and similar optimal thresholds ( $\pi^* = 0.2$  for  $M_{baseline}$  and  $\pi^* = 0.23$  for  $M_{full}$ ) and, most importantly, *null* variance. Despite this variability, the model still managed to prioritize the stable predictors instead of the random noise: in 87% of runs, all proteins from  $M_{baseline}$  in the input set were also in the final stable set.

---

<sup>6</sup>Recall that, in a nutshell, the decision rule of SS is based on the results of many LASSO models.

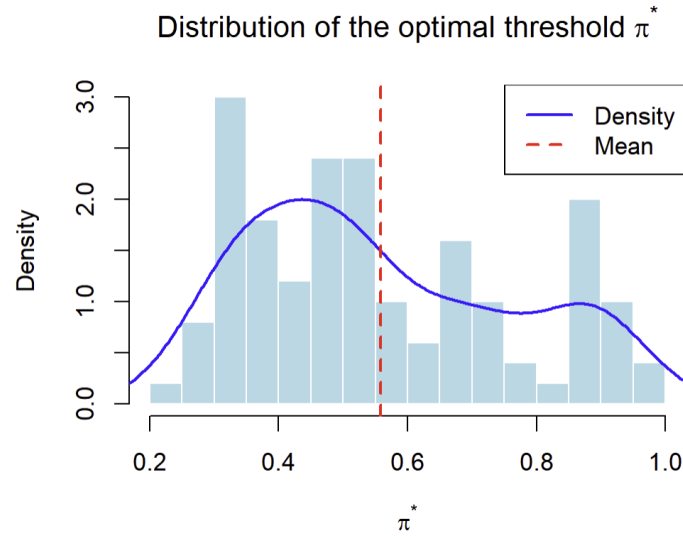


Figure 19: The distribution of  $\pi^*$  for  $M_{random}$  across the 100 runs.

- Despite our initial expectations,  $M_{perturbed}$  did not consistently select the same number of proteins across runs (Figure 20), although it approached the target of  $M_{baseline}$ 's 20 more closely, with a mean of 20.41 and a variance of 1.56.

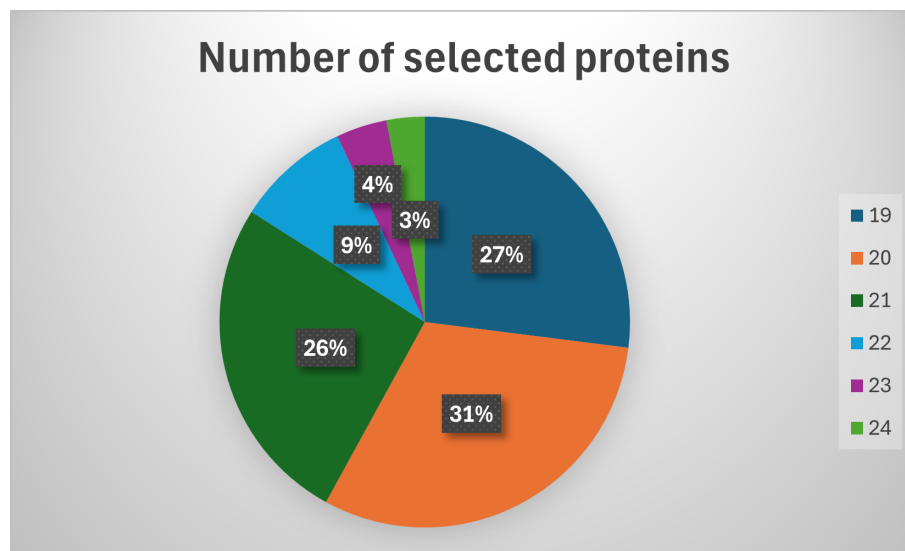


Figure 20: Distribution of the number of selected proteins for  $M_{perturbed}$ .

Regarding  $M_{perturbed}$ 's performance, only in 20% of the runs all 20 proteins from  $M_{baseline}$  were successfully selected. Although this result might appear worse than that of  $M_{random}$ , it is important to note that among the 20 proteins in  $M_{baseline}$ , only 18 are *unique*. Two of them (*Kallistatin* and *Cholinesterase*) were recorded twice,

each measured in two different body regions. We refer to these as the "strong" and "weak" measurements, with the *strong* measurement defined as the one having the highest selection probability in  $M_{baseline}$ .

Thus, if we instead ask how often all 18 *unique*  $M_{baseline}$  proteins are present in the selected set, the answer is 100% of the times, although with some extra proteins. This happens because some of these "noise" proteins are correlated with the others, so depending on the input set they may compete for selection. Overall, while randomness introduces variability and may lead to the inclusion of additional proteins, the essential signal from  $M_{baseline}$  (i.e., its set of 18 unique proteins) is consistently preserved.

As a comparison with  $M_{random}$ , it can be observed that  $\pi^*$  for  $M_{perturbed}$  is considerably more stable (mean of 0.32 and standard deviation of 0.027), with the few extreme values corresponding to the poorest performances (i.e., cases in which both weak signals are lost). This observation further reinforces the conclusions drawn from  $M_{random}$ : since the average number of truly predictive signals in the input set of  $M_{random}$  is lower than in  $M_{perturbed}$  (where the number of stable predictors in the input set is *fixed* at 20 instead of being random in every run), the optimal threshold tends to assume smaller values in  $M_{perturbed}$ , as shown in Figure 21.

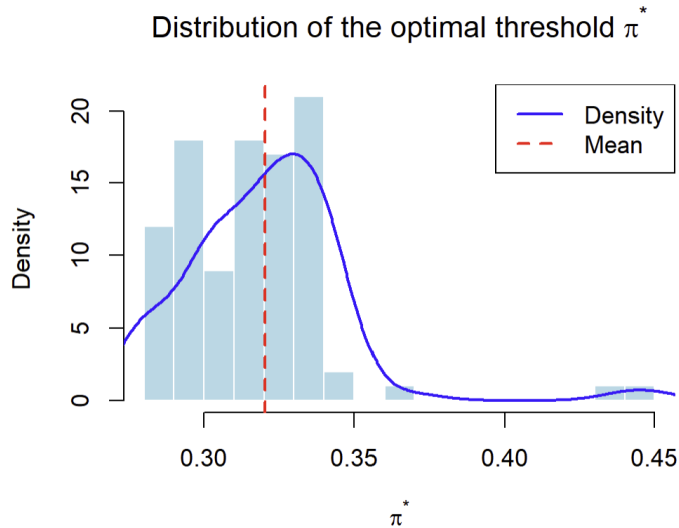


Figure 21: The distribution of  $\pi^*$  for  $M_{perturbed}$  across the 100 runs.

All these analyses showed that the automatically calibrated SS method is stable when applied to a single dataset, unlike LASSO. On average, automatically calibrated SS also selects a slightly smaller set of variables than LASSO. When the input dimension is fixed,

changes in the correlation structure among variables only slightly affect the selection results. Notably, as the dimension increases (i.e., in the full model,  $M_{full}$ ), automatically calibrated SS tends to select a smaller number of predictors, while remaining consistent (since the set of variables selected in the smaller dataset is a subset of those selected in the larger one). This is in clear contrast with classical LASSO, where the number of selected proteins changed significantly between  $LASSO_{full}$  and  $LASSO_{reduced}$ .

As expected, the probability threshold tends to be higher when fewer true signals are present; however, this does not prevent the model from accurately identifying the vast majority of those true signals. Overall, automatically calibrated SS exhibited very stable behavior across different settings (correlation and dimensionality). As discussed previously, this stability may indicate that the selected variables are likely to have a causal relationship with the outcome. The identified set of the 20 stable proteins, which from now on we will refer to as Stability Selection proteins (or simply SS proteins), is shown in Table 5: for each protein, we show its name, its selection probability (as it appears in  $M_{baseline}$ ), its p-value from the preselection, as long as additional information if needed.

	Selection probability	PWAS p-value	Additional notes
<i>Adiponectin</i>	0.86	6.123956e-17	/
<i>Growth Hormone Receptor</i>	0.63	1.264172e-21	/
<i>Insulin-like growth factor-binding protein 2 (IGFBP-2)</i>	0.58	6.224619e-21	/
<i>Cholinesterase</i>	0.56	5.537798e-17	strong measurement
	0.26	2.473688e-16	weak measurement, dropped 80% of the times in $M_{perturbed}$

	Selection probability	PWAS p-value	Additional notes
<i>Group XIIB secretory phospholipase A2-like protein (sPLA(2)-XIII)</i>	0.49	6.028407e-15	/
<i>tRNA-splicing endonuclease subunit Sen15 (SEN15)</i>	0.49	1.734590e-16	/
<i>Complement component C9</i>	0.42	8.285347e-13	/
<i>Heparan-sulfate 6-O-sulfotransferase 2 (H6ST2)</i>	0.42	1.018462e-12	/
<i>N-terminal pro-BNP</i>	0.35	7.487599e-15	/
<i>Stromal cell-derived factor 1 (SDF-1)</i>	0.35	1.346431e-14	/
<i>Kallistatin</i>	0.35	2.644079e-14	strong measurement
	0.21	3.128656e-13	weak measurement, dropped 3% of the times by $M_{perturbed}$ and always dropped by $M_{full}$
<i>SH3 and PX domain-containing protein 2B (SPD2B)</i>	0.33	1.336031e-14	/

	Selection probability	PWAS p-value	Additional notes
<i>Inter-alpha-trypsin inhibitor heavy chain H3 (ITIH3)</i>	0.3	1.080160e-13	/
<i>Small ubiquitin-related modifier 3 (SUMO3)</i>	0.28	1.279207e-13	/
<i>T-cell surface antigen CD2 (CD2)</i>	0.26	1.070880e-10	/
<i>Afamin</i>	0.24	1.750798e-13	/
<i>Epidermal growth factor receptor (ERBB1)</i>	0.22	3.011829e-15	Dropped by $M_{full}$
<i>Sushi, von Willebrand factor type A, EGF and pentraxin domain-containing protein 1:Sushi 15-18 (SVEP1:Sushi 15-18)</i>	0.21	1.288670e-12	Dropped by $M_{full}$

Table 5: The stable set of proteins as detected by Stability Selection.

Notably, variables with the highest selection probabilities do not always exhibit the lowest p-values, reinforcing the lack of direct correlation between stability and association strength.

### 3.2.2. Reliability analysis

Having established the consistency of Stability Selection, we proceeded to assess its reliability. Specifically, three main aspects were investigated:

1. the extent to which the output sets produced by Stability Selection and LASSO are comparable, both in terms of the number and identity of selected predictors;
2. the relative predictive performance of Stability Selection compared to LASSO;
3. how many of the detected SS proteins are hinted at being causal according to the literature.

First, we compared both the number and identity of the predictors selected by the two methods. On average, LASSO selected a larger number of proteins compared to Stability Selection. Moreover, in our experiments, the set of proteins identified by LASSO never exactly matched those selected by SS, indicating that the two methods tend to prioritize different subsets of predictors.

Regarding the predictive performance, by definition, LASSO is expected to perform better in this context. However, our aim is to quantify the extent of this performance gap, i.e. to clearly see if prioritizing stability over predictive performance has a deep impact in the predictive power of SS. In Table 6 we show a comparison in terms of McFadden’s  $R^2$  and adjusted  $R^2$ . All models in the table were run on the same reference seed, ensuring reproducibility.

	$n_{\text{proteins}}$	McFadden’s $R^2$	McFadden’s adjusted $R^2$
$M_{\text{baseline}}$	20	0.201	0.167
$M_{\text{full}}$	17	0.195	0.167
$LASSO_{\text{full}}$	18	0.2	0.17
$LASSO_{\text{reduced}}$	22	0.225	0.188

Table 6: Predictive power comparison between Stability Selection and LASSO.

The results are as follows: for LASSO, the preselection performed via PWAS leads to improved predictive performance. In contrast, for Stability Selection, the preselection has no significant impact, once again highlighting the distinction between stability and predictive power: the 3 additional predictors selected by  $M_{\text{baseline}}$  thanks to the preselection do not

contribute to better prediction. Moreover, Stability Selection is outperformed by LASSO by only 2% at worst, which is a minimal loss when considering that Stability Selection is more consistent and robust, identifies a stable set of predictors, and still achieves satisfactory predictive performance. Recall that, for McFadden's pseudo  $R^2$ , values around 0.2 already indicate a good model fit.

Finally, we conducted a literature review to assess how many and which of the proteins identified through Stability Selection have supporting evidence for a potential causal role. Note however, that some false negatives were expected, as the absence of evidence in the literature does not necessarily rule out potential causality. To ensure a meaningful comparison (i.e., going beyond the dichotomy causal role / non-causal role), we fitted a *glm* including all the 20 SS proteins as covariates, obtaining the odds ratio for each of the proteins<sup>7</sup>. The odds ratio provided an estimate of the direction of the effect (risk or protective), which we could then compare to the literature findings. Table 7 summarizes our findings.

	Causal evidence from the literature	Estimated effect
<i>N-terminal pro-BNP</i>	yes	protective factor
<i>Adiponectin</i>	contradictory results (1 yes and 1 no)	protective factor
<i>Kallistatin</i>	strong	risk factor (strong measurement), protective factor (weak measurement)
<i>Growth Hormone Receptor</i>	strong	protective factor
<i>Epidermal growth factor receptor (ERBB1)</i>	strong	risk factor
<i>Afamin</i>	strong	risk factor

<sup>7</sup>This is because SS alone does not provide coefficients. Therefore, the only way to obtain an estimate was to run the *glm* and obtain the odds ratio from there.

	Causal evidence from the literature	SS-estimated effect
<i>Insulin-like growth factor-binding protein 2 (IGFBP-2)</i>	strong	risk factor
<i>Stromal cell-derived factor 1 (SDF-1)</i>	strong	protective factor
<i>Group XIIB secretory phospholipase A2-like protein (sPLA(2)-XIII)</i>	weak	risk factor
<i>Heparan-sulfate 6-O-sulfotransferase 2 (H6ST2)</i>	weak	protective factor
<i>Small ubiquitin-related modifier 3 (SUMO3)</i>	weak	protective factor
<i>T-cell surface antigen CD2 (CD2)</i>	weak	protective factor
<i>SH3 and PX domain-containing protein 2B (SPD2B)</i>	no	protective factor
<i>Cholinesterase</i>	no	risk factor (strong measurement), protective factor (weak measurement)
<i>Complement component C9</i>	no	protective factor
<i>Inter-alpha-trypsin inhibitor heavy chain H3 (ITIH3)</i>	no	protective factor

	Causal evidence from the literature	SS-estimated effect
<i>tRNA-splicing endonuclease subunit Sen15 (SEN15)</i>	no	protective factor
<i>Sushi, von Willebrand factor type A, EGF and pentraxin domain-containing protein 1:Sushi 15-18 (SVEP1:Sushi 15-18)</i>	no	protective factor

Table 7: Literature review to assess the possible causal role of the SS proteins.

We classified the identified proteins into four groups based on the existing evidence of their potential causal role reported in the literature: verified, strong, weak and none.

**Proteins with a verified causal role** Two proteins belong to this group. Specifically:

- *N-terminal pro BNP* has a confirmed causal role from the literature (see Birukov *et al.* [51]). In the study, the protein is considered a protective factor, a role that is also confirmed by SS.
- *Adiponectin* (which was the most frequently selected protein by SS) has contradicting results: the MR study by Yaghootkar *et al.* [52] flags it as non-causal, while the MR study from Nielsen *et al.* [53] (and our SS findings) detects it as a causal protective factor. This is probably due to a problem of statistical power, considering that the two studies had very different sample sizes: circa 31000 for the former and 756219 for latter.

**Proteins with a strong suggestion of a causal role** Six proteins belong to this group. Specifically:

- the proteins *Kallistatin*, *Growth Hormone Receptor* and *Epidermal Growth Factor Receptor* have several studies that show a strong association with type 2 diabetes, with some of them hinting at a possible causal link. In particular, we refer to Lőrincz

*et al.* [54] for *Kallistatin* (it's an association study that flags the protein as a risk factor, coherently with what SS suggests), Sørensen *et al.* [55] for *Growth Hormone Receptor* (association study that identifies the protein as a protective factor, again in concordance to SS), and the animal study in Li *et al.* [56] for *Epidermal Growth Factor Receptor*, in which the protein is identified as a risk factor. Again, SS confirms this finding.

- the proteins *Afamin*, *Insulin-like growth factor-binding protein 2* and *Stromal cell-derived factor 1* have strong evidences in the literature that suggest a possible causal link. In particular, we refer to Kollerits *et al.* [57] for *Afamin*, a mechanistic and observed populational association study by Wittenbecher *et al.* [58] for *Insulin-like growth factor-binding protein 2* and both the animal study by Tatsuya *et al.* [59] and experimental study by Chen *et al.* [60] for *Stromal cell-derived factor 1*.

**Proteins with weak evidence of a causal role** Four proteins (*Group XIIB secretory phospholipase A2-like protein*, *Heparan-sulfate 6-O-sulfotransferase 2*, *Small ubiquitin-related modifier 3*, and *T-cell surface antigen CD2*) belong to this group. For them, no evidence of a direct causal role in type 2 diabetes was found. However, these proteins are biologically connected to broader metabolic pathways [61–64], such as lipid metabolism, immune signaling, or insulin resistance, which may contribute indirectly to the disease and justify their selection in the Stability Selection step.

**Proteins with no evidence of a causal role** For the final six proteins, no direct evidence of a causal role was found. However, some of them are still associated with type 2 diabetes, but in a way that excludes causal links.

- for *SH3 and PX domain-containing protein 2B* and *Cholinesterase* there is no particular strong evidence of causation.
- *Complement component C9* is linked with the vascular complications of diabetes [65]. Therefore, we are in the situation of reverse causality, i.e diabetes is causing the protein level to vary and not viceversa.
- Two proteins (*Inter-alpha-trypsin inhibitor heavy chain H3* and *tRNA-splicing endonuclease subunit Sen15*) currently lack evidence supporting either a causal role or a known association with type 2 diabetes.
- The role of *Sushi*, *von Willebrand factor type A*, *EGF* and *pentraxin domain-containing protein 1 (SVEP1)* as a biomarker for type 2 diabetes has been suggested in the study by Li *et al.* [66]. However, as a biomarker, SVEP1 may reflect

disease status or progression without necessarily playing a causal role in disease development. Therefore, it is not expected to be a causal predictor.

As a side note, it is interesting that both of the proteins with duplicate measurements (*Kallistatin* and *Cholinesterase*) showed contrasting estimated effects. This pattern is likely due to the weaker measurements being less capable of capturing the true effect of the proteins. This interpretation is further supported by the case of *Kallistatin*: according to the literature, it is a potential causal protective factor, and indeed, our estimate consistent with the literature corresponds to the stronger measurement rather than the weaker one. Overall, we showed that Stability Selection outperforms LASSO in terms of the stability of the selected predictors under changes such as variations in input sets and correlations, while also maintaining competitive predictive performance. Moreover, we are strongly confident that the predictors identified by Stability Selection may also have a causal role, as 8 out of the 18 unique predictors show at least some strong evidence in this direction from the literature. We will now proceed with Mendelian Randomization analyses to compare these findings to the usual practice in causal inference.

### 3.3. Mendelian Randomization

In this Section, we present the results obtained through Mendelian Randomization. Section 3.3.1 reports the MR analyses conducted on the 20 proteins selected by Stability Selection, with the aim of assessing how many of these proteins show evidence of a causal effect. Section 3.3.2 addresses the main clinical research question by extending the MR analyses to all available proteins. The goal of this section is to complement the previous one: indeed, even if all 20 SS proteins were confirmed as causal by MR, this would not ensure that all causal predictors were captured by Stability Selection. Hence, additional causal proteins not identified by SS might still be detected through MR. Given the limited size of our cohort (1270 patients), all analyses were performed using both one sample and two sample MR approaches, leveraging external outcome data from Loh et al. [44, 45]. However, it should be noted that the external dataset was also relatively small (50533 individuals) compared with the typical standards in genetic studies, and most importantly, it introduced potential population bias, as participants were drawn from across South Asia rather than exclusively from Bangladesh. Considering these limitations in both the one sample and two sample settings, we did not expect MR to identify many (if any) causal predictors. To control for multiple testing, we applied Bonferroni corrections to reduce false positives and increase confidence in any detected effects; however, with limited sample size, we could not mitigate the resulting loss of power and the increased

risk of false negatives. All analyses relied on single-SNP MR (Wald estimator) and were restricted to *cis*-SNP measurements.

### 3.3.1. MR on the SS proteins

We began by performing MR on the 20 proteins selected via Stability Selection.

#### One sample MR

First, we considered one sample MR, in which both exposure and outcome data are derived from our initial patient cohort. To account for multiple testing (and therefore to mitigate the number of false positives), we applied a Bonferroni correction of  $0.05/N$  (with  $N = 20$  tested proteins). With this threshold, no protein showed any causal effect. The forest plot in Figure 22 presents the raw (uncorrected) results: estimated causal effects are expressed as odds ratios since the response variable is binary, and proteins with duplicate measurements have two confidence intervals each.

The only protein showing a slightly significant MR estimate, *N-terminal pro-BNP* (p-value of 0.025 before the Bonferroni correction) exhibited a positive effect, which is contradicted by the literature (see Birukov *et al.* [51]) and the *glm* estimate (with an OR of 0.9). This effect is therefore probably due to chance, and considering that it was not detected when corrected for multiple testing, it can not be trusted.

#### Two sample MR

Due to preprocessing steps (see Chapter 2.3), only 11 of the 20 SS proteins were retained for this analysis. Figure 23 shows the results: in this case, even before the Bonferroni correction, no protein is flagged as causal. Interestingly, *N-terminal pro-BNP*, although not classified as causal (probably because of the population bias introduced with two sample MR) displays a point estimate consistent with both the *glm* and the literature, in contrast to the one sample MR estimate.

Overall, neither of the MR analyses provides strong evidence that the SS proteins are causal predictors. However, given the very limited statistical for both one sample and two sample MR (due to the small sample size) and the population bias introduced by two sample MR, the risk of false negatives is high. With the current data, Stability Selection's ability to detect true causal predictors can only be validated through the existing literature, since the MR results can not be trusted. Indeed, larger and more targeted cohorts would be required to robustly confirm our hypothesis.

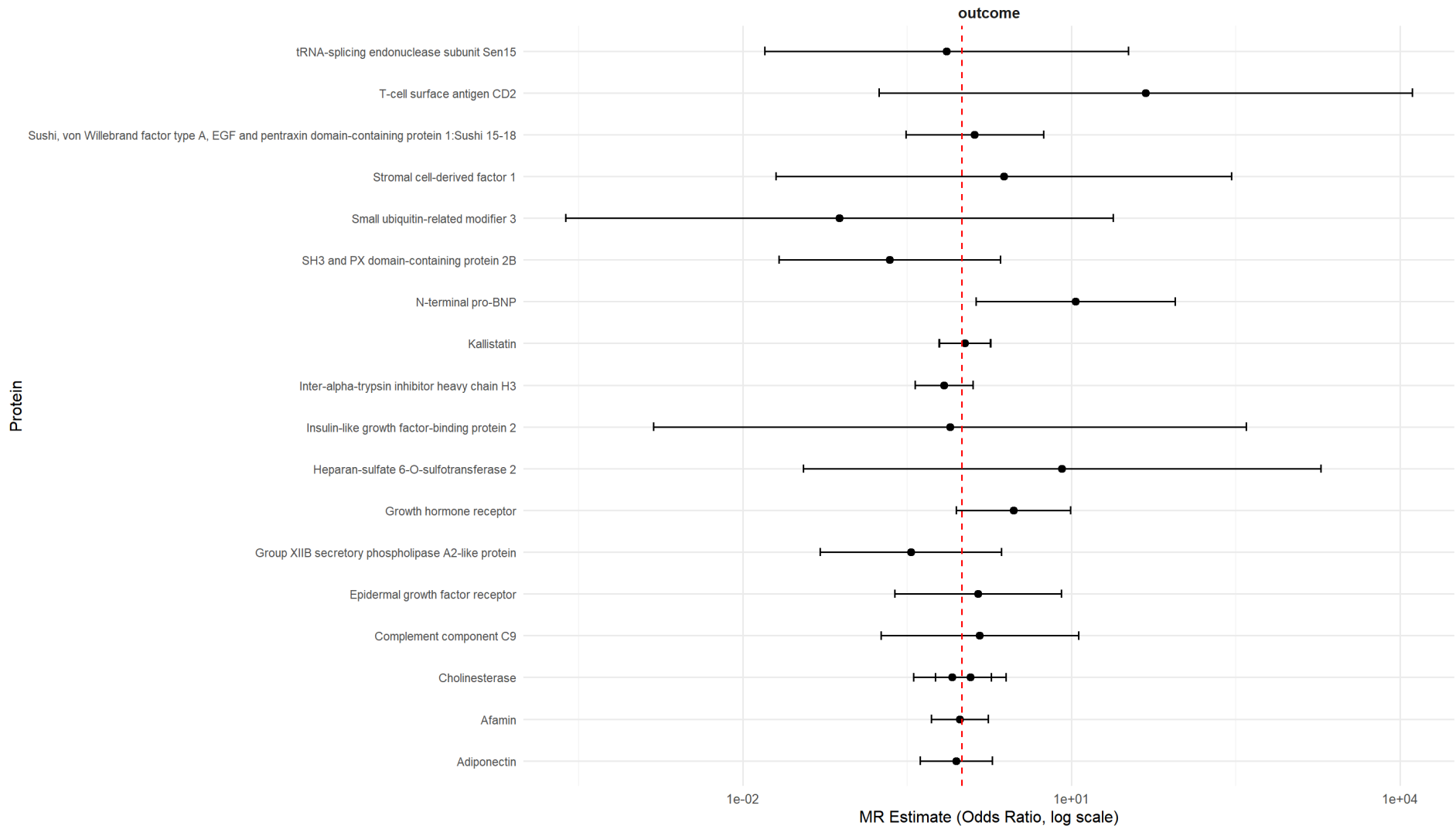


Figure 22: Forest plot for the estimated causal effect of all SS proteins in one sample MR.

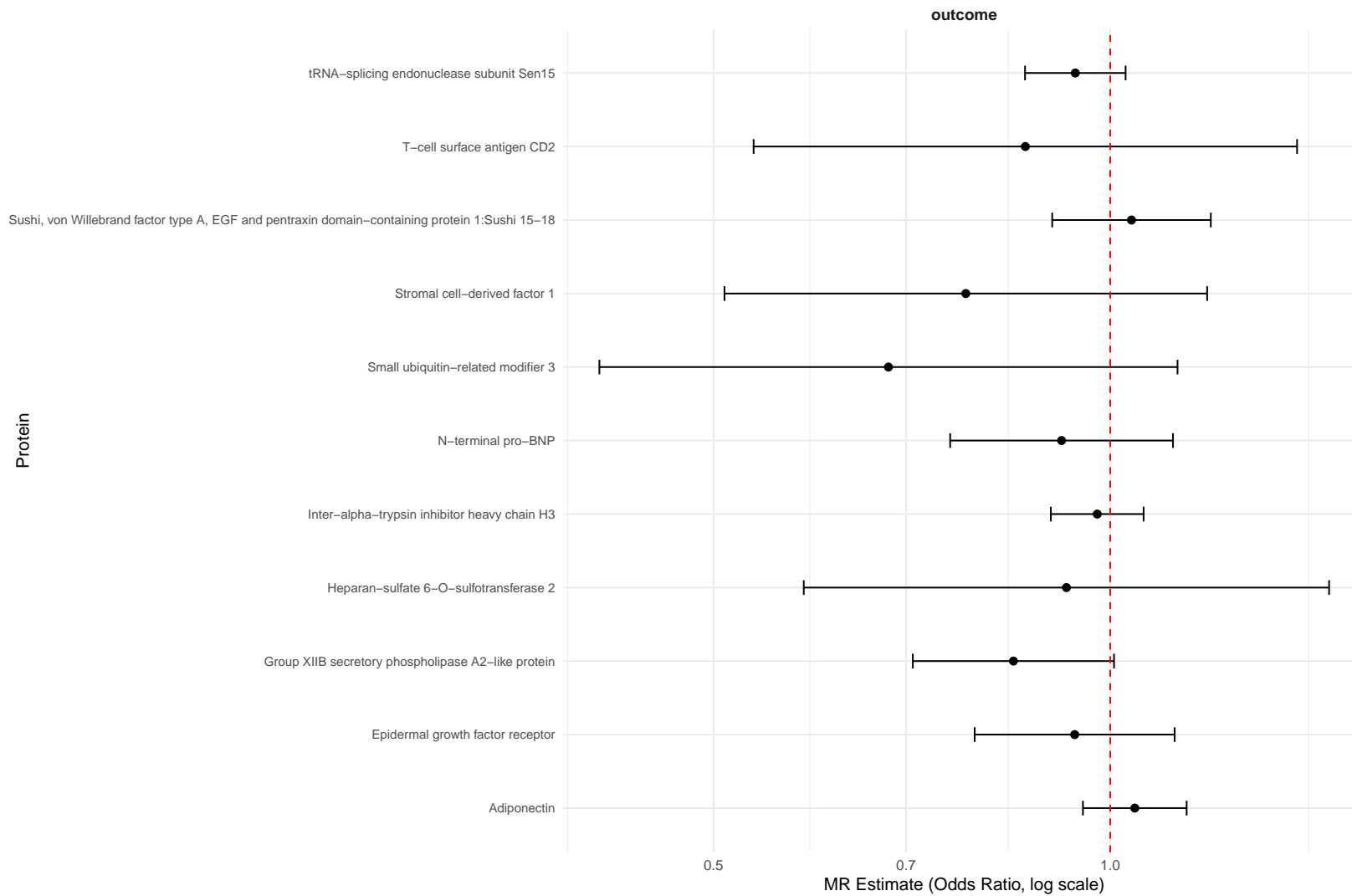


Figure 23: Forest plot for the estimated causal effect for the 11 SS proteins included in the two-sample MR analysis.

### 3.3.2. MR on all the available proteins

We performed MR analyses including all 1542 available proteins, using both one sample and two sample MR approaches. Due to the preprocessing procedure (see Chapter 2.3), we remind that one sample MR considered 1477 proteins and two sample MR 855.

#### One sample MR

In the one sample MR analysis, the results were consistent with our expectations: after correcting for multiple testing (Bonferroni correction threshold =  $0.05/1477$ ), no causal effects were detected. Figure 24 presents the forest plot of the uncorrected results. Among the 37 unique proteins identified (one measured twice and one three times), only 10 were supported by evidence in the literature, while the others were likely false positives, an expected pattern when multiple testing correction is not applied. Of these 10 proteins, 6 have a confirmed causal role, and 4 show a strong association with type 2 diabetes that may suggest a potential causal relationship. These proteins are described below.

**Protein with a confirmed causal role** These 6 proteins have confirmed Mendelian Randomization studies supporting a causal role in type 2 diabetes. Among these,

- 3 act as protective factors: *Natriuretic peptides B* (Pfister *et al.* [67]), *Glucosamine 6-phosphate N-acetyltransferase* (Zhou *et al.* [68]), and *Carbonyl reductase [NADPH] 1* (Zhang *et al.* [69]).
- 3 act as risk factors: *Angiotensin-converting enzyme* (Pigeyre *et al.* [70]), *DnaJ homolog subfamily B member 11* (Li *et al.* [71]), and *Alpha-2-HS-glycoprotein* (Ali *et al.* [72]). Notably, for *DnaJ homolog subfamily B member 11* (just like *N-terminal proBNP*), our MR yielded an opposite effect compared to literature findings. Without correcting for multiple testing, this association is likely due to chance. Similarly, the case of *Alpha-2-HS-glycoprotein* highlights the importance of large sample sizes: while the extensive MR by Ali *et al.* ( $n = 412444$ ) identified it as a risk factor, a smaller MR by Kröger *et al.* ( $n = 10020$ ) reported no causal effect.

**Proteins with strong associations with type 2 diabetes** These 4 proteins are equally divided into:

- 2 risk factors, being *Tyrosine-protein phosphatase non-receptor type substrate 1* (Flores *et al.* [43]) and *Phosphoenolpyruvate carboxykinase, cytosolic [GTP]* (Gómez-Valadés *et al.* [73]), although for the latter our MR analysis found the opposite

direction of effect.

- 2 protective factors, being *Interleukin-15 receptor subunit alpha* (Quinn *et al.* [74]) and *Galectin-3* (Vora *et al.* [75]).

Although some consistency with the literature was observed, it is important to note that this agreement was limited to the *raw* (uncorrected) results, with the smallest p-value being 0.00459. This suggests that the apparent associations may represent false positives, and larger studies will be required to validate them. For now, only the Bonferroni-corrected results can be considered reliable, and they indicate no evidence of causal predictors.

## Two sample MR

Following a similar procedure, the two sample MR analysis did not reveal any significant causal effects after correcting for multiple testing (Bonferroni threshold = 0.05/855). As shown in Figure 25, 47 raw associations were initially detected, with only three proteins supported by previous literature.

- *Peptidyl-glycine alpha-amidating monooxygenase* has been proven to be a therapeutic target for type 2 diabetes in the MR study by Yi *et al.* [76]. Here however, our MR analysis has again the opposite effect problem.
- *Liver-expressed antimicrobial peptide 2* is considered a risk factor due to its strong association to insulin-resistance (see Stark *et al.* [77]).
- *Insulin-like growth factor-binding protein 1* is a protective factor, since high levels are associated with a smaller risk of developing type 2 diabetes, as shown by Petersson *et al.* [78].

Although the smallest uncorrected p-value (0.0002) was slightly lower than in the one-sample MR, the fact that one of these associations was even in the opposite direction further supports that the uncorrected results are due to chance and may be false positives. The absence of significant causal effects is plausibly explained by population differences in the external dataset and by its still modest sample size (50533 individuals), which remains small by genetic standards.

In both one sample and two sample MR analyses, the main limitation lies in the outcome data rather than in the instrumental variables, as the SNP instruments were consistently strong, with F-statistics well above the conventional threshold of 10.

Overall, although both MR approaches identified a few proteins whose inferred causal

roles are consistent with previously published evidence, we can not trust them. Table 8 summarizes the results, with and without Bonferroni correction for multiple testing. For the raw results, we report only the proteins that were confirmed by the literature.

	<b>With Bonferroni correction</b>	<b>Without Bonferroni correction</b>
<b>One sample MR - SS proteins (20 tested proteins)</b>	0	1 (reverse effect)
<b>Two sample MR - SS proteins (11 tested proteins)</b>	0	0
<b>One sample MR - all proteins (1477 tested proteins)</b>	0	10 (3 reverse effects)
<b>Two sample MR - all proteins (855 tested proteins)</b>	0	3 (1 reverse effect)

Table 8: MR results with and without Bonferroni correction.

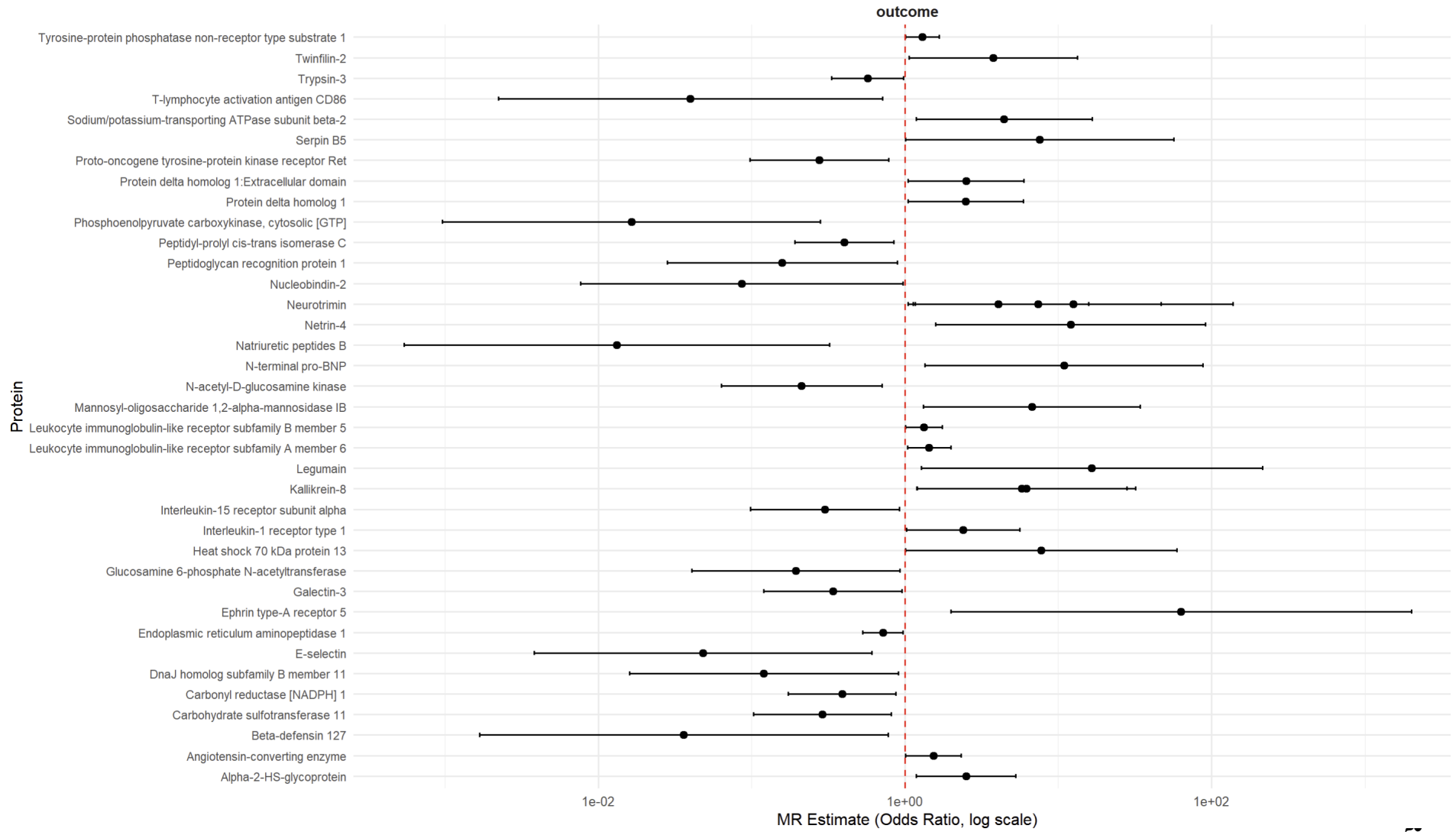


Figure 24: Forest plot including all proteins selected by one-sample MR.

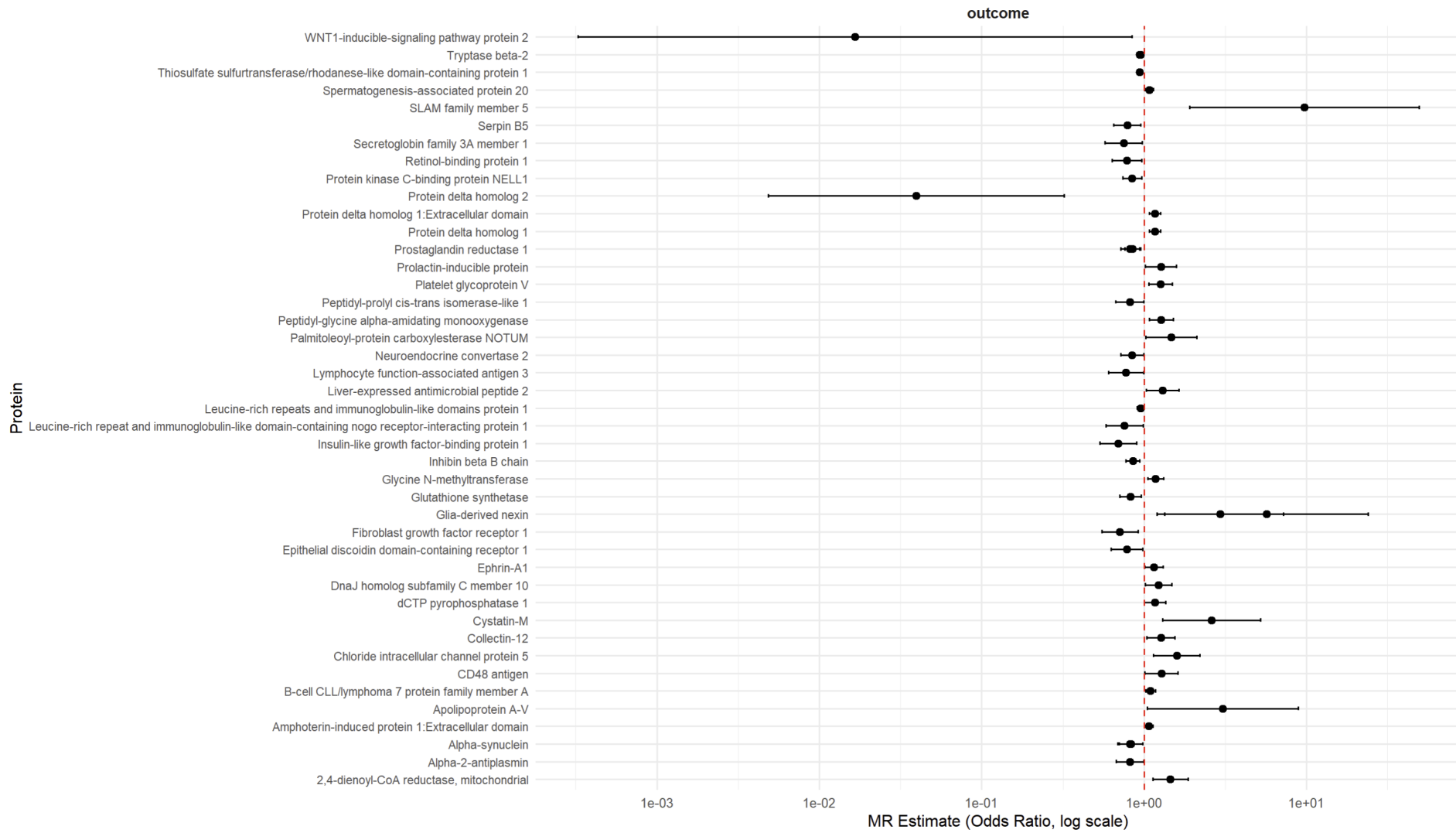


Figure 25: Forest plot including all proteins selected by two-sample MR.



## 4 | Discussion and conclusion

Type 2 diabetes is a complex, multifactorial disease affecting hundreds of millions of people worldwide. Despite extensive research, there remains a growing need to identify causal drivers to improve prevention, diagnosis, and treatment strategies. In particular, blood proteins are promising causal candidates, both because of their functional relevance and their suitability as drug targets for early intervention on the onset of the disease. This thesis therefore addressed two different research questions. From a methodological point of view, it employed the Stability Selection algorithm to identify potential causal predictors among 7244 plasma proteins. While Stability Selection is not a formal causal-inference method, it was expected to perform well in this regard given the documented link between stability and causality in the literature. From a clinical point of view, the second aim of the thesis was to uncover possible causal predictors among the proteins using both Stability Selection and Mendelian Randomization. Regarding the methodological question, the starting point was assessing the performances of SS. First, consistency was addressed: the model was tested across multiple input configurations (including the full protein set, a dimensionally reduced subset, and randomly sampled subsets) and it reliably identified a stable group of 18 unique proteins. Then, SS was compared to the classical LASSO regression, demonstrating greater model stability while maintaining competitive predictive performance. Finally, we investigated whether or not the SS proteins were potential causal candidates in two different ways: a literature review and a comparison with the set of proteins selected by Mendelian Randomization (both one sample and two sample). The literature review rendered encouraging results, highlighting 8 out of the 18 unique SS proteins as potential causal predictors. When moving to MR, however, the limited statistical power of both one sample and two sample MR and the population bias of two sample MR compromised the strength of the final conclusions. None of the SS proteins managed to withstand the correction for multiple testing. When considering the *raw*, uncorrected results, only *N-terminal pro BNP* initially showed suggestive evidence of a causal effect in the one sample MR; however, this association was not consistent with the SS estimates or with previous literature. Similarly, the two sample MR did not identify any significant associations. These results indicate that, under the current data

constraints, MR lacks sufficient power to validate the SS proteins as causal predictors.

To address the clinical research question, MR analyses (both one sample and two sample) were conducted on all the available proteins. Indeed, since SS itself already provided an answer to the question, with the 8 confirmed proteins from the literature, the MR analyses should have either re-established the SS results or find new causal predictors that SS could have missed. However, the MR analyses were not powerful enough: when correcting for multiple testing, none of the *raw* associations remained significant. Indeed, the main constraint of this study lies in the limited statistical power, largely due to the relatively small sample size (for one sample MR and, to a certain extent, also to two sample MR) and heterogeneity of the available cohorts for two sample MR.

In conclusion, although all methods have their own strengths and weaknesses (see Table 9), they should be applied appropriately according to the specific research context and data constraints.

	<b>Advantages</b>	<b>Disadvantages</b>
<b>Stability Selection</b>	Does not suffer of small sample size as much as MR; does not need genetic data	Can not distinguish causality from reverse causality; sensitive to strong confounding
<b>One sample MR</b>	No population bias	Limited statistical power for small datasets; overfitting risk ("bias towards the estimate")
<b>Two sample MR</b>	(Slightly) higher statistical power; smaller number of false positives ("bias towards the null")	Population bias

Table 9: Final comparison of the methods.

Stability Selection is a particularly viable option when genetic data are unavailable or when sample sizes are limited, as its performance is not strongly influenced by sample size. However, it is highly sensitive to confounding and cannot distinguish between causality and reverse causality. In contrast, Mendelian Randomization provides a well-established framework for causal inference but requires substantially larger sample sizes, especially

in the one sample setting. One sample MR is less affected by population bias but tends to overfit, whereas two sample MR is less prone to overfitting but more sensitive to population structure.

With a larger cohort available, we advise to apply both SS and MR (one sample and two sample) to identify causal predictors. In case of overlap between selection, we can have high confidence that the proteins in the overlap are causal. Indeed, SS could especially be useful in situations in which two sample MR is not possible (population or outcome very specific), given the known limits of one sample MR.



# A | Appendix A

In this appendix, we report the analyses conducted on prevalent diabetes. It is important to recall that, in this context, diabetes and protein levels are measured at the same point in time. As a result, we are limited to investigating classical associations rather than causal relationships. No causal link can be established from these associations, as the concurrent measurement of both the exposure (proteins) and the outcome (diabetes) introduces ambiguity in the direction of the effect. Several scenarios are possible: the protein may not be causal at all and simply reflect a downstream consequence of an unmeasured causal process (i.e., it acts as a biomarker); the protein could be causal, but the timing of measurement being after the onset of diabetes prevents us from detecting its true role; or we may observe a spurious association driven by reverse causation, where it remains unclear whether the protein influences diabetes or viceversa. The structure of the Appendix is as follows: Section A.1 presents the preprocessing procedure, that must be re-done because of the use of a slightly different dataset (i.e., more patients), and Section A.2 shows the results of Stability Selection and LASSO on prevalent diabetes.

## A.1. Prevalent diabetes data engineering

Since we are now considering prevalent diabetes, the number of patients is significantly larger than before. In the case of incident diabetes, the analysis was restricted to the 1391 patients who were healthy at baseline (time  $t$ ) and had a measurement at time  $t^*$  (i.e., with a value of *ep1\_diab2* different from "nonfatal censor"). This time, however, we are not concerned with patients' health status at time  $t^*$ , resulting in a much larger initial cohort of 9934 subjects. Three patients were immediately removed due to ambiguous values in the *hxdiab2* (prevalent diabetes) variable, which were recorded as "poss/susp", a "maybe" classification that offers no clear value for analysis.

A different cohort size also necessitated a new pre-processing procedure. While the methodology followed the same steps described in Chapter 2.1, the outcomes differed.

- The *fasting\_time* variable was constructed in the same way as before, i.e by summing the differences in days (*fasting\_time\_1*) and in seconds (*fasting\_time\_2*), then converting the result into total seconds. However, this time we identified some implausible entries: one patient had a two-day eating gap, which seemed highly unlikely, and two others probably misunderstood the question, reporting the time of the meal *after* rather than *before* the blood sample. These three patients were removed.
- Outlier detection was repeated for the variables *age* and *BMI*. As illustrated in Figure 26, five extreme outliers were found: one in *age* (105.04) and four in *BMI* (111.50, 68.31, 65.54, and 56.76). Given that a BMI above 30 already indicates obesity, these values appeared implausible. All outliers were replaced with the median of their respective variables.

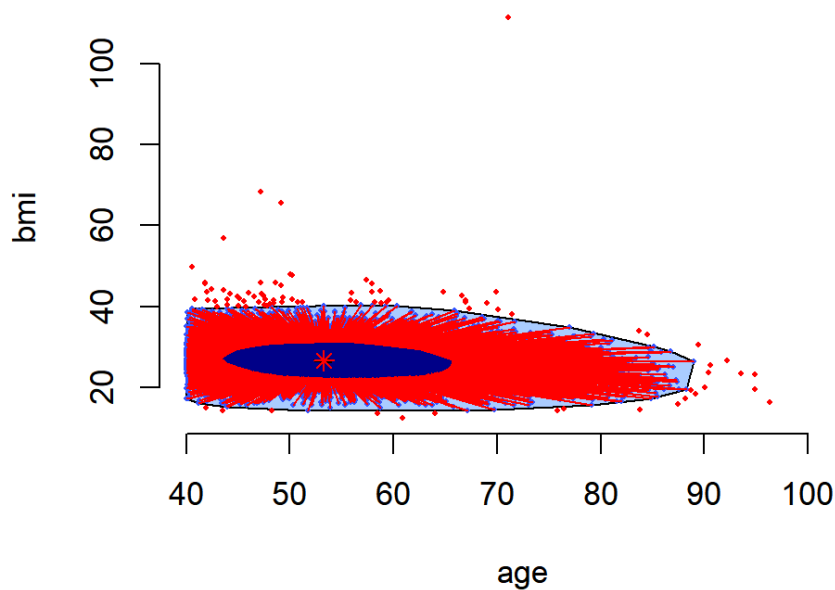


Figure 26: Bagplot for outlier identification for *BMI* vs *age* in the prevalent diabetes dataset.

- Unlike in the incident diabetes dataset, the *diadstat* variable was retained because its distribution was no longer heavily unbalanced (see Figure 27). Although the naming conventions remained unclear, the statistical distinction between the categories justified keeping them as-is.

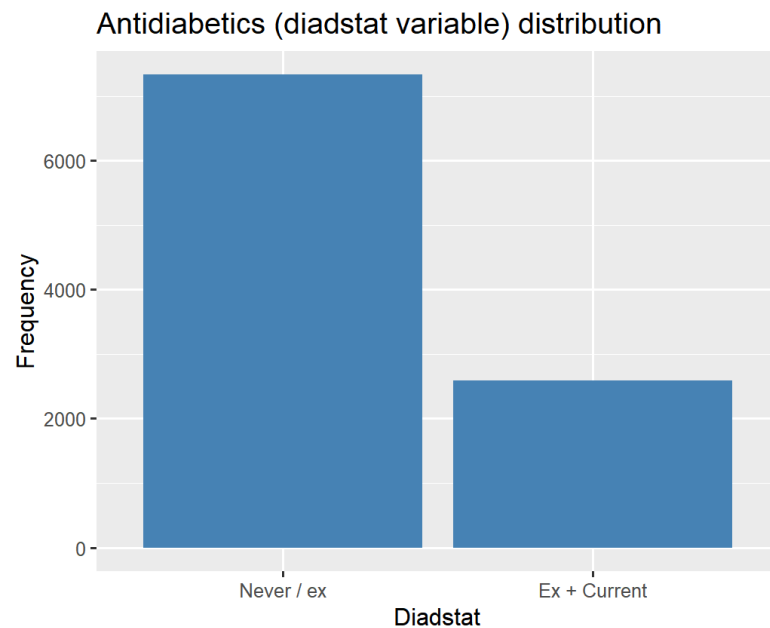


Figure 27: Distribution of the *diadstat* variable in the prevalent diabetes dataset.

- The *smokstat* variable was re-examined for group differences among categories ("Never", "Ex", "Ex/Current", "Current"). No entries were found for "Ex", so we analyzed the three remaining categories using ANOVA followed by Tukey's HSD post-hoc test to assess differences in *BMI*. The ANOVA test yielded a p-value of 0, indicating that at least one group differed significantly from the others. Subsequent Tukey comparisons also returned p-values of 0 for all pairwise contrasts, confirming that all three groups were statistically distinct, as shown in Figure 28.

The variables *sex*, *KidneyDisease* and all 7244 protein measurements required no specific additional preprocessing with respect to what was done in Chapter 2.1. All missing values were imputed using the median (for *BMI*) or mode (for *smokstat*), depending on the variable type.

The final dataset comprised 9928 patients and 7252 variables.

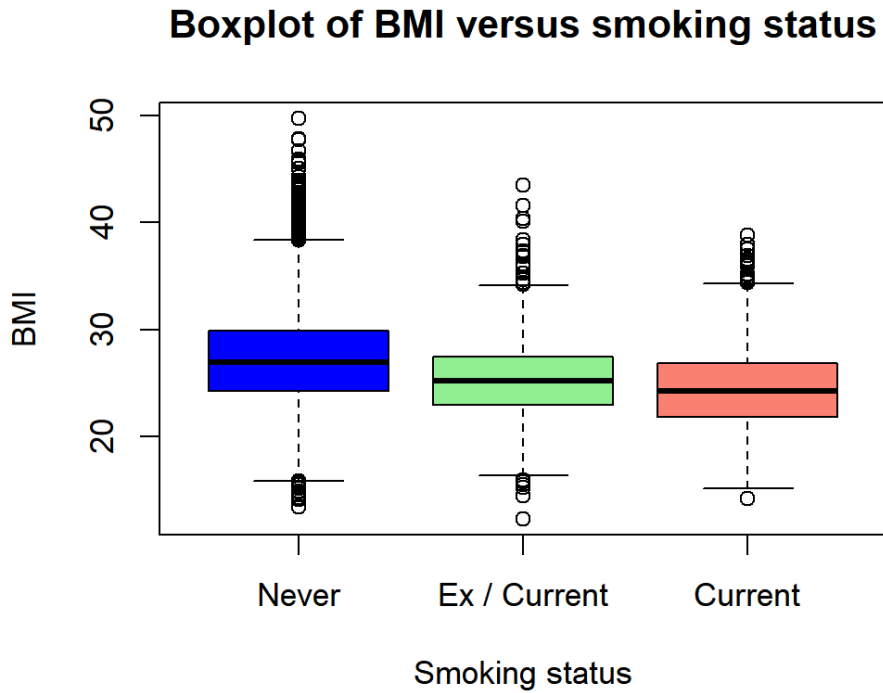


Figure 28: Boxplot of *BMI* by *smokstat* category in the prevalent diabetes dataset.

## A.2. Stability Selection on prevalent diabetes

The aim of this Section is to present the results of both SS and LASSO on the prevalent diabetes dataset. Since the consistency of the method has already been demonstrated in the thesis, we will no longer run the models 100 times. Instead, we will run each of the four models once, to allow for comparison, and run LASSO twice: once with preselection and once without. The rationale behind remains the same:  $M_{baseline}$  has an input set made of the PWAS preselected proteins (this time they are 387, see Figure 29),  $M_{perturbed}$  and  $M_{random}$  have a variable input set and  $M_{full}$  uses all 7244 proteins.

The calibration plots for all the models are shown in Figures 30 - 33: this time, the colored regions appear more vertical, suggesting that  $\lambda$  played a much bigger role than  $\pi$  in the optimization procedure.

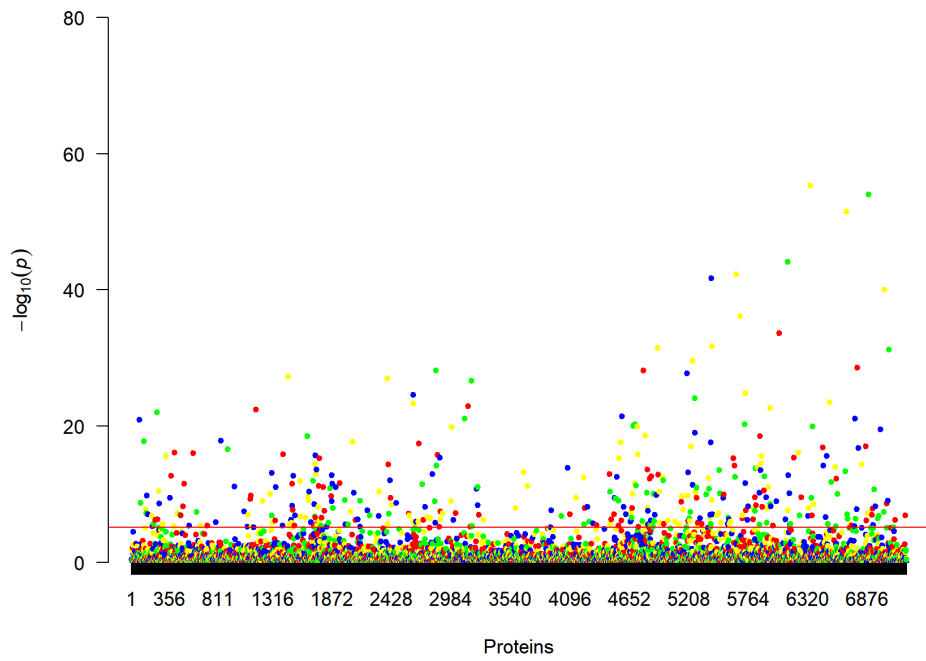


Figure 29: Manhattan plot for prevalent diabetes.

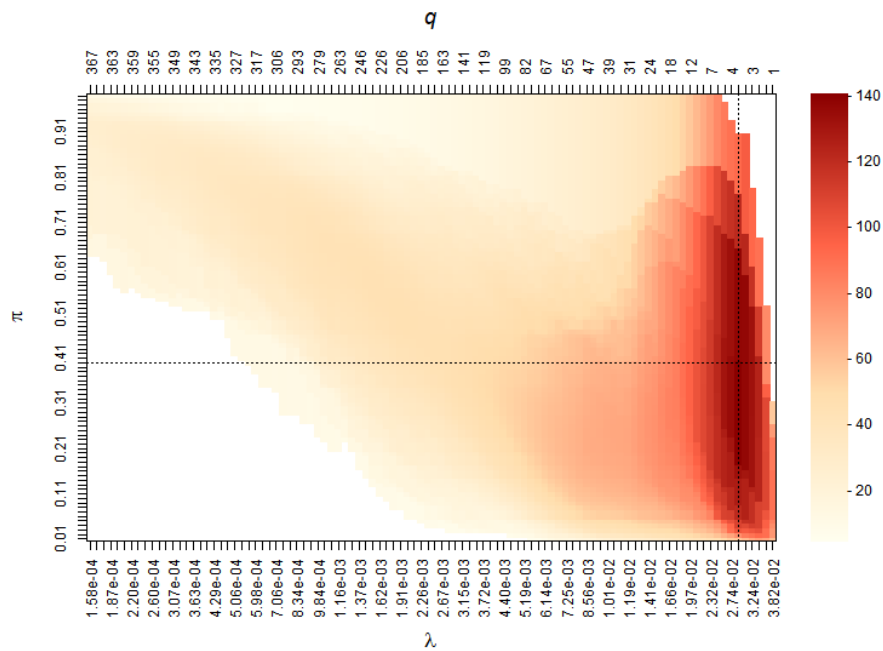


Figure 30: Calibration plot for  $M_{baseline}$  for prevalent diabetes.

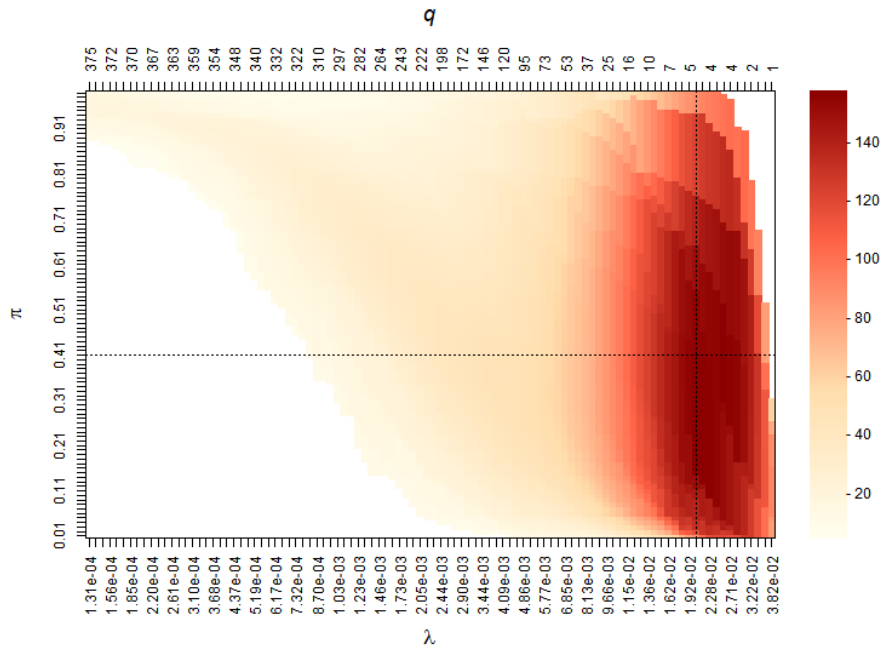


Figure 31: Calibration plot for  $M_{perturbed}$  for prevalent diabetes.

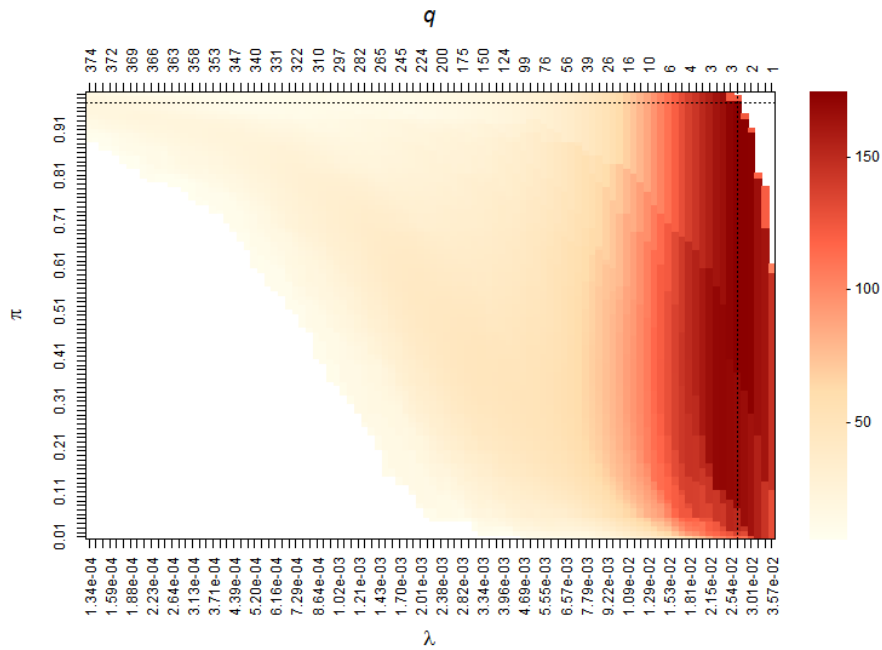


Figure 32: Calibration plot for  $M_{random}$  for prevalent diabetes.

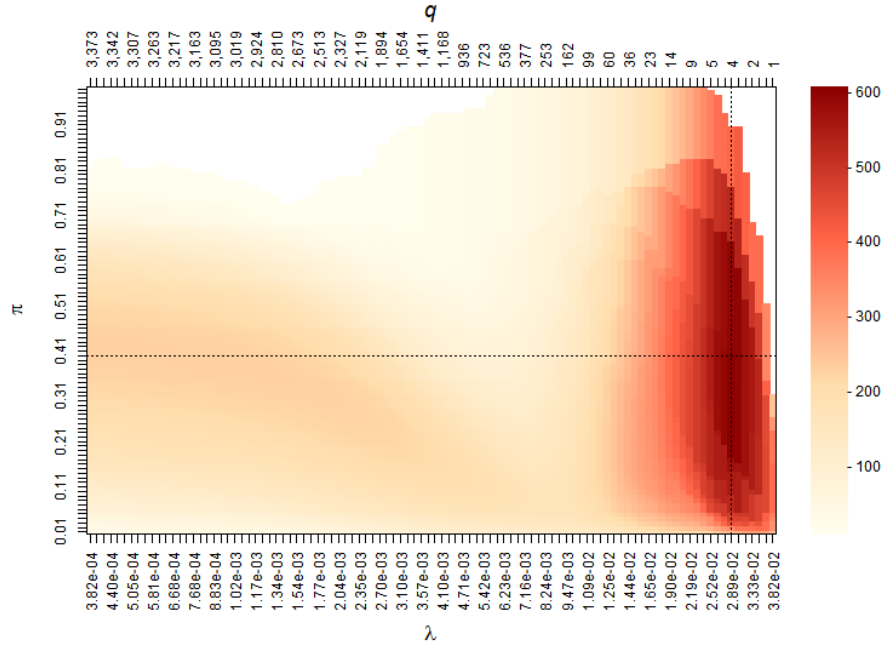


Figure 33: Calibration plot for  $M_{full}$  for prevalent diabetes.

The results continue to support the overall consistency of the model: both  $M_{baseline}$  and  $M_{full}$  select the same four proteins;  $M_{perturbed}$  selects six proteins, being the four identified by  $M_{baseline}$  plus two additional randomly selected proteins; finally,  $M_{random}$  selects only the two proteins from  $M_{baseline}$  that are included in its input set. As previously noted, the threshold has significantly less importance than with incident diabetes: not only do  $M_{baseline}$ ,  $M_{perturbed}$  and  $M_{full}$  have basically the same  $\pi^*$  (0.4, 0.41 and 0.4 again), but  $M_{random}$  manages to capture the  $M_{baseline}$  proteins even with a value of  $\pi^*$  of 0.97. Table 10 shows our final stable set of predictors.

	Selection probability	PWAS p-value
<i>Cartilage intermediatelaye protein 2 (CILP2)</i>	0.90	3.447954e-52
<i>Semaphorin-6A</i>	0.77	5.153841e-56
<i>Plexin-B2</i>	0.65	9.467723e-55
<i>Neurofascin (NFASC)</i>	0.41	8.499782e-45

Table 10: Stable set of proteins for prevalent diabetes.

Both LASSO models selected only one protein, *Semaphorin-6A*. This further reinforces the view of Stability Selection as an extension of LASSO, and highlights the importance of running a model multiple times in order to reliably assess the stability of selected predictors. Regarding the predictive power of the selected proteins, Table 11 shows the results.

	$n_{\text{proteins}}$	McFadden's $R^2$	McFadden's adjusted $R^2$
$M_{\text{baseline}}$	4	0.011	0.010
<b>LASSO</b>	1	0.005	0.005
<b>PWAS</b>	387	0.026	-0.037

Table 11: Predictive power comparison between Stability Selection and LASSO for prevalent diabetes.

It is worth noting that the predictive performance of the models, as measured by McFadden's  $R^2$ , is relatively low. This is likely due to the fact that we are modeling prevalent diabetes (that is, both the protein levels and the outcome are measured at the same time point). In this context, the models capture association rather than true prediction, which inherently limits predictive power. Nevertheless, Stability Selection still outperforms standard LASSO in terms of variable selection (and even predictive power), as it focuses on identifying stable and replicable predictors across subsamples. This makes SS particularly valuable in settings where predictive performance is limited but interpretability and robustness are crucial. Indeed, all the variables detected by Stability Selection are known in the literature as highly associated with prevalent (and even incident) diabetes, as shown in the association studies for *CILP2* [79], *Semaphorin-6A* [80], *Plexin-B2* and *Neurofascin* [81]. Note that *NFASC* and *PLXB2*, although known to be associated with incident diabetes, were not selected by Stability Selection in the analysis of incident diabetes. This highlights once again the distinction between association and causality: in this appendix, Stability Selection was applied in its "associative" form, whereas the main body of the thesis demonstrated its utility for identifying potential causal relationships.

## Bibliography

- [1] Lascar N, Brown J, Pattison H, Barnett A. H, Bailey CJ, and Bellary S. Type 2 diabetes in adolescents and young adults. *Lancet Diabetes Endocrinol*, 6:69–80, January 2018.
- [2] Schlienger JL. Complications du diabète de type 2 [type 2 diabetes complications]. *Presse Med*, 42:839–48, May 2013.
- [3] International Diabetes Federation. Idf diabetes atlas – 10th edition, 2024.
- [4] World Health Organization. Diabetes - who fact sheet, 2023.
- [5] Shahina Pardhan, Md. Saiful Islam, and Raju Sapkota. Knowledge, attitude, and diabetes self-care among individuals at high-risk of diabetes-related blindness in bangladesh: a cross-sectional study. *BMC Public Health*, 2024.
- [6] International Diabetes Federation. Bangladesh | south-east asia region, 2024.
- [7] Dhaka Tribune. Over 8 million people have diabetes in bangladesh, 2020.
- [8] Hossain Z, Khanam M, and Razzaque Sarker A. Out-of-pocket expenditure among patients with diabetes in bangladesh: A nation-wide population-based study. *Health Policy Open*, 13 September 2023.
- [9] Julia Carrasco Zanini, Maik Pietzner, and Claudia Langenberg. Integrating genetics and the plasma proteome to predict the risk of type 2 diabetes. *Current Diabetes Reports*, 20, 8 October 2020.
- [10] Shuai Yuan, Fengzhe Xu, Xue Li, Jie Chen, Jie Zheng, Christos S Mantzoros, and Susanna C Larsson. Plasma proteins and onset of type 2 diabetes and diabetic complications: Proteome-wide mendelian randomization and colocalization analyses. *Cell reports Medicine*, 4, 19 September 2023.
- [11] Valborg Gudmundsdottir, Shaza B Zaghlool, Valur Emilsson, Thor Aspelund, Marjan Ilkov, Elias F Gudmundsson, Stefan M Jonsson, Nuno R Zilhão, John R Lamb, Karsten Suhre, Lori L Jennings, and Vilmundur Gudnason. Circulating protein

- signatures and causal candidates for type 2 diabetes. *Diabetes*, 69:1843–1853, August 2020.
- [12] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.
- [13] Barbara Bodinier, Sarah Filippi, Therese Haugdahl Nøst, Julien Chiquet, and Marc Chadeau-Hyam. Automated calibration for stability selection in penalised regression and graphical models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2023. Published online: 13 July 2023.
- [14] Debbie A. Lawlor, Roger M. Harbord, Jonathan A. C. Sterne, Nic Timpson, and George Davey Smith. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Wiley InterScience*, 27:1133–1163, 20 September 2007.
- [15] <https://www.believestudy-bangladesh.org/>.
- [16] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 1:267–288, 1996.
- [17] Daniel McFadden. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, page 104–142, November 1972.
- [18] J. Scott Long. *Regression Models for categorical and limited dependent variables*. Sage Publications, 1997.
- [19] Jordan Louviere, David Hensher, and Joffre Swait. *Stated choice methods: analysis and application*, volume 17. 01 2000.
- [20] Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 01-09-2009.
- [21] Miguel A. Hernàn, Sonia Hernández-Díaz, and James M. Robins. A structural approach to selection bias. *Epidemiology*, 15(5), September 2004.
- [22] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK, 2nd edition, 2009.
- [23] Solène Cadiou and Rémy Slama. Instability of variable-selection algorithms used to identify true predictors of an outcome in intermediate-dimension epidemiologic studies. *Epidemiology*, 32:402–411, May 2021.
- [24] Luis H. Braga, Forough Farrokhyar, M Irfan Dönmez, Caleb P. Nelson, Bernhard Haid, Kathy Herbst, Massimo Garriboli, Salvatore Cascio, Anka Nieuwhof-Leppink,

- Martin Kaefer, Darius J. Bägli, Nicholas Kalfa, Christina Ching, Magdalena Fossum, and Luke Harper. Randomized controlled trials - the what, when, how and why. *Journal of Pediatric Urology*, 21:397–404, 2025.
- [25] Matthew S. Thiese. Observational and interventional study design types; an overview. *National Library of Medicine*, 24, 15 June 2014.
- [26] George Tegan. What is a cohort study? | definition & examples. <https://www.scribbr.com/methodology/cohort-study/>, 24 February 2023.
- [27] Kenneth J. Rothman, Sander Greenland, and Timothy L. Lash. *Modern Epidemiology*. Lippincott Williams & Wilkins, Philadelphia, 3rd edition, 2008.
- [28] David A. Grimes and Kenneth F. Schulz. Bias and causal associations in observational research. *The Lancet*, 359(9302):248–252, 2002.
- [29] John Concato, Nirav Shah, and Ralph I. Horwitz. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine*, 342(25):1887–1892, 2000.
- [30] Vanessa Didelez and Nuala Sheehan. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical methods in Medical research*, 16:309–330, August 2007.
- [31] Gregor Mendel. Experiments in plant hybridization. <http://www.mendelweb.org/archive/Mendel.Experiments.txt>, 1865.
- [32] Zixuan Wu and Jingshu Wang. Interpretation of two-sample mendelian randomization for binary exposures and outcome. <http://doi.org/10.1101/2024.06.09.598150>, June 2024. Preprint.
- [33] Mendelian Randomization Dictionary. One-sample mr. <https://mr-dictionary.mrcieu.ac.uk/term/one-sample/>.
- [34] Brandon L Pierce, Habibur Ahsan, and Tyler J VanderWeele. Power and instrument strength requirements for mendelian randomization studies using multiple genetic variants. *International Journal of Epidemiology*, 40:740–752, 2011.
- [35] Victoria Garfield, Antoine Salzman, Stephen Burgess, and Nish Chaturvedi. A guide for selection of genetic instruments in mendelian randomisation (mr) studies of type-2 diabetes and hba1c: towards an integrated approach. *Diabetes*, 72:175–183, 1 February 2023.
- [36] Jing Mao, Yanqiong Gan, Xinlin Tan, Yuhan He, Qiao Jing, and Qi Shi. A two-sample

- mendelian randomization study of basophil count and risk of gestational diabetes mellitus. *International Journal of Women's Health*, 17:517–527, 26 February 2025.
- [37] Douglas Staiger and James H. Stock. Instrumental variables regression with weak instruments. *Econometrica*, 65:557–586, May 1997.
- [38] Marcus R Munafò, Kate Tilling, and Yoav Ben-Shlomo. Smoking status and body mass index: a longitudinal study. *Nicotine tobacco research: official journal of the Society for Research on Nicotine and Tobacco*, 11:765–761, June 2009.
- [39] Carlo Emilio Bonferroni. *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, Firenze, Italy, 1936.
- [40] Joelle Mbatchou, Leland Barnard, Joshua Backman, Anthony Marcketta, Jack A. Kosmicki, Andrey Ziyatdinov, Christian Benner, Colm O'Dushlaine, Mathew Barber, Boris Boutkov, Lukas Habegger, Manuel Ferreira, Aris Baras, Jeffrey Reid, Goncalo Abecasis, Evan Maxwell, and Jonathan Marchini. Computationally efficient whole-genome regression for quantitative and binary traits. *nature genetics*, 53:1097–1103, 20 May 2021.
- [41] Sebastian Schönherr, Johanna F Schachtl-Riess, Silvia Di Maio, Marvin Mark Michele Filosi, Christian Fuchsberger Claudia Lamina, Florian Kronenberg, and Lukas Forer. Performing highly parallelized and reproducible gwas analysis on biobank-scale data. *NAR Genomics and Bioinformatics*, 6, 2024.
- [42] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38:904–909, 2006.
- [43] <https://www.ebi.ac.uk/gwas/studies/GCST90093109>.
- [44] Marie Loh, Weihua Zhang, Hong Kiat Ng, Katharina Schmid, Amel Lamri, Lin Tong, Meraj Ahmad, Jung-Jin Lee, Maggie C Y Ng, Lauren E Petty, Cassandra N Spracklen, Fumihiko Takeuchi, Md Tariqul Islam, Farzana Jasmine, Anuradhani Kasturiratne, Muhammad Kibriya, Karen L Mohlke, Guillaume Paré, Gauri Prasad, Mohammad Shahriar, Miao Ling Chee, H Janaka de Silva, James C Engert, Hertzfel C Gerstein, K Radha Mani, Charumathi Sabanayagam, Marijana Vujkovic, Ananda R Wickremasinghe, Tien Yin Wong, Chittaranjan S Yajnik, Salim Yusuf, Habibul Ahsan, Dwaipayan Bharadwaj, Sonia S Anand, Jennifer E Below, Michael Boehnke, Donald W Bowden, Giriraj R Chandak, Ching-Yu Cheng, Norihiro Kato, Anubha Mahajan, Xueling Sim, Mark I McCarthy, Andrew P Morris, Jaspal S Kooner, Dan-

- ish Saleheen, and John C Chambers. Identification of genetic effects underlying type 2 diabetes in south asian and european populations. *Communication biology*, 5, 7 April 2022.
- [45] Marie Loh, Weihua Zhang, Hong Kiat Ng, Katharina Schmid, Amel Lamri, Lin Tong, Meraj Ahmad, Jung-Jin Lee, Maggie C Y Ng, Lauren E Petty, Cassandra N Spracklen, Fumihiko Takeuchi, Md Tariqul Islam, Farzana Jasmine, Anuradhani Kasturiratne, Muhammad Kibriya, Karen L Mohlke, Guillaume Paré, Gauri Prasad, Mohammad Shahriar, Miao Ling Chee, H Janaka de Silva, James C Engert, Hertzfel C Gerstein, K Radha Mani, Charumathi Sabanayagam, Marijana Vujkovic, Ananda R Wickremasinghe, Tien Yin Wong, Chittaranjan S Yajnik, Salim Yusuf, Habibul Ahsan, Dwaipayan Bharadwaj, Sonia S Anand, Jennifer E Below, Michael Boehnke, Donald W Bowden, Giriraj R Chandak, Ching-Yu Cheng, Norihiro Kato, Anubha Mahajan, Xueling Sim, Mark I McCarthy, Andrew P Morris, Jaspal S Kooner, Danish Saleheen, and John C Chambers. Author correction: Identification of genetic effects underlying type 2 diabetes in south asian and european populations. *Communications biology*, 5, 5 May 2022.
- [46] Elin Grundberg, Kerrin S Small, Åsa K Hedman, Alexandra C Nica, Alfonso Buil, Sarah Keildson, Jordana T Bell, Tsun-Po Yang, Eshwar Meduri, Amy Barrett, James Nisbett, Magdalena Sekowska, Alicja Wilk, So-Youn Shin, Daniel Glass, Mary Travers, Josine L Min, Sue Ring, Karen Ho, Gudmar Thorleifsson, Augustine Kong, Unnur Thorsteindottir, Chrysanthi Ainali, Antigone S Dimas, Neelam Hassanali, Catherine Ingle, David Knowles, Maria Krestyaninova, Christopher E Lowe, Paola Di Meglio, Stephen B Montgomery, Leopold Parts, Simon Potter, Gabriela Surdulescu, Loukia Tsaprouni, Veronique Bataille Sophia Tsoka, Richard Durbin, Frank O Nestle, Stephen O’Rahilly, Nicole Soranzo, Cecilia M Lindgren, Krina T Zondervan, Kouros R Ahmadi, Eric E Schadt, Kari Stefansson, George Davey Smith, Mark I McCarthy, Panos Deloukas, Emmanouil T Dermitzakis, Tim D Spector, and Multiple Tissue Human Expression Resource (MuTHER) Consortium. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature genetics*, 44:1084–9, October 2012.
- [47] Xiaoling Zhang, Hincó J Gierman, Daniel Levy, Andrew Plump, Radu Dobrin, Harald HH Goring, Joanne E Curran, Matthew P Johnson, John Blangero, Stuart K Kim, Christopher J O’Donnell, and Valur Emilsson Andrew D Johnson. Synthesis of 53 tissue and cell line expression qtl datasets reveals master eqtls. *BMC Genomics*, 15, 27 June 2014.

- [48] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.
- [49] Barbara Bodinier, Sabrina Rodrigues, Maryam Karimi, Sarah Filippi, Julien Chiquet, and Marc Chadeau-Hyam. Stability selection and consensus clustering in R: The R package sharp. *Journal of Statistical Software*, 112(5):1–27, 2025.
- [50] G. Hemani, K. Tilling, and G. Davey Smith. Orienting the causal relationship between imprecisely measured traits using gwas summary data. *PLoS Genetics*, 13(11):e1007081, 2017.
- [51] Anna Birukov, Fabian Eichelmann, Olga Kuxhaus, Elli Polemiti, Andreas Fritsche, Janine Wirth, Heiner Boeing, Cornelia Weikert, and Matthias B Schulze. Opposing associations of nt-probnp with risks of diabetes and diabetes-related complications. *Diabetes Care*, 43, December 2020.
- [52] Hanieh Yaghootkar, Claudia Lamina, Robert A Scott, Zari Dastani, Marie-France Hivert, Liling L Warren, Alena Stancáková, Sarah G Buxbaum, Leo-Pekka Lyytikäinen, Peter Henneman, Ying Wu, Chloe Y Y Cheung, James S Pankow, Anne U Jackson, Stefan Gustafsson, Jing Hua Zhao, Christie M Ballantyne, Weijia Xie, Richard N Bergman, Michael Boehnke, Fatiha el Bouazzaoui, Francis S Collins, Sandra H Dunn, Josee Dupuis, Nita G Forouhi, Christopher Gillson, Jaeyoung Hong Andrew T Hattersley, Mika Kähönen, Johanna Kuusisto, Lyudmyla Kedenko, Florian Kronenberg, Alessandro Doria, Themistocles L Assimes, Ele Ferrannini, Torben Hansen, Ke Hao, Hans Häring, Joshua W Knowles, Cecilia M Lindgren, John J Nolan, Jussi Paananen, Oluf Pedersen, Thomas Quertermous, Ulf Smith, GENESIS Consortium, RISC Consortium, Terho Lehtimäki, Ching-Ti Liu, Ruth J F Loos, Mark I McCarthy, Andrew D Morris, Ramachandran S Vasani, Tim D Spector, Tanya M Teslovich, Jaakko Tuomilehto, Ko Willems van Dijk, Jorma S Viikari, Na Zhu, Claudia Langenberg, Erik Ingelsson, Robert K Semple, Alan R Sinaiko, Colin N A Palmer, Mark Walker, Karen S L Lam, Bernhard Paulweber, Karen L Mohlke, Cornelia van Duijn, Aurelian Bidulescu Olli T Raitakari, Nick J Wareham, Markku Laakso, Dawn M Waterworth, Debbie A Lawlor, James B Meigs, J Brent Richards, and Timothy M Frayling. Mendelian randomization studies do not support a causal role for reduced circulating adiponectin levels in insulin resistance and type 2 diabetes. *Diabetes*, 62, 1 October 2013.
- [53] Maria B. Nielsen, Yunus Çolak, Marianne Benn, and Børge G. Nordestgaard. Low plasma adiponectin in risk of type 2 diabetes: Observational analysis and one- and

- two-sample mendelian randomization analyses in 756,219 individuals. *Diabetes*, 70, November 2021.
- [54] Hajnalka Lőrincz, Sára Csiha, Balázs Ratku, Sándor Somodi, Ferenc Sztanek, and György Paragh and Mariann Harangi. Associations between serum kallistatin levels and markers of glucose homeostasis, inflammation, and lipoprotein metabolism in patients with type 2 diabetes and nondiabetic obesity. *International Journal of Molecular Sciences*, 25, 6 June 2024.
- [55] Kaspar Sørensen, Lise Aksglaede, Thor Munch-Andersen, Henrik Leffers Niels Jacob Aachmann-Andersen, Jørn Wulff Helge, Linda Hilsted, and Anders Juul. Impact of the growth hormone receptor exon 3 deletion gene polymorphism on glucose metabolism, lipids, and insulin-like growth factor-i levels during puberty. *The Journal of Clinical Endocrinology metabolism*, 94, 1 August 2009.
- [56] Zhilian Li, Yan Li, Jessica M Overstreet, Sungjin Chung, Aolei Niu, Xiaofeng Fan, Suwan Wang, Yinqiu Wang, Ming-Zhi Zhang, and Raymond C Harris. Inhibition of epidermal growth factor receptor activation is associated with improved diabetic nephropathy and insulin resistance in type 2 diabetes. *Diabetes*, 67, September 2018.
- [57] Barbara Kollerits, Claudia Lamina, Cornelia Huth, Pedro Marques-Vidal, Stefan Kiechl, Ilkka Seppälä, Jackie Cooper, Steven C Hunt, Christa Meisinger, Christian Herder, Ludmilla Kedenko, Johann Willeit, Barbara Thorand, Doreen Dähnhardt, Doris Stöckl, Karin Willeit, Michael Roden, Wolfgang Rathmann, Bernhard Paulweber, Annette Peters, Mika Kähönen, Terho Lehtimäki, Olli T Raitakari, Steve E Humphries, Peter Vollenweider, Hans Dieplinger, and Florian Kronenberg. Plasma concentrations of afamin are associated with prevalent and incident type 2 diabetes: A pooled analysis in more than 20,000 individuals. *Diabetes Care*, 40:1386–1393, 12 September 2017.
- [58] Clemens Wittenbecher, Meriem Ouni, Olga Kuxhaus, Markus Jähnert, Pascal Gottmann, Andrea Teichmann, Karina Meidtner, Jennifer Kriebel, Harald Grallert, Tobias Pischon, Heiner Boeing, Matthias B. Schulze, and Annette Schürmann. Insulin-like growth factor binding protein 2 (igfbp-2) and the risk of developing type 2 diabetes. *Diabetes*, 68(1):188–197, 5 November 2018.
- [59] Tatsuya Yano, Zhengyu Liu, Jennifer Donovan, Melissa K. Thomas, and Joel F. Habener. Stromal cell-derived factor-1 (sdf-1)/cxcl12 attenuates diabetes in mice and promotes pancreatic  $\beta$ -cell survival by activation of the prosurvival kinase akt. *Diabetes*, 56(12):2946–2957, 1 December 2007.

- [60] Xiang-Yu Chen, Ying-Xin Shi, Ya-Ping Huang, Min Ding, Qi ling Shen, Chun-Jun Li, and Jing-Na Lin. Sdf-1 inhibits the dedifferentiation of islet cells in hyperglycaemia by up-regulating foxo1 via binding to cxcr4. *Journal of Cellular and Molecular Medicine*, 26, 21 December 2021.
- [61] Naoko Nagai, Hiroko Habuchi, Noriko Sugaya, Masao Nakamura, Toru Imamura, Hideto Watanabe, and Koji Kimata. Involvement of heparan sulfate 6-o-sulfation in the regulation of energy metabolism and the alteration of thyroid hormone levels in male mice. *Glycobiology*, 23(8):980–992, May 2013.
- [62] Kamynina E and Stover PJ. The roles of sumo in metabolic regulation. *Advances in Experimental Medicine and Biology*, 963:143–168, 2017.
- [63] Sato H, Taketomi Y, and Murakami M. Metabolic regulation by secreted phospholipase a2. *Inflammation and Regeneration*, 36, 21 May 2016.
- [64] Nikolajczyk BS, Jagannathan-Bogdan M, Shin H, and Gyurko R. State of the union between metabolism and the immune system in type 2 diabetes. *Genes and Immunity*, 12:239–50, June 2011.
- [65] J Acosta, J Hettinga, R Flückiger, N Krumrei, A Goldfine, L Angarita, and J Halperin. Molecular basis for a link between complement and the vascular complications of diabetes. *Proceedings of the National Academy of Sciences for the United States of America*, 97, 9 May 2000.
- [66] Jiayong Li, Yuncheng Ma, Lingling Xie, Kaichen Zhuo, Yuxian He, Xin Ma, Shufen Zheng, Shicheng Guo, Yizhen Tang, Guzainuer Muhetaer, Mireayi Aizezi, Dan Zhang, Aizezi Wumaier, Chao Tang Xu Zhang 7, Wei Wang, Wenyong Huang, and Xinbo Gao. Comprehensive proteomic profiling of exfoliation glaucoma via mass spectrometry reveals svep1 as a potential biomarker. *Investigative Ophthalmology and Visual Science*, 66, 7 March 2025.
- [67] Roman Pfister, Stephen Sharp, Robert Luben, Paul Welsh, Inês Barroso, Veikko Salomaa, Aline Meirhaeghe, Kay-Tee Khaw, Naveed Sattar, Claudia Langenberg, and Nicholas J Wareham. Mendelian randomization study of b-type natriuretic peptide and type 2 diabetes: Evidence of causal association from population studies. *PLoS Medicine*, 8, 25 October 2011.
- [68] Shuai Zhou, Peiwen Zhou, Tianshi Yang, Junzhuo Si, Wenyan An, and Yanfang Jiang. Glucosamine supplementation contributes to reducing the risk of type 2 diabetes: Evidence from mendelian randomization combined with a meta-analysis. *The Journal of international medical research*, 53, April 2025.

- [69] Yue-Yang Zhang, Bing-Xue Chen, and Qin Wan. Association of lipid-lowering drugs with the risk of type 2 diabetes and its complications: a mendelian randomized study. *Diabetology & Metabolic Syndrome*, 16, 2024.
- [70] Marie Pigeyre, Jennifer Sjaarda, Michael Chong, Sibylle Hess, Jackie Bosch, Salim Yusuf, Hertzell Gerstein, and Guillaume Paré. Ace and type 2 diabetes risk: A mendelian randomization study. *Diabetes Care*, 43:835–842, April 2020.
- [71] Yang Li, Yahu Miao, Qing Feng, Weixi Zhu, Yijing Chen, Qingqing Kang, Zhen Wang, Fangting Lu, and Qiu Zhang. Mitochondrial dysfunction and onset of type 2 diabetes along with its complications: a multi-omics mendelian randomization and colocalization study. *frontiers in Endocrinology*, 15, 30 August 2024.
- [72] Lawien Al Ali, Yordi J van de Vegte, M Abdullah Said, Hilde E Groot, Tom Hendriks, Ming Wai Yeung, Erik Lipsic, and Pim van der Harst. Fetuin-a and its genetic association with cardiometabolic disease. *Scientific reports*, 13, 6 December 2023.
- [73] Alicia G Gómez-Valadés, Anna Vidal-Alabré, Maria Molas, Jordi Boada, Jordi Bermúdez, Ramon Bartrons, and José C Perales. Overcoming diabetes-induced hyperglycemia through inhibition of hepatic phosphoenolpyruvate carboxykinase (gtp) with rnaï. *Molecular therapy: the Journal of the American Society of Gene Therapy*, 13:401–10, February 2006.
- [74] LeBris S Quinn and Barbara G Anderson. Interleukin-15, il-15 receptor-alpha, and obesity: Concordance of laboratory animal and human genetic studies. *Journal of Obesity*, 29 March 2011.
- [75] Amy Vora, James A de Lemos, Colby Ayers, Justin L Grodin, and Ildiko Lingvay. Association of galectin-3 with diabetes mellitus in the dallas heart study. *The Journal of clinical endocrinology and metabolism*, 104:4449–4458, 1 October 2019.
- [76] Ming Yi, Xingrong Feng, Qiuyue Guan, Yin Liu, Yunqiang Liu, and Zhiguang Su. A multi-omics mendelian randomization study reveals pam as a potential therapeutic target for type 2 diabetes. *Journal of Transational Medicine*, 23, 8 October 2025.
- [77] Romana Stark, Jack Feehan, Aya Mousa, Zane B Andrews, and Barbora de Courten. Liver-expressed antimicrobial peptide 2 is associated with improved pancreatic insulin secretion in adults with overweight and obesity. *Diabetes, obesity & metabolism*, 25:1213–1220, May 2023.
- [78] U Petersson, C J Ostgren, L Brudin, K Brismar, and P M Nilsson. Low levels of insulin-like growth-factor-binding protein-1 (igfbp-1) are prospectively associated

with the incidence of type 2 diabetes and impaired glucose tolerance (igt): the söderåkra cardiovascular risk factor study. *Diabetes & metabolism*, 35:198–205, June 2009.

- [79] Tong Wu, Qin Zhang, Shaobo Wu, Wenjing Hu, Tingting Zhou, Ke Li, Dongfang Liu, Harvest F Gu, Hongting Zheng, Zhiming Zhu, Ling Li, and Gangyi Yang. Cilp-2 is a novel secreted protein and associated with insulin resistance. *Journal of molecular cell biology*, 11(12):1083–1094, 19 December 2019.
- [80] Qiongyu Lu and Li Zhu. The role of semaphorins in metabolic disorders. *International Journal of Molecular Sciences*, 6 August 2020.
- [81] Helene T. Cronje, Michael Y. Mi, Thomas R. Austin, Mary L. Biggs, David S. Siscovick, Rozenn N. Lemaitre, Bruce M. Psaty, Russell P. Tracy, Luc Djousse, Jorge R. Kizer, Joachim H. Ix, Prashant Rao, Jeremy M. Robbins, Jacob L. Barber, Mark A. Sarzynski, Clary B. Clish, Claude Bouchard, Kenneth J. Mukamal, Robert E. Gerszten, and Majken K. Jensen. Plasma proteomic risk markers of incident type 2 diabetes reflect physiologically distinct components of glucose-insulin homeostasis. *diabetes*, 72(5):666–673, 2023.

## List of Figures

1	Example of a calibration plot for a generic Stability Selection model (source: Bodinier <i>et al.</i> [13]). . . . .	8
2	A causal DAG in its simplest form (source: Hernàn <i>et al.</i> [21]). . . . .	10
3	A causal DAG with a common cause (source: Hernàn <i>et al.</i> [21]). . . . .	11
4	A causal DAG with a common effect (source: Hernàn <i>et al.</i> [21]). . . . .	11
5	The Randomized Control Trial scheme (source: Braga <i>et al.</i> [24]). . . . .	13
6	Comparison between Mendelian Randomization and Randomized Control Trials. . . . .	15
7	Causal DAG representing the definition of the IV Z (source: Lawlor <i>et al.</i> [14]). . . . .	16
8	Underlying scheme of MR with binary exposure $X$ and outcome $Y$ (source: Wu and Wang [32]). . . . .	17
9	Patient recruitment procedure for the BELIEVE study (source: the BELIEVE study <a href="https://www.believestudy-bangladesh.org/">https://www.believestudy-bangladesh.org/</a> )[15]. . . . .	21
10	Bagplot of age vs BMI. . . . .	24
11	Distribution of the <i>diadstat</i> variable. . . . .	25
12	Boxplot of <i>BMI</i> vs <i>smokstat</i> . . . . .	26
13	Manhattan plot. . . . .	28
14	Pie charts for the selected number of proteins throughout the 100 runs for each of the two LASSO models. . . . .	35
15	Calibration plot for $M_{baseline}$ . . . . .	38
16	Calibration plot for $M_{perturbed}$ . . . . .	38
17	Calibration plot for $M_{random}$ . . . . .	39
18	Calibration plot for $M_{full}$ . . . . .	39
19	The distribution of $\pi^*$ for $M_{random}$ across the 100 runs. . . . .	41
20	Distribution of the number of selected proteins for $M_{perturbed}$ . . . . .	41
21	The distribution of $\pi^*$ for $M_{perturbed}$ across the 100 runs. . . . .	42

22	Forest plot for the estimated causal effect of all SS proteins in one sample MR. . . . .	53
23	Forest plot for the estimated causal effect for the 11 SS proteins included in the two-sample MR analysis. . . . .	54
24	Forest plot including all proteins selected by one-sample MR. . . . .	58
25	Forest plot including all proteins selected by two-sample MR. . . . .	59
26	Bagplot for outlier identification for <i>BMI</i> vs <i>age</i> in the prevalent diabetes dataset. . . . .	66
27	Distribution of the <i>diadstat</i> variable in the prevalent diabetes dataset. . . . .	67
28	Boxplot of <i>BMI</i> by <i>smokstat</i> category in the prevalent diabetes dataset. . . . .	68
29	Manhattan plot for prevalent diabetes. . . . .	69
30	Calibration plot for $M_{baseline}$ for prevalent diabetes. . . . .	69
31	Calibration plot for $M_{perturbed}$ for prevalent diabetes. . . . .	70
32	Calibration plot for $M_{random}$ for prevalent diabetes. . . . .	70
33	Calibration plot for $M_{full}$ for prevalent diabetes. . . . .	71

## List of Tables

1	Comparison between one sample and two sample MR. . . . .	18
2	Frequency table of patients. . . . .	23
3	<i>smokstat</i> categories distribution. . . . .	25
4	Summary of the models' main features. . . . .	37
5	The stable set of proteins as detected by Stability Selection. . . . .	45
6	Predictive power comparison between Stability Selection and LASSO. . . . .	46
7	Literature review to assess the possible causal role of the SS proteins. . . . .	49
8	MR results with and without Bonferroni correction. . . . .	57
9	Final comparison of the methods. . . . .	62
10	Stable set of proteins for prevalent diabetes. . . . .	71
11	Predictive power comparison between Stability Selection and LASSO for prevalent diabetes. . . . .	72



## Ringraziamenti

Desidero ringraziare innanzitutto la professoressa Francesca Ieva, che mi ha permesso di svolgere questa tesi, e le mie correlatrici, Solène Cadiou e Nicole Fontana, per l'aiuto fornito durante tutto il percorso, per i preziosi consigli e per il supporto continuo. Insieme a loro ringrazio tutto il personale di Human Technopole (in particolar modo Deborah Zani e Giulia Pontali) per l'assistenza e per l'accoglienza calorosa.

Ringrazio di cuore tutta la mia famiglia. Grazie alla mia mamma, la mia roccia, che mi ha sostenuto per tutti questi anni e che mi ha sempre compreso nel profondo senza neanche bisogno di chiedermi come andasse. Grazie a papà, che anche con poche parole e qualche sguardo riesce a farmi capire che è fiero di me qualunque cosa io faccia. Grazie alle mie zie e con loro a tutto il resto della famiglia per avermi circondato sempre di un affetto immenso. Grazie alla mia nonna, che mi ha sempre infuso e continuerà ad infondermi calma, serenità e saggezza, e che ora vorrei che fosse qui a vedere il suo *picciriddu*. Mi manchi.

Ringrazio i miei amici che riescono sempre a farmi svagare quando ne ho più bisogno. Ringrazio Alessandro, Carlo, Alessandro ed Elisa: con voi mi sento sempre a casa e al sicuro. Ringrazio Simone, che mi ha insegnato che quando c'è una vera amicizia la distanza non conta. Infine, ringrazio Enrica, la migliore amica che potessi desiderare e grazie a cui ho imparato a vedere le cose con un po' più di ottimismo e col sorriso.

Grazie a tutti, avete reso questo percorso speciale.