# Biology-guided algorithms

*Improved cardiovascular risk prediction and biomarker discovery*

Belo Pereira, J.P.

**Publication date**
2022

**Document Version**
Final published version

Link to publication

**Citation for published version (APA):**
Belo Pereira, J. P. (2022). *Biology-guided algorithms: Improved cardiovascular risk prediction and biomarker discovery*. [Thesis, fully internal, Universiteit van Amsterdam].

# Biology-guided Algorithms

## Improved Cardiovascular Risk Prediction and Biomarker Discovery

João Pereira

# Biology-guided algorithms Improved cardiovascular risk prediction and biomarker discovery

João Pedro Belo Pereira

Colophon
Biology-guided algorithms Improved cardiovascular risk prediction and biomarker discovery
João Pedro Belo Pereira
ISBN/EAN: 978-90-9036615-9

The investigations were supported by the European Research Area Network on Cardiovascular Diseases (ERA-CVD) grant (ERA-CVD JTC2017) and HorAIzon B.V.

Cover design by Cláudia Pinhão.
Printing: Print & Bind — https://www.printenbind.nl/.

# Biology-guided algorithms

## Improved cardiovascular risk prediction and biomarker discovery

## ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Aula der Universiteit
op woensdag 12 oktober 2022, te 11.00 uur

door João Pedro Belo Pereira
geboren te Entroncamento

***Promotiecommissie***

| | | |
|---|---|---|
| *Promotor:* | prof. dr. E.S.G. Stroes | AMC-UvA |
| *Copromotor:* | dr. E. Levin | AMC-UvA |
| *Overige leden:* | prof. dr. N.A.W. van Riel | Technische Universiteit Eindhoven |
| | prof. dr. A.H. Zwinderman | AMC-UvA |
| | prof. dr. W.J. de Jonge | AMC-UvA |
| | dr. C.J. Veenman | TNO/Leiden University |
| | dr. H.J. Herrema | AMC-UvA |
| | prof. dr. P.A.N. Bosman | CWI |

Faculteit der Geneeskunde

# Contents

## I Applied Machine Learning in Clinical Research 9

# 1

# Introduction

## 1.1 General Introduction

Cardiovascular disease (CVD) is the number one cause of death globally (WHO). Identification of cardiovascular risk remains a major challenge both in primary and secondary event prevention causing immense burden in Western societies. Despite well known causal risk factors such as blood apolipoprotein-B-containing lipoproteins, high blood pressure, cigarette smoking, and Diabetes Mellitus Members: et al. (2022), the clinically used algorithms such as the Framingham risk score, pooled cohort equations and Systemic Coronary Risk Evaluation (SCORE) suffer from limited event prediction accuracy Fernández-Friera et al. (2017). Presumably, this limitation stems from the narrow pathophysiological view these risk factors offer, while CVD risk is the result of complex interactions between comorbidities and exogeneous risk factors Hoogeveen et al. (2018). Targeted proteomics provide a promising risk prediction alternative since they capture a snapshot of the current individual physiology reflecting the genetic background but also lifestyle and metabolic pathways Williams et al. (2019). Unlike traditional statistical methods which help understand relationships between a limited number of variables, the use of high-throughput technology such as proteomics requires the use of more flexible and scalable multivariate methods that Machine Learning (ML) offers.

To further improve the prediction, combining proteomics with other classes of biomarkers such as genomics, transcriptomics and phenotypic markers provides a systems overview of the disease progression thereby encompassing a larger set of etiologies and population generalization. Despite the critical role better prediction accuracy plays on mitigating the impact of CVD, there are several ways in which ML can advance Medical/Biological research. In this thesis, we will present novel biology-guided algorithms to tackle domain knowledge integration, multi-domain learning and unbiased feature importance.

## 1.2 Machine Learning Preliminaries

**Notation**

We denote matrices, 1-dimensional arrays, and scalars/functions with capital bold, bold, and regular text, respectively (e.g. $\mathbf{X}$, $\mathbf{x}$, $\alpha/f$). Given a dataset $\mathbf{X}_{M\times N}$, we will denote the set of all random variables by $\mathcal{X}$, its individual random variables by capital regular text with a subscript and the values using lowercase (e.g. $X_i$ and $x_i$), while the joint density/mass will be represented as $p(x)$. We will refer to the input random variables as features or variables interchangeably. The indicator function will be denoted by $\mathbb{I}(\cdot)$ and its argument will be expressed as a boolean evaluation such as $\mathbb{I}(x=y)$ mapping to 1 if $x=y$ and to 0 otherwise.

### 1.2.1 Supervised Learning

In supervised tasks, a dataset $\mathbf{X}_{M\times N}$ consisting of $M$ measurements for $N$ different variables like glucose, blood pressure and so on is used to predict an outcome of interest $\mathbf{y}$ such as diabetes. The goal of the model is to reduce the empirical risk given by :

$$\mathcal{L}(\mathbf{X}, \mathbf{y}, f_{\boldsymbol{\theta}}) \equiv \frac{1}{M}\sum_{i=1}^{M} l\left(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i\right), \tag{1.1}$$

where $f_{\boldsymbol{\theta}}$ is a model parameterized by $\boldsymbol{\theta} = [\theta_1, ..., \theta_k]$ and $l$ is a loss function designed to measure how different the prediction of the model and the output $y_i$ is. Note that the model is merely trying to approximate the true mapping function from the input random variables to the output $f_{\text{true}} : \mathcal{X} \rightarrow \mathcal{Y}$. When the outcome variable is a continuous variable, then this task is called *regression* and when it is discrete it is called *classification*. Common choices for loss functions include (ommiting the function arguments for compactness):

- Mean squared error (Regression): $l = (f_{\boldsymbol{\theta}}(\mathbf{x}_i) - y_i)^2$
- Mean absolute error (Regression): $l = |f_{\boldsymbol{\theta}}(\mathbf{x}_i) - y_i|$
- Accuracy (Classification): $l = \mathbb{I}\,(f_{\boldsymbol{\theta}}(\mathbf{x}_i) = y_i)$
- Cross-entropy (Classification): $l = -y_i log\, f_{\boldsymbol{\theta}}(\mathbf{x}_i) + (1 - y_i) log\,(1 - f_{\boldsymbol{\theta}}(\mathbf{x}_i))$

### 1.2.2 Parameter optimization

The model parameters are optimized by minimizing equation 1.1:

$$\boldsymbol{\theta}^* = \text{argmin}_{\boldsymbol{\theta}}\, \mathcal{L}(\mathbf{X}, \mathbf{y}, f_{\boldsymbol{\theta}}), \tag{1.2}$$

producing a single parameter estimate and the prediction for a new example $\mathbf{x}_{\text{new}}$ is then given by $f_{\boldsymbol{\theta}^*}(\mathbf{x}_{\text{new}})$. In contrast, Bayesian methods estimate the parameters' uncertainty by performing inference over the parameters' posterior distribution using the Bayes rule:

$$p(\boldsymbol{\theta}|\mathbf{X}, f) = \frac{p(\mathbf{X}|\boldsymbol{\theta}, f)p(\boldsymbol{\theta}|f)}{\int p(\mathbf{X}|\boldsymbol{\theta}, f)p(\boldsymbol{\theta}|f)d\boldsymbol{\theta}}, \tag{1.3}$$

and the uncertainty can be incorporated into the prediction by marginalizing out the parameters:

$$p(\mathbf{x}_{\text{new}}|\mathbf{X}) = \int_{\boldsymbol{\theta}} p(\mathbf{x}_{\text{new}}, \boldsymbol{\theta}|\mathbf{X})d\boldsymbol{\theta} = \int_{\boldsymbol{\theta}} p(\mathbf{x}_{\text{new}}|\boldsymbol{\theta}, \mathbf{X})p(\boldsymbol{\theta}|\mathbf{X})d\boldsymbol{\theta}, \tag{1.4}$$

which assuming future observations are conditionally independent given $\boldsymbol{\theta}$ becomes: $\int_{\boldsymbol{\theta}} p(\mathbf{x}_{\text{new}}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{X})d\boldsymbol{\theta}$. This is known as posterior predictive distribution. Since the denominator in equation 1.3 is generally intractable, one usually resorts to asymptotically correct sampling methods like MCMC, or approximations methods like variational inference. Although knowing the uncertainty around the parameters' estimation is desirable, inference is typically very computationally expensive which in practice becomes prohibitive for larger datasets.

**Bias-variance trade-off**

Suppose the true function mapping the input variables to the outcome can be described as $y_i = f_{\text{true}}(\mathbf{x}_i) + \epsilon$, with $\epsilon$ representing noise with zero mean and variance $\sigma^2$. Since $f_{\text{true}}$ is deterministic, $E_{\sim\mathcal{D}}[f_{\text{true}}(x)] = f_{\text{true}}(x)$. The expected model mean squared error is given by:

$$\begin{aligned} E_{\sim\mathcal{D}}\left[(y - f_{\boldsymbol{\theta}}(x))^2\right] &= E_{\sim\mathcal{D}}\left[(f_{\text{true}}(x) + \epsilon - f_{\boldsymbol{\theta}}(x))^2\right] \\ &= \underbrace{(f_{\text{true}}(x) - E_{\sim\mathcal{D}}[f_{\boldsymbol{\theta}}(x)])}_{\text{Bias}} + \sigma^2 + \underbrace{Var(f_{\boldsymbol{\theta}}(x))}_{\text{Variance}}. \end{aligned} \tag{1.5}$$

This is an important result in Machine Learning, called the *Bias-Variance trade-off.* Equation 1.5 expresses supervised learning as a balance between two sources of error: inadequate assumptions of the model (bias) and sensitivity to small perturbations in the input, meaning the model may be modelling random noise in the training data (variance). Variance is especially problematic in the presence of many uninformative features as is the case in biological datasets.

To mitigate the model variance, it is common to apply *regularization* techniques for example by adding a *penalization* term to the objective function 1.2:

$$f^* = \text{argmin}_f\, \mathcal{L}(\mathbf{X}, \mathbf{y}, f_{\boldsymbol{\theta}}) + \lambda R(f), \tag{1.6}$$

where $R(f)$ is typically a function of model complexity (for example the number of features the model uses for prediction) and $\lambda$ controls the extent of this penalization. In Bayesian methods, this regularization can be imposed using narrow priors in eq. 1.3.

Another way of reducing variance which we will make extensive use of in this thesis is by using ensemble methods. In ensemble methods like *random forests* Breiman (2001), several models are trained on random subsets of the data and their individual predictions are aggregated to form a final prediction. Similarly, in gradient boosting several "weak" (i.e. low bias) models $h_{\boldsymbol{\theta}}$ are combined to make a final prediction: $f_{\gamma,\boldsymbol{\theta}} = \sum_{i=1}^{K} \gamma_i h_i(x)$, but these are trained sequentially by performing gradient descent on the objective function 1.1 reducing both bias and variance.

### 1.2.3 Model evaluation

After estimating the optimal model parameter using equation 1.1, the model should be tested on an independent dataset to prevent "data leakage", that is, preventing fictitiously inflating the models' performance by measuring how well it performs on data it has "seen" before. Thus, to estimate the model's true performance using the dataset at hand, consider a splitter function to divide the dataset into a training, validation and testing set $s_i(\mathbf{X}) = \left\{\boldsymbol{\pi}^{(i)}_{\text{train}}, \boldsymbol{\pi}^{(i)}_{\text{val}}, \boldsymbol{\pi}^{(i)}_{\text{test}}\right\}$, $i \in \{1, .., K\}$, where $\boldsymbol{\pi}$ are permutations of the set $\{1, 2, ..., M\}$ and for all $i$:

$$p\left(\mathbf{X}_{\boldsymbol{\pi}^{(i)}_{\text{train}}}, \mathbf{X}_{\boldsymbol{\pi}^{(i)}_{\text{val}}}, \mathbf{X}_{\boldsymbol{\pi}^{(i)}_{\text{test}}}\right) = p\left(\mathbf{X}_{\boldsymbol{\pi}^{(i)}_{\text{train}}}\right) p\left(\mathbf{X}_{\boldsymbol{\pi}^{(i)}_{\text{val}}}\right) p\left(\mathbf{X}_{\boldsymbol{\pi}^{(i)}_{\text{test}}}\right) \tag{1.7}$$

$$\boldsymbol{\pi}^{(i)}_{\text{train}} \cap \boldsymbol{\pi}^{(i)}_{\text{val}} = \boldsymbol{\pi}^{(i)}_{\text{train}} \cap \boldsymbol{\pi}^{(i)}_{\text{test}} = \boldsymbol{\pi}^{(i)}_{\text{val}} \cap \boldsymbol{\pi}^{(i)}_{\text{test}} = \emptyset. \tag{1.8}$$

Using the splitter function $s_i(\mathbf{X})$, the model performance estimator is then commonly chosen as:

$$\hat{\mathcal{P}}(\mathcal{X}, \mathcal{Y}, f_{\boldsymbol{\theta}}) \equiv \frac{1}{K}\sum_{i=1}^{K} \mathcal{P}\left(\mathbf{X}_{\boldsymbol{\pi}^{(i)}_{\text{test}}}, \mathbf{y}_{\boldsymbol{\pi}^{(i)}_{\text{test}}}, f^{(i)}_{\boldsymbol{\theta}^*}, l\right), \tag{1.9}$$

where $f^{(i)}_{\boldsymbol{\theta}^*}$ is the model trained on $\left\{\mathbf{X}_{\boldsymbol{\pi}^{(i)}_{\text{train}}}, \mathbf{y}_{\boldsymbol{\pi}^{(i)}_{\text{train}}}\right\}$ and optimized on $\left\{\mathbf{X}_{\boldsymbol{\pi}^{(i)}_{\text{val}}}, \mathbf{y}_{\boldsymbol{\pi}^{(i)}_{\text{val}}}\right\}$ using the loss function $l$, and $\mathcal{P}$ is a performance metric not necessarily equal to $l$. This process is called cross-validation and measures how well the model will generalize to an independent dataset.

When data is assumed to be independent and identically distributed (i.i.d.), then splitting the data indices into different non-overlapping "blocks" is sufficient to satisfy 1.7. However, suppose this assumption is not met because measurements for the same patient are taken at different time points, then the splitter function $s_i(\mathbf{X})$ should be chosen such that it does not place the same patient in different sets.

### 1.2.4 Kernel methods

There is a class of algorithms whose output can be cast as a function of the inner product with instances in the data: $f_{\boldsymbol{\theta}}(\mathbf{x}_i) = g_{\boldsymbol{\theta}}(\{\mathbf{x}_j^T\mathbf{x}_i\})$ with $\mathbf{x}_j \in \mathcal{D}$. The inner product here can be seen as a measure of similarity between $\mathbf{x}_j$ and $\mathbf{x}_i$, but it may be the case that two points are very similar in the input space $\mathcal{X}$ despite having a large difference in the output space (or belonging to different classes): $y_j \neq y_i$, $\mathbf{x}_j^T\mathbf{x}_i \leq \delta$. A powerful extension is to use a map $\phi : \mathcal{X} \to \mathcal{X}'$ to project the data into a higher dimensional space where the data dissimilarity becomes more clear, that is, to use $f_{\boldsymbol{\theta}}(\mathbf{x}_i) = g_{\boldsymbol{\theta}}\left(\{\phi(\mathbf{x}_j)^T\phi(\mathbf{x}_i)\}\right)$ with $\mathbf{x}_j \in \mathcal{D}$.

We can define a function designed to measure the similarity between two points called a kernel $k(\mathbf{x}_j, \mathbf{x}_i)$, and if this function is positive definite, then the map $\phi$ exists and the kernel is given by $k(\mathbf{x}_j, \mathbf{x}_i) = \phi(\mathbf{x}_j)^T \phi(\mathbf{x}_i)$. The significance of this result is that a kernel can be specified to measure similarity in a higher dimensional space, and the points are projected to this space without the need to explicitly compute the projection. In the case of the gaussian kernel: $k(\mathbf{x}_j, \mathbf{x}_i) = e^{\frac{(\mathbf{x}_i - \mathbf{x}_j)^2}{2\sigma^2}}$, it can be shown that the data is mapped into an infinite dimensional space.

## 1.3 Thesis Outline

### 1.3.1 Part I: Applied Machine Learning in Clinical Research

In the first part of this thesis, we demonstrate how Machine Learning can be used to improve CVD prediction compared to the clinically used algorithms. In chapter 2, we use tree-based gradient boosting techniques trained on targeted proteomics to predict primary events which resulted in improved accuracy compared to the clinical risk model. We then extend these results to a secondary prevention setting in chapter 3. Finally, in chapter 4 we adopt a systems biology approach to heart failure prediction by using a model trained on clinical phenotypic markers, proteomics, transcriptomics and genetic data and explore the implied biological pathophysiology.
Over the course of these projects, we considered other ways in which we could use ML to advance medical research. This endeavor resulted in novel algorithms which we discuss in part II of this thesis.

### 1.3.2 Part II: Novel Machine Learning Algorithms for Clinical Research

#### Domain Knowledge Integration

Following the extensive use of proteomics in the clinical projects, we decided to develop a way to incorporate domain knowledge on protein-protein interactions to further improve prediction. Because medical data is often short on individual samples and large in random variables $M << N$, introducing prior knowledge into the learning pipeline to constrain the learning problem can be beneficial for many classes of algorithms. Thus, in chapter 5 we present a novel graph kernel (see section 1.2.4) to incorporate protein-protein interactions and demonstrate its superior performance compared to the same algorithm trained on the dataset alone.

### Multi-Domain learning

In chapters 2 and 3 we show the added benefit of using targeted plasma proteomics for event prediction, while in chapter 4 we show further improvement by taking a multi-domain approach. To circumvent the issues raised by different domain statistical properties, we used an approach called stacking Wolpert (1992). In the stacking framework, the data is first passed to a layer of independent models whose predictions are then used to train a meta-model that learns how to optimally combine the first layer's predictions. This model can itself be a stacking model and so the final prediction becomes: $f(\mathbf{x}) = p\left(y|g_1^L(\mathbf{x}), ..., g_{W_L}^L(\mathbf{x}), \boldsymbol{\theta}^{L+1}\right)$, where $g_{W_L}^L$ is the model at the final layer $L$, $W_L$ is the number of models in this layer and each intermediate layer model is defined as $g_i^k(\mathbf{x}) = p\left(y|g_1^{k-1}(\cdot), ..., g_{W_k}^{k-1}(\cdot), \boldsymbol{\theta}_i^k\right)$ , $k \in [1, L]$. In a multi-domain setting, one can pass each domain dataset to a different model in the base layer: $g_i^0(\mathbf{x}) = p\left(y|\mathbf{x}^i, \boldsymbol{\theta}_i^0\right)$, so that the width of the first layer $W_0$ is equal to the number of domains $M$.

Viewing biological systems as stacked layers of connected information is more likely to produce fruitful models though, and thus we would like to use all the modalities while considering their interactions. In chapter 6, we describe a method that uses the topology of each domains' data to induce deformations in the other related modalities, effectively tying the different layers together.

### Feature importance

One of the core goals in medical research is to identify the main drivers of a particular disease. Traditionally, the isolated influence of each variable on the input was inspected via the coefficients of linear regression $f(x) = \sum_{i=1}^{N} \alpha + \beta_i \cdot x_i$. In this scenario, it is easy to interpret the influence of each random variable since $\frac{\partial f(x)}{\partial x_i} = \beta_i$ and thus one unit change in $x_i$ causes a change of magnitude $\beta_i$ in the output. One of the hardest challenges in applying ML methods in biological data is the number of high order non-linear interactions between the features, that is, the outcome cannot be decomposed into functions of individual features as in $Y = \sum_{i=1}^{N} f_i(X_i)$ but rather depends on multi-feature functions like $Y = \sum_{i=1}^{|C|} f_i(\mathbf{X}_{c_i})$ with $C = \{c_1, ..., c_{|C|}\}$ and $c_i \subset \{1, ..., N\}$. This warrants the use of sophisticated models, which generally comes at the expense of model transparency. In chapters 2 and 3 we used tree-based models which implicitly compute feature importance by determining the optimal features to split the data. However, in chapter 4 we make use of a stacked model which obscures how much the model relies on each feature. To address this problem, global or local post-hoc methods like *permutation importance* Breiman (2001) or *LIME* Ribeiro, Singh, and Guestrin (2016) can be employed. Local methods address the problem of explaining what the model is doing for a particular instance while global methods try to estimate which are the most important variables for the model prediction. In chapter 5, we describe an extension of *LIME* to explain local model behavior in higher dimensional spaces.

Despite the intuitiveness and wide application of *permutation importance*, it produces biased results when the variables are highly correlated. In chapter 7 we describe how permuting correlated features together can correct this issue, but it is an incomplete solution since biological systems are characterized by interactions between more than two variables. Recent extensions to the popular importance method SHAP Štrumbelj and Kononenko (2014); Covert, Lundberg, and Lee (2020) take into account all interactions but they are very computationally expensive and they are still not completely bias free. In chapter 8, we describe a truly unbiased, model-agnostic, global feature importance method which considers all feature interactions with a fast computing time.

# Part I

# Applied Machine Learning in Clinical Research

# 2

# Improved cardiovascular risk prediction using targeted plasma proteomics in primary prevention

Renate M Hoogeveen*, **João P Belo Pereira***, Nick S Nurmohamed, Veronica Zampoleri, Michiel J Bom, Andrea Baragetti, S Matthijs Boekholdt, Paul Knaapen, Kay-Tee Khaw, Nicholas J Wareham, Albert K Groen, Alberico L Catapano, Wolfgang Koenig, Evgeni Levin, Erik S G Stroes

*Both authors contributed equally to this work

European Heart Journal, Volume 41, Issue 41, 1 November 2020, Pages 3998–4007, https://doi.org/10.1093/eurheartj/ehaa648

## 2.1 Abstract

### Aims

In the era of personalized medicine, it is of utmost importance to be able to identify subjects at the highest cardiovascular (CV) risk. To date, single biomarkers have failed to markedly improve the estimation of CV risk. Using novel technology, simultaneous assessment of large numbers of biomarkers may hold promise to improve prediction. In the present study, we compared a protein-based risk model with a model using traditional risk factors in predicting CV events in the primary prevention setting of the European Prospective Investigation (EPIC)-Norfolk study, followed by validation in the Progressione della Lesione Intimale Carotidea (PLIC) cohort.

### Methods and results

Using the proximity extension assay, 368 proteins were measured in a nested case–control sample of 822 individuals from the EPIC-Norfolk prospective cohort study and 702 individuals from the PLIC cohort. Using tree-based ensemble and boosting methods, we constructed a protein-based prediction model, an optimized clinical risk model, and a model combining both. In the derivation cohort (EPIC-Norfolk), we defined a panel of 50 proteins, which outperformed the clinical risk model in the prediction of myocardial infarction [area under the curve (AUC) 0.754 vs 0.730; $p < 0.001$] during a median follow-up of 20 years. The clinically more relevant prediction of events occurring within 3 years showed an AUC of 0.732 using the clinical risk model and an AUC of 0.803 for the protein model ($p < 0.001$). The predictive value of the protein panel was confirmed to be superior to the clinical risk model in the validation cohort (AUC 0.705 vs 0.609; $p < 0.001$).

### Conclusion

In a primary prevention setting, a proteome-based model outperforms a model comprising clinical risk factors in predicting the risk of CV events. Validation in a large prospective primary prevention cohort is required to address the value for future clinical implementation in CV prevention.

## 2.2 Introduction

Identification of asymptomatic people at the greatest cardiovascular (CV) risk remains a major challenge in primary prevention.Yusuf et al. (2014); Fernández-Friera et al. (2017) Clinically used risk algorithms, including the Framingham risk score, pooled cohort equations, and Systemic Coronary Risk Evaluation (SCORE) system, are based on traditional risk factors for CV disease and predict future events with limited accuracy.Piepoli et al. (2016); Goff et al. (2014) Accordingly, a substantial proportion of the general population at risk remains unidentified until their first clinical event.Fernández-Friera et al. (2017) Despite adding individual plasma biomarkers such as pro-brain natriuretic peptide (BNP), high sensitivity troponins, and high sensitivity C-reactive protein (CRP) to clinical risk engines, the overall improvement has been limited.Piepoli et al. (2016) This may be explained by the fact that the vast majority of single markers are selected based on specific pathophysiological concepts, which do not reflect the true complexity of atherosclerosis. In fact, CV risk is the result of an interplay between comorbidities (chronic inflammatory diseases, metabolic derangements) and exogenous risk factors, propagated by a variety of pathophysiological axes, comprising but not limited to lipids, coagulation, and inflammation.Hoogeveen et al. (2018)

Simultaneous assessment of a large number of plasma proteins may hold a promise to further refine risk assessment.Lindsey et al. (2015) To this end, either discovery proteomics, aiming to identify new diagnostic markers or therapeutic targets, or targeted proteomics, aimed at quantification of proteins of specific interest, can be applied.Lindsey et al. (2015) Widespread use of proteomics has been precluded by labour intensiveness, high costs, and the complex clinical interpretation of the bulky results. More recently, these limitations have largely been resolved. Technical advances now allow for high-throughput proteomic analysis in a reproducible and cost-effective manner.Assarsson et al. (2014) In parallel, advanced computational modelling has facilitated the interpretation of large data sets for clinical implementation.Deo (2015); Rajkomar, Dean, and Kohane (2019) Using these innovations, a targeted protein panel was found to modestly improve the prediction of incident atherosclerotic CV disease in primary prevention,Yin et al. (2014) whereas Ganz et al.Ganz et al. (2016) substantiated that targeted proteomics also outperformed refit Framingham in predicting recurrent coronary events. In support, we recently identified two complementary protein signatures predicting the presence of high-risk plaque and the absence of coronary atherosclerosis in subjects referred for the analysis of anginal complaints,Danad et al. (2017) clearly outperforming the traditional risk algorithm.Bom et al. (2019)

In the present study, we hypothesized that a protein-based risk model can outperform prediction using traditional risk factors in the primary prevention setting. Therefore, we tested the ability of a targeted proteomics panel comprising 368 proteins, related to pathways and/or risk factors involved in atherogenesis, to predict CV event risk in a nested case–control sample of the European Prospective Investigation (EPIC)-Norfolk population study,Day et al. (1999) using advanced machine learning techniques. The findings were subsequently validated in the independent, external primary prevention cohort [Progressione della Lesione Intimale Carotidea (PLIC)].Olmastroni et al. (2019)

## 2.3 Methods

### 2.3.1 Study populations

The derivation cohort was a nested case–control sample derived from the EPIC-Norfolk prospective population study, comprising 25633 individuals recruited from general practices in the Norfolk area, UK.Day et al. (1999) Study participants aged between 39 and 79 years were enrolled between 1993 and 1997. At baseline, patients completed general health questionnaires and a panel of measurements was performed. During follow-up, all individuals were flagged for mortality at the UK Office of National Statistics and vital status was ascertained for the entire cohort. Data on all hospital contacts throughout England and Wales were obtained using National Health Service numbers through linkage with the East Norfolk Health Authority (ENCORE) database. Hospital records and death certificates were coded by trained nosologists and categorized according to the International Classification of Disease 10th revision (ICD-10). The study protocol was approved by the Norwich District Health Authority Ethical Committee. All individuals gave written informed consent. For the current study, we selected 822 apparently healthy individuals in a nested case–control sample from the EPIC-Norfolk study. Apparently healthy individuals were defined as study participants who did not report a history of CV disease. A total of 411 individuals who developed an acute myocardial infarction (either hospitalization or death with ICD code I21-22 coded as the underlying cause) between baseline and follow-up through 2016 were selected together with 411 apparently healthy individuals who remained free of any CV disease during follow-up (Figure 2.1).Boekholdt et al. (2004); Saleheen et al. (2015)

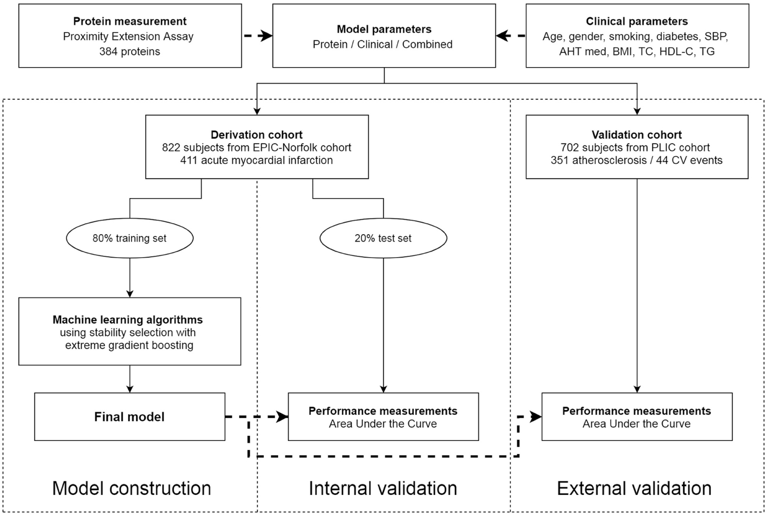

Protein measurement
Proximity Extension Assay
384 proteins
Model parameters
Protein / Clinical / Combined
Clinical parameters
Age, gender, smoking, diabetes, SBP, AHT med, BMI, TC, HDL-C, TG
Derivation cohort
822 subjects from EPIC-Norfolk cohort
411 acute myocardial infarction
Validation cohort
702 subjects from PLIC cohort
351 atherosclerosis / 44 CV events
80% training set
20% test set
Machine learning algorithms
using stability selection with extreme gradient boosting
Final model
Performance measurements
Area Under the Curve
Performance measurements
Area Under the Curve
Model construction
Internal validation
External validation

Figure 2.1: Machine learning workflow of model construction and validation. AHT med, antihypertensive medication; BMI, body mass index; CV, cardiovascular; EPIC, European Prospective Investigation; HDL-C, high-density lipoprotein cholesterol; PLIC, Progressione della Lesione Intimale Carotidea; SBP, systolic blood pressure; TC, total cholesterol; TG, triglycerides.

## 2.4 Introduction

The validation cohort was the PLIC cohort, a single-centre, observational, cross-sectional, and prospective study of subjects enrolled on a voluntary basis in 1998–2000 and followed for 11 years on average in the northern area of Milan.Olmastroni et al. (2019) The 2606 Caucasian subjects who were enrolled in the study underwent four periodic visits. Data about clinical, pathological, familial, and pharmacological history and lifestyle habits were collected based on medical records and self-reporting during these visits. Blood samples were withdrawn, and subjects underwent carotid ultrasound to assess the presence or absence of carotid vascular damage. The presence of documented stenosis or vascular damage on aorta and limb arteries was included in the definition of subclinical atherosclerosis. For the validation cohort, 702 subjects were selected, of whom 351 developed atherosclerosis, comprising subclinical atherosclerosis and 44 subjects who suffered from a CV event, and 351 gender-matched controls during follow-up (Figure 1). Cardiovascular events were defined as a combined endpoint of coronary heart disease (myocardial infarction, unstable angina, coronary revascularization, silent ischaemia) and cerebrovascular disease (ischaemic stroke and transient ischaemic attack). This study was approved by the ethics committee and was performed in accordance with the Declaration of Helsinki. All participating subjects signed informed consent.

### 2.4.1 Biochemical analyses

In EPIC-Norfolk, non-fasting blood was drawn at baseline from study participants, from which total cholesterol, high-density lipoprotein (HDL) cholesterol, and triglycerides were determined with the RA1000 analyser (Bayer Diagnostics, Basingstoke, UK). The Friedewald formula was used for the calculation of low-density lipoprotein (LDL) cholesterol levels.Friedewald, Levy, and Fredrickson (1972) After blood withdrawal, ethylene diamine tetra acetic acid (EDTA) samples were kept overnight at room temperature before transporting to the EPIC-Norfolk laboratory for centrifugation. Hereafter, the remaining plasma was stored at -80°C for future analyses.

In the PLIC cohort, blood samples were collected after overnight fasting. Samples were kept on ice after blood withdrawal and centrifuged within 1 h at 3000 rounds per minute for 12 min (Eppendorf 5810R centrifuge). Plasma samples were subsequently stored in 200 $\mu$L aliquots at -80°C. Since multiple aliquots were stored, multiple freeze/thaw cycles were prevented. Total cholesterol, HDL cholesterol, triglyceride and glucose levels were determined in serum samples with the Cobas Mira Plus Analyser (Horiba, ABX, France). Again, the Friedewald formula was used for the calculation of LDL cholesterol levels.Friedewald, Levy, and Fredrickson (1972)

In 2019, we selected cases and controls from both cohorts, whereupon aliquots were thawed and plasma was transferred, on ice, to 96-well plates. The 96-well plates were shipped to Olink proteomics AB (Uppsala, Sweden) on dry ice for analysis using the proximity extension assay technology. Levels of 368 proteins were measured from the CV II, CV III, Cardiometabolic, and Inflammation panels. These panels were selected for their known associations with CV disease. Cases and controls were randomly distributed across plates and assays were performed in a blinded fashion. Data are Normalized Protein eXpression values. Using an internal extension control and an interpolate control, data quality is controlled and normalized. All assay validation data are available on the manufacturer's website (www.olink.com).

### 2.4.2 Statistical analysis

Data are presented as mean ± standard deviation for normally distributed variables or median with inter-quartile range for skewed data. Categorical variables are expressed as absolute number and percentages. Independent sample t-tests and Mann–Whitney U tests were used where appropriate. Two-sided p-values $\leq 0.05$ were considered statistically significant. Data were analysed using R version 3.5.1 (R Foundation, Vienna, Austria).

### 2.4.3 Model construction

A combination of stacking generalization framework,Wolpert (1992); Caruana et al. (2004) tree-based ensemble methods, and multiple gradient boosting classifiersChen and Guestrin (2016) was used to best discriminate between cases and controls. Using these techniques, explained in detail below, different models were constructed. First, a clinical risk model was built. The clinical risk model included parameters of different validated risk scores, the Framingham Risk Score, pooled cohort equations, and SCORE. Parameters included in the clinical risk model were age, gender, body mass index, systolic blood pressure, smoking status, and presence of diabetes, the use of antihypertensive medication, total cholesterol levels, HDL cholesterol levels, and triglyceride levels. Second, a protein-based model was constructed using the measured plasma proteins only. A third model was formed by stacking the clinical risk parameters with the protein parameters. The proteins and clinical parameters were allowed to compete in the formation of this model. All three models were validated in the validation cohort without adjustments.

Next, considering the long-term follow-up of subjects and the fact that proteins are subject to change due to lifestyle and medical interventions, we assessed the optimal time point of prediction of acute myocardial infarction in the derivation cohort using Markov-Chain Monte Carlo techniques. For this optimal time point of prediction, similar to the long-term modelling, a clinical risk model, protein model, and a combined model were formed. Specifically for this time point, we calculated the net reclassification improvement (NRI) as described by Pencina et al.Pencina, D'Agostino, and Steyerberg (2011) for case–control studies. Theretofore, we used the acute myocardial infarction prevalence of the total EPIC-Norfolk cohort in the same period.

In addition, we constructed survival models for both the protein model and the clinical risk model in the derivation cohort, to compare model performance across all possible time points. This time-to-event analysis was performed using identical machine learning techniques as the binary models, with the implementation of a survival loss function. Inverse probability of censoring weighting was used to cope with the right-censored data.Vock et al. (2016) Using these survival models, time-dependent area under the curves (AUCs) were calculated with a 2-year interval starting from 3 years up to the median follow-up of 20 years.

### 2.4.4 Machine learning techniques

All binary models were constructed using the same machine learning techniques (Figure 1). First, to avoid overfitting of the models, the derivation data set was split into two sets: a training set of 80% and a test set of 20%. The model was not exposed to data from the 20% test set; this was only used for the performance measurements. Ten percentage of the 80% training set was used for model refinement before the model performance was tested in the test set. In construction of the models and identification of the most reliable biomarker signature in our datasets (both proteomics and clinical), we used stability selection with extreme gradient boosting. Gradient boosting is a statistical learning technique, which produces a non-linear model in the form of an ensemble of weak prediction tree-based models. It builds the model in a stage-wise fashion, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. The extreme gradient boosting classification algorithm optimizes a cost function by iteratively choosing a weak hypothesis that points in the negative gradient direction.Caruana et al. (2004); Chen and Guestrin (2016) Using a fivefold cross-validation by random reshuffling of the training set, overfitting was avoided. For increased confidence, this procedure was repeated multiple times on a completely reshuffled dataset. Furthermore, the method was coupled with a rigorous stability selection procedure to ensure the reliability and robustness of the obtained parameters.Meinshausen and Bühlmann (2010) Finally, we applied a permutation (randomization test) to evaluate statistical validity of the results,Marques et al. (2010) since standard univariate significance tests cannot be applied to the used models due to the large number of features. The permutation test comprised 1000 reruns of the model, every time randomly permuting the output variable (presence/absence of the event). By evaluating the distribution of all the results obtained in these simulations and comparing it to the true outcomes, we computed statistical significance associated with the joint panel of the selected markers. We also reported importance scores for each of the proteins that demonstrate preferences of model when constructing non-linear prediction function based on the selected biomarkers. Python version 3.7 (www.python.org), with packages Numpy, Scipy, and Scikits-learn, was used for machine learning models and visualizations.

## 2.5 Results

### 2.5.1 Study populations

Baseline characteristics of individuals from the derivation and validation cohort are provided in Table A.1. In short, cases in the derivation cohort were more likely to have traditional risk factors for CV disease (older, more likely male, smokers and had higher blood pressure and cholesterol levels). In the validation cohort, cases were more likely to be older, to smoke, and have high blood pressure. In all participants, 368 preselected proteins were measured; proteins were excluded if $\geq$ 90% of the values were below lower limit of detection. Due to the latter and overlap between the panels, the final analysis included 333 unique proteins (we refer the reader to the Supplementary material, Table S1 found online in https://doi.org/10.1093/eurheartj/ehaa648).

### 2.5.2 Prediction of acute myocardial infarction

Prediction of myocardial infarction using a machine learning model consisting of 50 plasma proteins over a median follow-up of 20 years resulted in an receiver operating characteristic (ROC) AUC of 0.754 ± 0.011 (permutation test $p = 0.0099$; Figures 2.2 and 2.3A and Table 2.1). In comparison, the use of the clinical risk model resulted in an ROC AUC of 0.730 ± 0.015 (permutation test $p = 0.0099$). Combining the protein panel with clinical risk model resulted in an ROC AUC of 0.764 ± 0.015 (permutation test $p = 0.0099$). The biomarker model was superior to the clinical risk model ($p < 0.001$). The combined protein and clinical risk model showed a small incremental AUC of 0.01 in comparison with the protein model alone. Using Markov-Chain Monte Carlo techniques, the optimal time point for prediction was found at 1132 days ($\sim$3 years), which included 66 acute myocardial infarctions and all 411 non-myocardial infarction controls. Of the 50 proteins that were selected for the $\sim$3 years prediction model, 33 overlapped with the original 20 years model (Supplementary material online, Figure A.1). Focusing on the events occurring within ∼3 years after baseline blood withdrawal, the ROC AUC increased to $0.803 \pm 0.093$ (permutation test $p = 0.0145$; Figure 3B), as opposed to $0.732 \pm 0.164$ (permutation test $p = 0.0099$) using the clinical risk model. The combination of the protein and clinical risk parameters resulted in an ROC AUC of $0.808 \pm 0.085$ (permutation test $p = 0.0178$). Now, the biomarker model was superior to the clinical risk model with an incremental AUC of 0.07 ($p = 0.025$) but not to the combination of the protein and clinical risk model ($p = 0.721$). For the short-term prediction, the NRI of the protein model in comparison to the clinical risk model was 6.6%. In the survival analysis, the protein model resulted in a mean time-dependent AUC of $0.717 \pm 0.027$, which was superior across all time points compared to the clinical risk model mean AUC of $0.653 \pm 0.031$ ($p < 0.001$; Supplementary material online, Figure A.2).

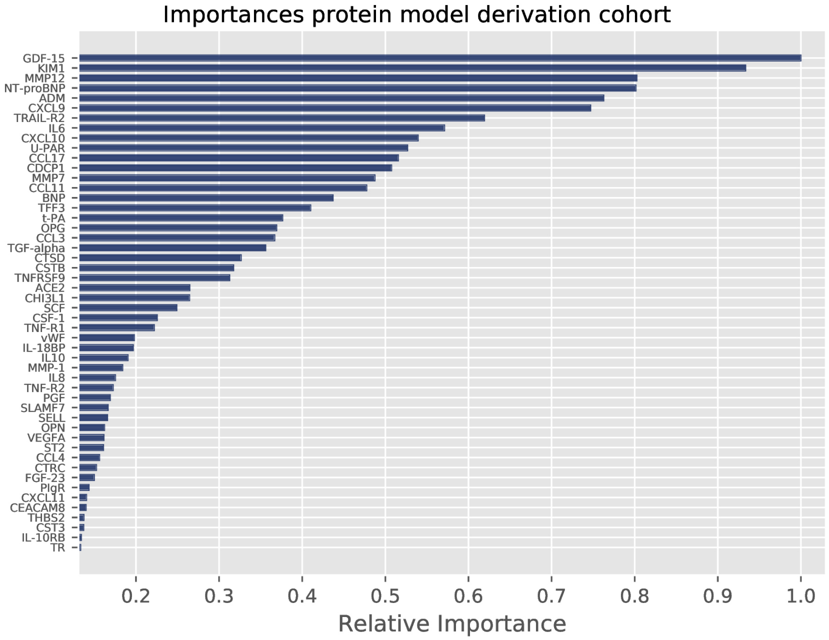


Figure 2.2: Importance plot of proteins. Relative importance of 50 proteins predictive in derivation cohort.

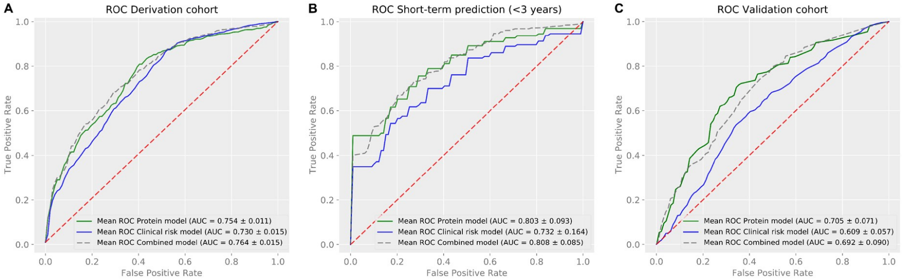


Figure 2.3: Receiver operating characteristics of prediction models. (A) Prediction of events with protein, clinical risk, and combined model in derivation cohort. (B) Short-term prediction (<3 years) of events with protein, clinical risk, and combined model in derivation cohort. (C) Prediction of events with protein, clinical risk, and combined model in validation cohort. AUC, area under the curve; ROC, receiver operating characteristic.

| | Derivation cohort | Derivation (<3 years) | Validation cohort |
|---|---|---|---|
| Protein model | $0.754 \pm 0.011$ | $0.803 \pm 0.093$ | $0.705 \pm 0.071$ |
| Clinical risk model | $0.730 \pm 0.015$ | $0.732 \pm 0.164$ | $0.609 \pm 0.057$ |
| Combined model | $0.764 \pm 0.015$ | $0.808 \pm 0.085$ | $0.692 \pm 0.090$ |

Table 2.1: Average receiver operating characteristics area under the curve of the prediction models.

### 2.5.3 Validation of the predictive value

We validated the discriminatory ability of the 50 proteins from the derivation cohort in the validation cohort. First, we investigated the ability of the proteins to predict subclinical atherosclerosis. The prediction was relatively poor with an ROC AUC of 0.648 ± 0.056 (permutation test $p = 0.0297$; Supplementary material online, Figure A.3). When validating the proteins in the 44 participants who suffered from CV events vs. the 351 participants with no signs of atherosclerosis, the protein model resulted in an ROC AUC of 0.705 ± 0.071 (permutation test $p = 0.0099$; Figure 3C), compared to the clinical risk model ROC AUC 0.609 ± 0.057 (permutation test $p = 0.0700$; Table 2.1). The protein model was significantly better than the clinical risk model in the validation cohort ($p < 0.001$). The combined protein and clinical risk model resulted in an ROC AUC of $0.692 \pm 0.090$ (permutation test $p = 0.0099$), which was not better than the protein model alone ($p = 0.618$).

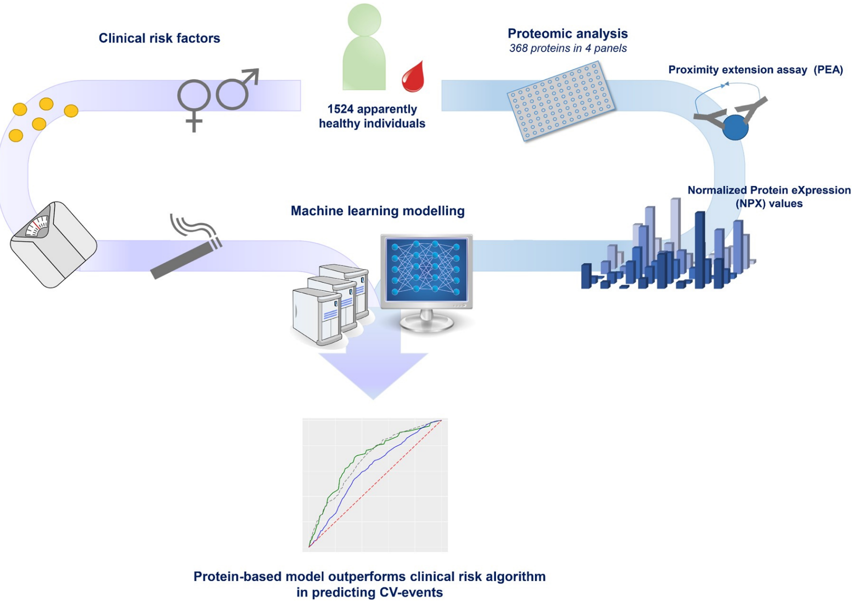

Clinical risk factors
1524 apparently healthy individuals
Proteomic analysis
368 proteins in 4 panels
Proximity extension assay (PEA)
Normalized Protein eXpression (NPX) values
Machine learning modelling
Protein-based model outperforms clinical risk algorithm in predicting CV-events

Figure 2.4: **Take-home figure**: Derivation and validation of a plasma proteomic model improves cardiovascular risk prediction in a primary prevention setting, demonstrating the potential of a proteomics panel to further refine risk assessment. CV, cardiovascular; NPX, Normalized Protein eXpression; PEA, proximity extension assay.

## 2.6 Discussion

Using targeted proteomics, we show that a panel of 50 proteins outperforms the clinical risk model in predicting the risk of myocardial infarction ($<$ 3 years) in a primary prevention setting with an AUC increase in the ROC curve of 0.07. Improvement in predicting CV events during the entire (median) 20-year follow-up period was significant, albeit modest. In an external independent validation cohort, the predictive value of the protein panel for CV events was confirmed and superior to the clinical risk model (incremental AUC 0.10). Survival analysis showed superiority of the protein model to the clinical risk model at all tested time points ($p < 0.001$). Collectively, these data show that a novel proteomic panel offers a significant improvement in CV risk discrimination compared to a clinical risk model based on traditional risk factors (Figure 2.4)

### 2.6.1 Protein-based risk prediction outperforms traditional risk factors

We substantiate the predictive value of a panel comprising 50 proteins for a first MI with an ROC AUC of 0.754 ± 0.011, using targeted proteomics. Although outperforming the prediction by the clinical risk model ($p < 0.001$), the AUC increase of 0.02 is very modest. Interestingly, the prediction of earlier MI (within 3 years after baseline blood sampling) using the plasma protein panel performed better, with an incremental ROC of 0.07. Where genetic prediction models are advocated to predict lifelong risk, the ability of our protein model particularly in shorter-term risk prediction most likely highlights the property of plasma proteomics to reflect a more proximate timeframe. Confronted with continuous changes in lifestyle as well as medical interventions during the course of a life, repeated proteome-based risk estimation as a 'liquid health check' may help to further improve lifetime risk estimation.Williams et al. (2019) The prediction of predominantly short-term MI substantiates our previous findings that proteomics also predicts the presence of high-risk plaques in patients, which are closely associated with an increased risk for ensuing MI.Nerlekar et al. (2018) Previous cohort studies have also reported benefit of proteins in risk prediction. The Framingham Heart Study investigators evaluated a panel of 85 plasma proteins in relation to CV events in primary care setting.Ho et al. (2018) Using a multi-marker analysis, they reported eight biomarkers predictive for incident CVD, which on top of clinical parameters achieved an ROC of 0.758. These data are in line with data reported for the prediction of recurrent coronary events, where a panel of 9 out of 1130 proteins modestly improved risk prediction (AUC 0.70) compared to the clinical risk algorithm (AUC 0.64).Ganz et al. (2016)

### 2.6.2 Proteins predictive of cardiovascular events

Based on previous findings,Bom et al. (2019) we used targeted proteomics using proteins relating to cardiometabolic disease, CV disease, and inflammation/immune responses. The majority of proteins in our model were related to immune system response; particularly proteins involved in chemotaxis, migration, apoptosis, and angiogenesis. Most of the proteins found to predict early vs. all events overlapped (33/50). Several proteins merit further attention considering their marked contribution to the final model. Growth Differentiation Factor 15 (GDF-15) was the protein with the largest contribution. In chronic diseases, GDF-15 produced by leukocytes has been shown to enhance inflammation.Kempf et al. (2011) Other prominent candidates involve the N-terminal pro-B-type natriuretic peptides and BNP, which are established markers for heart failure and predictors of CV events.Wang et al. (2006) There is also a preponderance of inflammatory proteins, comprising metalloproteinase-12 (MMP-12), TRAIL receptor 2 and interleukin-6. These proteins, involved in matrix degradation, apoptosis and inflammation induction, reflect major pathways contributing to atherosclerotic lesion formation and destabilization. Interestingly, there is clear overlap in proteins and pathways when comparing our data to previous CVD-proteomic studies. Thus, GDF-15 was also identified as a predictive candidate in previous studies.Bom et al. (2019); Ho et al. (2018); Lind et al. (2015) Similarly, the relevance of plasma MMP12Ganz et al. (2016); Bom et al. (2019); Stenemo et al. (2018); Nowak et al. (2018) and various chemo/cytokinesHo et al. (2018); Lind et al. (2015); Nowak et al. (2018) underscore consistency between these studies.

### 2.6.3 Validation in the Progressione della Lesione Intimale Carotidea cohort

Validation of our findings was performed in the primary prevention PLIC cohort, in which both repetitive non-invasive measures for atherosclerosis and CV events were collected during an 11-year follow-up.Olmastroni et al. (2018) The 50-protein model from the derivation cohort showed reasonable prediction of CV events with an ROC AUC of 0.705 ± 0.071, with an incremental AUC of 0.10 compared to the clinical risk model in the PLIC cohort (ROC AUC of 0.609 ± 0.057). We also assessed the value of the proteomic model to predict the presence of subclinical atherosclerotic lesions assessed using ultrasound, revealing an ROC AUC of 0.648 ± 0.056. The failure of plasma proteomics to accurately predict the presence of subclinical atherosclerosis is in line with the findings in the derivation cohort, where the protein signature performed better for early/mid-term CV events than for long-term events.

### 2.6.4 Clinical perspective

In previous studies, adding single plasma markers to clinical risk algorithms resulted in only a modest improvements of risk prediction.Force et al. (2018); Mortensen et al. (2018) Here, we report a marked improvement in CVD risk prediction using a targeted proteomics approach. The hurdles for using proteomic panels in clinical practice have been largely removed with the advent of affordable high-throughput technology requiring only minimal amounts of plasma. More importantly, machine learning technology further facilitates the use of complex, massive data (such as proteomics) in clinical decision making.Deo (2015); Rajkomar, Dean, and Kohane (2019) The need for better discrimination of subjects at highest CV event risk is underscored by the advent of expensive medication in CVD preventive therapy beyond generic statins, among which PCSK9-antibodies,Schwartz et al. (2018); Sabatine et al. (2017) low-dose Xa inhibition,Eikelboom et al. (2017) SGLT2 inhibitors,Perkovic et al. (2019); Wiviott et al. (2019) and GLP1 agonists.Husain et al. (2019); Marso et al. (2016) Whereas a high-risk proteomic panel holds a promise to help identify higher-risk subjects, it is tempting to speculate that pathway analysis of the proteomic signature may also allow for the guidance of what medication to use in specific patient categories.Lindsey et al. (2015) This concept is underscored by the CANTOS study, where predominantly CRP responders demonstrated CV benefit of interleukin 1 beta-antibody administration.Ridker et al. (2017a) However, this concept needs further validation with special emphasis on relationships between biomarkers and protein network analysis.Lindsey et al. (2015); Johnson et al. (2018) Hypothetically, the development of a targeted-proteomic based risk score might enable a more patient-tailored approach for the primary prevention of CV events.

### 2.6.5 Strengths and limitations

The combination of proteomics with machine learning technology is highly promising.Deo (2015); Rajkomar, Dean, and Kohane (2019) Machine learning technology can process data that surpasses the capacity of traditional statistics and the human brain to comprehend.Rajkomar, Dean, and Kohane (2019) One of the most important differences is that our predictive machine learning model is based on multiple proteins in a panel, which collectively leads to a reliable prediction. Using machine learning, non-linear relationships and interactions among proteins are taken into account, in contrast to univariate models that only address up- and/or down-regulation of individual proteins. In the current analyses, we refitted the clinical risk factors from the Framingham risk score and SCORE to best fit our cohort data, aiming to improve the performance of traditional risk factors. By applying analogous machine learning methods for the traditional risk factors, the observed superiority of our protein model over the clinical risk model is distinct.

Several potential limitations deserve closer attention. First, the cohorts used in this study were collected over a decade ago. Over the years, risk factor management has improved, plaque characteristics have altered, and patient characteristics have changed.van Lammeren et al. (2014) Second, our validation cohort had a limited number of CVD events. However, validation of our protein model on these events was reasonable and the model outperformed the clinical risk model in the validation cohort, in an even stronger manner than in the derivation cohort. Third, we used targeted rather than untargeted proteomics in our study. Proteins were preselected as potential biomarkers for CV disease, since clinical verification, rather than protein discovery was the goal in our study. Despite analysing a broad range of proteins, we may have missed other predictors of CV event risk due to the use of targeted proteomics only. As a result, we may have underestimated the true potential of proteomics in CV risk estimation. Fourth, in contrary to other primary prevention risk scores such as the Pooled Cohort Equation,Goff et al. (2014) our constructed models do not predict lifetime risk, which could be useful in primary prevention patients characterized by a relatively low short-term CV risk, such as in subjects below 50 years of age. However, in the present study, we preferred shorter-term prediction for several reasons. Most importantly, the mean age of both our derivation and validation cohorts was well above the age of 50 years, resulting in a higher short-term risk even in primary prevention. Furthermore, diagnostic improvement in detecting high-risk patients is currently needed to make decisions on initiating novel medication on top of routine regimens, and for these decisions, relatively short-time horizons are routinely used. Fifth, the samples in the derivation cohort were non-fasting, while the samples in the validation cohort were collected after an overnight fast. Despite this difference, the protein model performed comparable in the validation cohort. Finally, the current analyses were performed in subjects primarily from European ancestry. Hence, the predictive power remains to be validated in different ethnicities.

## 2.7 Conclusions

In primary prevention, proteome-based risk prediction significantly outperforms prediction using clinical risk factors in predicting the risk of acute myocardial infarction and CV events, especially in the first 3 years. In the midst of novel, expensive drugs, prediction of individual CVD risk and treatment benefit is increasingly important. Further large prospective studies will have to determine the true value of proteome-based risk scores in primary prevention.

## Acknowledgements

The authors wish to thank the participants and staff of the EPIC-Norfolk prospective population study and PLIC cohort.

## Funding

This project has received funding from the European Research Area Network on Cardiovascular Diseases (ERA-CVD) grant (ERA-CVD JTC2017) and the European Union's Horizon 2020 research and innovation programme under grant agreement No 667837 (REPROGRAM). The EPIC-Norfolk Study is funded by Cancer Research UK (14136) and the Medical Research Council (G1000143).

### Conflict of interest:

K.-T.K. reports a research grant from the Medical Research council UK paid to institution. A.L.C. reports grants from Sanofi, Regeneron, Merck, Mediolanum, SigmaTau, Menarini, Kowa, Recordati, Eli Lilly, and personal fees from Merck, Sanofi, Regeneron, AstraZeneca, Amgen, SigmaTau, Recordati, Aegerion, Kowa, Menarini, Eli Lilly, and Genzyme, all outside the submitted work. W.K. reports personal fees from AstraZeneca, Novartis, Pfizer, The Medicines Company, DalCor, Kowa, Amgen, Corvidia, Berlin-Chemie, Sanofi, Bristol-Myers Squibb and grants and non-financial support from Beckmann, Singulex, Abbott, and grants from Roche Diagnostics, all outside the submitted work. E.S.G.S. reports ad-board/lecturing fees paid to his institution from Amgen, Sanofi-Regeneron, Esperion, Novartis, and grants from Athera; all outside the submitted work. All other authors declared no conflict of interest.

# 3

# Secondary prevention using targeted proteomics

Nick S. Nurmohamed*, **João P. Belo Pereira***, Renate M. Hoogeveen, Jeffrey Kroon, Jordan M. Kraaijenhof, Farahnaz Waissi, Nathalie Timmerman, Michiel J. Bom, Imo E. Hoefer, Paul Knaapen, Alberico L. Catapano, Wolfgang Koenig, Dominique de Kleijn, Frank L.J. Visseren, Evgeni Levin, Erik S.G. Stroes

*Authors contributed equally

European Heart Journal, Volume 43, Issue 16, 21 April 2022, Pages 1569–1577, https://doi.org/10.1093/eurheartj/ehac055

## 3.1 Abstract

### Aims

Current risk scores do not accurately identify patients at highest risk of recurrent atherosclerotic cardiovascular disease (ASCVD) in need of more intensive therapeutic interventions. Advances in high-throughput plasma proteomics, analysed with machine learning techniques, may offer new opportunities to further improve risk stratification in these patients.

### Methods and results

Targeted plasma proteomics was performed in two secondary prevention cohorts: the Second Manifestations of ARTerial disease (SMART) cohort (n = 870) and the Athero-Express cohort (n = 700). The primary outcome was recurrent ASCVD (acute myocardial infarction, ischaemic stroke, and cardiovascular death). Machine learning techniques with extreme gradient boosting were used to construct a protein model in the derivation cohort (SMART), which was validated in the Athero-Express cohort and compared with a clinical risk model. Pathway analysis was performed to identify specific pathways in high and low C-reactive protein (CRP) patient subsets. The protein model outperformed the clinical model in both the derivation cohort [area under the curve (AUC): 0.810 vs. 0.750; $p < 0.001$] and validation cohort (AUC: 0.801 vs. 0.765; $p < 0.001$), provided significant net reclassification improvement (0.173 in validation cohort) and was well calibrated. In contrast to a clear interleukin-6 signal in high CRP patients, neutrophil-signalling-related proteins were associated with recurrent ASCVD in low CRP patients.

### Conclusion

A proteome-based risk model is superior to a clinical risk model in predicting recurrent ASCVD events. Neutrophil-related pathways were found in low CRP patients, implying the presence of a residual inflammatory risk beyond traditional NLRP3 pathways. The observed net reclassification improvement illustrates the potential of proteomics when incorporated in a tailored therapeutic approach in secondary prevention patients.

### Key question

Does targeted plasma proteomics improve cardiovascular risk prediction in secondary prevention patients? Are different pathways contributing to cardiovascular risk in high and low C-reactive protein (CRP) patients?

### Key finding

Cardiovascular risk prediction with targeted plasma proteomics outperformed prediction with clinical risk factors resulting in major net reclassification improvement. Neutrophil-signalling-related proteins were associated with cardiovascular events in low CRP patients.

### Take-home message

Routine implementation of a targeted protein panel in cardiovascular risk prediction holds promise to improve risk stratification in secondary prevention. The involvement of neutrophil-related pathways in low CRP patients indicates residual inflammatory risk beyond NLRP3.

## 3.2 Introduction

The residual burden of atherosclerotic cardiovascular disease (ASCVD) remains large, despite the use of guideline-based preventive medication.Jernberg et al. (2015) The successful introduction of novel agents, comprising proprotein convertase subtilisin-like/kexin type 9 inhibitors,Sabatine et al. (2017); Schwartz et al. (2018) low-dose oral anticoagulants,Eikelboom et al. (2017) sodium-glucose cotransporter 2 inhibitors,Zinman et al. (2015) glucagon-like peptide-1 agonists,Marso et al. (2016); Husain et al. (2019) anti-inflammatory agents,Nidorf et al. (2020); Ridker et al. (2017a) and icosapent ethyl,Bhatt et al. (2019) offers an opportunity to further reduce the burden of recurrent ASCVD risk. However, the expanding choice of novel agents has also underscored the need to implement cost-effective therapeutic regimes, which mandates more accurate identification of patients at highest risk in order to solidify the highest absolute ASCVD benefit.Annemans et al. (2018) Epidemiological surveys have demonstrated a highly variable residual risk in patients with established ASCVD ranging from <5% to a more than 40% 10-year recurrence risk.Kaasenbrood et al. (2016) Clinical characteristics included in traditional risk prediction scores poorly discriminated individual recurrence of ASCVD,Jensen (2016) attested by the modest c-statistic of

0.64 [95% confidence interval (CI) 0.63 – 0.65] of the Second Manifestations of ARTerial disease (SMART) score in three independent secondary prevention cohorts.Kaasenbrood et al. (2016) RidkerRidker (2016, 2018) has argued to use C-reactive protein (CRP) as stratifying marker in order to identify residual inflammatory risk; however, it remains a matter of debate whether CRP reflects the entirety of inflammatory responses involved in atherogenesis.Soehnlein and Libby (2021) Therefore, improved methods to identify patients at highest recurrence risk are needed to help guide ASCVD risk-based therapeutic decisions.

Protein-based risk scores hold a major promise to improve ASCVD risk prediction, since proteins are not only influenced by the genetic background of an individual, but can also reflect adverse changes due to lifestyle alterations and specific pathways contributing to ASCVD risk.Williams et al. (2019); Lindsey et al. (2015) Improvements in machine learning techniques could allow clinical doctors to interpret the massive datasets emerging from proteomic analyses in an outpatient setting, which cannot be analysed using traditional statistical methods.Williams et al. (2019); Hoogeveen et al. (2020); Riley et al. (2020) Previously, we showed that the use of a targeted proteomics approach outperformed traditional ASCVD risk scores in a primary prevention setting.Hoogeveen et al. (2020) However, given their high recurrence risk, the most urgent need to identify highest-risk patients pertains to secondary prevention patients.Annemans et al. (2018); Ray et al. (2021)

In the present study, we evaluated the predictive value of targeted proteomics in a secondary prevention setting using advanced machine learning techniques. To this end, we performed plasma proteomics in two large secondary prevention cohorts. As a derivation cohort, we used a high-risk subset of secondary prevention patients included in the SMART cohort, followed by validation of these findings in an independent secondary prevention cohort; the Athero-Express.Simons et al. (1999); Verhoeven et al. (2004) In an exploratory analysis, inflammatory pathways were assessed by dividing patients into high or low residual inflammatory risk profiles based on baseline CRP levels.

## 3.3 Methods

### 3.3.1 Selection of patients

The SMART cohort is an ongoing prospective single-centre cohort of the University Medical Center Utrecht.Simons et al. (1999) Patients younger than 80 years were included from 1996 onwards, if they had clinically manifest atherosclerotic disease or marked risk factors for atherosclerosis. Previously, a clinical risk model (SMART) was developed and validated to estimate the absolute risk for recurrent ASCVD events.Dorresteijn et al. (2013) We selected all subjects who entered the SMART cohort for myocardial infarction, stroke, or transient ischaemic attack with a 10-year SMART risk score above 15% and had blood samples available. A total of 870 participants were included as a derivation cohort.

The Athero-Express study was initiated in 2002, and included patients undergoing carotid and femoral endarterectomy for previous ischaemic cerebral events or peripheral artery disease.Verhoeven et al. (2004) Patients were followed up until 3 years after the endarterectomy. We included 700 subjects who underwent a carotid endarterectomy following a stroke or transient ischaemic attack with plasma samples and complete follow-up data available as validation cohort.

### 3.3.2 Proteomic analyses

For both cohorts, the procedures for blood withdrawal and storage have been described previously.Simons et al. (1999); Verhoeven et al. (2004) In short, plasma samples were collected fasting at baseline in the derivation cohort, whereas samples were collected non-fasting on the preoperative day in the validation cohort. In both cohorts, plasma samples were directly centrifuged and stored at -80$\circ$C for future analyses. For this study, frozen plasma samples of selected subjects from both cohorts were collected from storage and transferred to Olink proteomics AB (Utrecht, The Netherlands) on dry ice for Proximity Extension Assay analysis. We measured levels of 276 proteins from the Cardiovascular II, Cardiovascular III, and Cardiometabolic panels. These panels were selected based on known associations with ASCVD. All samples with a quality control warning or with $\geq$40% of measurements below the lower limit of detection (LOD) were excluded from the analysis; separately per proteomic panel. In addition, proteins with $\geq$90% of samples below the LOD were excluded from the model.

### 3.3.3 Statistical and machine learning methods

In both cohorts, we defined the primary outcome as the first recurrent ASCVD event, comprising acute myocardial infarction, ischaemic stroke, and cardiovascular death.

In the derivation cohort, we constructed three classification models: first, measured proteins passing quality control (267 proteins) were used to construct a protein-based model with 50 proteins with the highest predictive value. Second, to compare the protein model with current clinical practice, a clinical risk model was constructed and optimized using the same approach as the protein model, including parameters of different validated risk scores such as SMART, Reynolds Risk Score, and Framingham Risk Score.Dorresteijn et al. (2013); Ridker et al. (2008, 2007); Wang et al. (2006) The clinical risk model comprised the following parameters: age, sex, body mass index, systolic blood pressure, total cholesterol, HDL cholesterol, CRP, smoking status, the presence of diabetes, the use of antihypertensive medication, and family history of cardiovascular disease. A third combined model was formed by stacking the clinical risk parameters with the protein parameters. For use in the validation cohort, all three models were recalibrated to allow an equal comparison and avoid miscalibration.Pencina et al. (2012)

All models were constructed using the same machine learning techniques. For the training and evaluation of the models as well as identification of the most reliable biomarker signature in our datasets (both proteomics and clinical), we used stability selection with extreme gradient boosting to predict a binary outcome (event/non-event).Caruana et al. (2004); Chen and Guestrin (2016) The model hyperparameters were selected using a Randomized Grid Search followed by classifier calibration using the Sigmoid method,Niculescu-Mizil and Caruana (2005) both performed on the validation set. To prevent overfitting, 'leave one out cross-validation' was employed on a random subset with half the dimension of the original dataset. For increased confidence, this process was repeated 20 times. This method was coupled with a rigorous stability selection procedure to ensure the reliability and robustness of the obtained parameters. Finally, a permutation test (randomization test) was applied to evaluate the statistical validity of the results,Ojala and Garriga (2010) since standard univariate significance tests cannot be applied to the used models due to the non-linear combination of feature functions.

To further explore the inflammatory pathways involved, we performed additional analyses by dividing the SMART cohort in a high CRP ($>2$ mg/L) and low CRP ($\leq 2$ mg/L) group. Patients with a suspected acute inflammatory episode (CRP $>$ 20 mg/L) were excluded. In both groups, a model comprising 50 proteins was constructed to predict recurrent ASCVD events. Protein-protein association networks were assessed and graphically displayed using STRING v11 (string-db.org).Szklarczyk et al. (2019) Normalized protein expression (NPX) values (relative quantification on log2 scale) for interleukin-6 (IL-6) were compared between high and low CRP groups. To identify high or low CRP-specific proteins, the top 10 proteins from both groups were compared with the overall model.

Model performance was reported by means of discrimination, calibration, and reclassification. Discrimination was assessed using the receiver operating characteristic (ROC) curve with an area under the curve (AUC). Relative protein importance was reported in a bar plot.Hastie, Tibshirani, and Friedman (2009) Calibration plots were constructed to display calibration performance. Reclassification performance was assessed using the category-free net reclassification improvement (NRI $> 0$) and integrated discrimination index (IDI).Pencina et al. (2012) 95% CI were reported using bootstrap intervals for point estimates of performance metrics when asymptotic intervals were not available.

Data are presented as mean ± standard deviation for normally distributed variables or median with interquartile range (IQR) for skewed data. Categorical variables are expressed as absolute numbers and percentages. Independent sample t-tests and Mann-Whitney U-tests were used where appropriate. Two-sided p-values of $\leq$0.05 were considered statistically significant. Data were analysed using Python version 3.7 (www.python.org) and RStudio version 3.6.1 (R Foundation, Vienna, Austria).

## 3.4 Results

Patient characteristics of both the derivation and validation cohort are listed in Table A.2. In the derivation cohort, 263 (30.2%) participants experienced a recurrent ASCVD event during a median follow-up of 8.0 (4.6-12.2) years. The primary recurrent event consisted of myocardial infarction in 48 (5.5%) patients, ischaemic stroke in 105 (12.1%) patients, and 110 (12.6%) patients died of cardiovascular causes. In the validation cohort, 130 (18.6%) participants experienced a recurrent ASCVD event during a median follow-up of 3.0 (2.2-3.1) years. In this cohort, the primary recurrent ASCVD event was a myocardial infarction in 39 (5.6%) patients, whereas 53 (7.5%) patients had an ischaemic stroke and 38 (5.4%) patients died of cardiovascular causes. The final proteomic analysis included 267 unique proteins after exclusion of nine proteins with $\geq$90% of values below the LOD (see the online supplementary material: section 3.7, Table 1).

### 3.4.1 Discriminatory value of proteomic risk model

In the derivation cohort, prediction of recurrent ASCVD events using the protein model resulted in an ROC AUC of 0.810 (95% CI $0.797 - 0.823$; Figure 3.1A and Table 3.1. The proteins with their relative importance are shown in

Figure 3.2. In comparison, the clinical risk model resulted in an ROC AUC of 0.750 (95% CI 0.734 – 0.765; Figure 3.1A and Table 3.1). Combination of both models led to an ROC AUC of 0.824 (95% CI 0.812 – 0.835; Figure 3.1A and Table 3.1). The protein model performed significantly better than the clinical risk model (delta AUC 0.060, 95% CI 0.040 – 0.083, $p < 0.001$), whereas the combination of both models was only slightly superior to the protein model alone (delta AUC 0.014, 95% CI 0.009 – 0.019, $p < 0.001$).

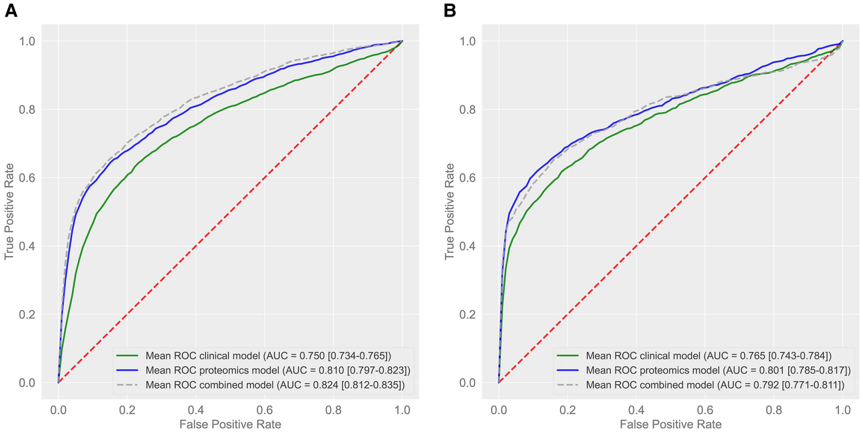


Figure 3.1: Discriminatory value in the derivation and validation cohort. Receiver operating characteristic curve of protein, clinical, and combined model in the derivation cohort (A) and in the validation cohort (B). The 95% confidence interval is shown between brackets. AUC, area under the curve.

After recalibration of all models, the discriminatory value was tested in the validation cohort. Validation of the prediction of recurrent ASCVD events using the protein model resulted in an ROC AUC of 0.801 (95% CI 0.785 – 0.817; Figure 3.1B and Table 2). In comparison, the clinical risk model resulted in an ROC AUC of 0.765 (95% CI 0.743 – 0.784; Figure 3.1B and Table 2). Combination of both models led to an ROC AUC of 0.792 (95% CI 0.771 – 0.811; Figure 3.1B and Table 2). In the validation cohort, the protein model also outperformed the clinical risk model (delta AUC 0.036, 95% CI 0.020 – 0.051, $p < 0.001$), whereas a combination of both models was not superior to the protein model alone (delta AUC −0.007, 95% CI −0.023 to 0.004, $p =$ 0.996).

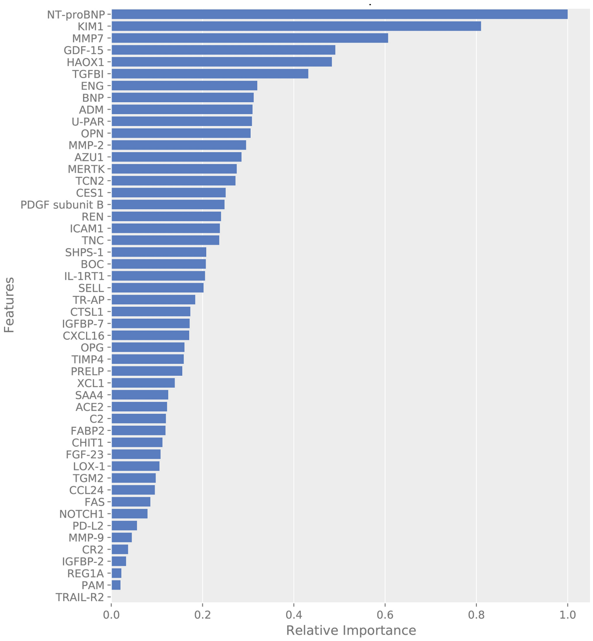


Figure 3.2: Importance plot of the protein model. Importance plot of the proteins in the protein model from the derivation cohort. The importance refers to the extent to which a model relies on a given protein. Shown is the relative importance of the 50 proteins in the model.

### 3.4.2 Calibration and reclassification of the proteomic risk model

The calibration plots of the proteomic, clinical, and combined model for both the derivation cohort and validation cohort (after recalibration) are shown in Figure 3.3. The six models were well calibrated, although risk was slightly underestimated in the highest-risk categories. We calculated the NRI and IDI by comparing the protein model with the clinical risk model (Table 1). In the derivation cohort, the NRI was 0.152 (95% CI 0.110 – 0.196) and the IDI was 0.098 (95% CI 0.073 – 0.122), compared with an NRI of 0.173 (95% CI 0.133 – 0.211) and an IDI of 0.085 (95% CI 0.068 – 0.101) in the validation cohort.

| | **Clinical model** | **Protein model** | **Combined model** |
|---|---|---|---|
| AUC | | | |
| Derivation cohort | 0.750(0.734 – 0.765) | 0.810(0.797 – 0.823) | 0.824(0.812 – 0.835) |
| Validation cohort | 0.765(0.743 – 0.784) | 0.801(0.785 – 0.817) | 0.792(0.771 – 0.811) |
| NRI | | | |
| Derivation cohort | Reference | 0.152(0.110 – 0.196) | 0.174(0.134 – 0.218) |
| Validation cohort | Reference | 0.173(0.133 – 0.211) | 0.146(0.099 – 0.188) |
| IDI | | | |
| Derivation cohort | Reference | 0.098(0.073 – 0.122) | 0.116(0.094 – 0.139) |
| Validation cohort | Reference | 0.085(0.068 – 0.101) | 0.070(0.049 – 0.090) |

Table 3.1: Summary statistics of performance: area under the curve (AUC), net reclassification improvement (NRI), and integrated discrimination index (IDI). 95% confidence interval is shown between parentheses.

### 3.4.3 Predictive value in high and low C-reactive protein subsets

In clinical practice, CRP is used to identify patients with 'residual inflammatory risk'. To evaluate the impact of CRP on the performance of the proteomic panel, we divided patients based on CRP levels in the SMART cohort, resulting in 373 patients classified as low CRP ($\leq$2 mg/L) vs 463 patients classified as high CRP (>2 mg/L). Thirty-four patients with a suspected acute inflammatory episode (CRP $>$ 20 mg/L) were excluded from the analysis. In the low CRP group, 27.3% of patients experienced an ASCVD event during follow-up, compared with 32.0% of patients in the high CRP group ($p$ = 0.13). Interleukin-6 levels were much higher in the high CRP group compared with the low CRP group [NPX (log2 scale) 13.50, IQR 10.24 – 18.45 vs. 8.63, IQR 6.71 – 11.27]. The overview of the network pathway analysis in the high and low CRP group is depicted in the online supplementary material (section 3.7), Figure 1. The high CRP group showed a central role for IL-6, which was not present in the low CRP protein model. Conversely, four different inflammatory proteins, which were neither in the initial model nor in the high CRP group, were identified in the top 10 predicting proteins of the low CRP group (Table 3.2).

| **Overall** | **High CRP subset** | **Low CRP subset** |
|---|---|---|
| NT-proBNP | NT-proBNP | KIM1 |
| KIM1 | HAOX1 | BNP |
| MMP-7 | OPN | ADM |
| GDF-15 | KIM1 | **AMBP** |
| HAOX1 | PSGL-1 | **NID1** |
| TGFBI | GDF-15 | TIMP4 |
| ENG | TIMD4 | FABP2 |
| BNP | MMP-2 | NT-proBNP |
| ADM | CTSL1 | **VASN** |
| U-PAR | XCL1 | **TF** |

Table 3.2: Overview of the 10 most important proteins in the overall group as well as in the high and low CRP groups. Marked bold are proteins not in the overall 50-protein model. CRP, C-reactive protein; NT-proBNP, N-terminal prohormone brain natriuretic peptide; KIM-1, kidney injury molecule 1; MMP-7, matrix metalloproteinase 7; GDF-15, growth/differentiation factor 15; HAOX1, hydroxyacid oxidase 1; TGFBI, transforming growth factor-β-induced protein ig-h3; ENG, endoglin; BNP, brain natriuretic peptide; ADM, adrenomedullin; U-PAR, urokinase plasminogen activator surface receptor; OPN, osteopontin; PSGL-1, P-selectin glycoprotein ligand 1; TIMD4, T-cell immunoglobulin and mucin domain-containing protein 4; MMP-2, matrix metalloproteinase-2; CTSL1, cathepsin L1; XCL1, lymphotactin; AMBP, α1-microglobulin-bikunin precursor; NID1, nidogen-1; TIMP4, metalloproteinase inhibitor 4; FABP2, intestinal-type fatty acid-binding protein; VASN, vasorin; TF, tissue factor.

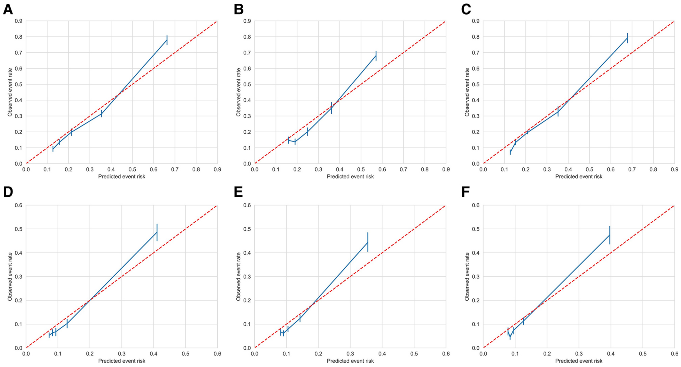


Figure 3.3: Calibration in the derivation and validation cohort. Calibration plots for the protein (A), clinical (B), and combined (C) model in the derivation cohort (SMART) and the protein (D), clinical (E), and combined (F) model in the validation cohort (Athero-Express). Predicted event risk vs. observed event rate per risk category quintiles.

## 3.5 Discussion

Using targeted proteomics in two cohorts comprising 1570 patients with established arterial disease, we show that a panel of 50 proteins is superior to a clinical risk model in predicting recurrent ASCVD events. In both the derivation and the validation cohort, the proteomic model performed better in terms of discrimination, was similarly well calibrated and provided a significant NRI over the clinical risk model (Structured Graphical Abstract). Collectively, these data confirm the potential of improved, proteome-supported risk stratification in a secondary prevention setting.

Atherosclerotic cardiovascular disease risk prediction using clinical characteristics performs relatively poor in terms of discrimination.Kaasenbrood et al. (2016); Jensen (2016) We previously showed that a targeted proteomics panel improves the prediction of ASCVD events in a primary prevention setting.Hoogeveen et al. (2020) Ganz et al.Ganz et al. (2016) illustrated that a nine-protein risk score also predicted recurrent ASCVD events in patients with coronary heart disease with modest discrimination (C-statistic 0.70 in validation). With improved proteomic and machine learning techniques, we now show that the use of proteomics significantly outperforms clinical risk prediction in two large secondary prevention cohorts (AUC of 0.801 in the validation cohort, delta AUC 0.036). Whereas in the highest-risk groups the models tended to underestimate ASCVD recurrence risk, the protein, clinical, and combined models were similarly and well calibrated.

### 3.5.1 Recurrent cardiovascular events: predictive proteins

A targeted proteomics panel was used comprising proteins related to ASCVD, metabolism, and inflammation. N-terminal pro-B-type natriuretic peptide (NT-proBNP), an established marker for heart failure,Wang et al. (2006) was the protein with the strongest predictive value. NT-proBNP was also found among the top proteins predicting primary ASCVD events in an earlier study.Hoogeveen et al. (2020) Kidney injury molecule-1 (KIM-1) was the second most predicting protein, and has been associated with cardiorenal syndrome.Figarska et al. (2018) The top three proteins were completed by matrix metalloproteinase 7 (MMP-7), which was also found in the primary prevention population.Hoogeveen et al. (2020) MMP-7 and its family of matrix metalloproteinases, the main group of enzymes responsible for degradation of the extracellular matrix, are associated with plaque instability, through macrophage-related pathways.Abbas et al. (2014) Lastly, growth differentiation factor 15 (GDF-15), as the top predictive protein in the earlier primary prevention cohort,Hoogeveen et al. (2020) was the fourth most predictive protein in this study. GDF-15 has been shown to play an important role in leucocyte integrin activation after myocardial infarction.Kempf et al. (2011) The other proteins in the panel were primarily related to immune system involvement in atherosclerosis, including chemotaxis, migration, apoptosis, and angiogenesis.Hoogeveen et al. (2020); Bom et al. (2019)

### 3.5.2 Residual inflammatory atherosclerotic cardiovascular disease risk

With respect to the residual inflammatory ASCVD risk, attention has primarily focused on the NLRP3 inflammasome with CRP as a reliable downstream marker.Ridker (2018) In a recent sub-study from low-dose colchicine for secondary prevention of cardiovascular disease (LoDoCo2),Opstal et al. (2020) evaluating the impact of colchicine in secondary prevention, we observed colchicine-induced changes in a panel of 37 inflammatory proteins; the majority of which were, however, unrelated to CRP change. To evaluate the impact of CRP on the performance of a proteomic panel containing multiple inflammatory proteins, we compared the predictive value of our proteomic panel between patients with high

(>2 mg/L) vs. low baseline ($\leq$2 mg/L) CRP. As observed in the online supplementary material (section 3.7), Figure 2, the central protein in the high CRP group, linked to many other crucial proteins in the model, is IL-6 with much higher levels in the high CRP group compared with the low CRP group, substantiating the involvement of the NLRP3-IL6 pathway leading to CRP elevation. To further evaluate a potential role of inflammatory factors in patients with low CRP, we compared the 10 most important proteins in both high and low CRP groups with the overall 50 protein model. In contrast to the top 10 proteins in the high CRP protein model, which were all present in the overall protein model, the top 10 proteins in the low CRP group comprised four proteins not represented in the initial model nor in the high CRP model: $\alpha$1-microglobulin-bikunin precursor (AMBP), nidogen-1 (NID1; also known as entactin), tissue factor (TF), and vasorin (VASN). All four proteins are related to neutrophil signalling, implying a role for pro-inflammatory innate immunity activation in the low CRP group independent from the NLRP3-IL6 inflammasome pathway.Opstal et al. (2020) Thus, $\alpha$1-microglobulin, which is a plasma and tissue protein derived from AMBP, has been shown to inhibit oxidation of LDL through the inhibition of myeloperoxidase (MPO).Cederlund et al. (2015) MPO, abundantly present in neutrophilic granules,Aratani (2018) has been shown to oxidize LDL, aggravating atherogenesis.Delporte et al. (2014) NID-1 (entactin) is a component of basement membranes stimulating neutrophil adhesion and chemotaxis.Senior et al. (1992) Tissue factor has been shown to contribute to thrombosis at the site of plaque rupture via release from neutrophil extracellular traps and is critical in the formation of arterial thrombosis.Stakos et al. (2015) Vasorin directly binds to and attenuates signalling of transforming growth factor beta (TGF$\beta$).Ikeda et al. (2004) TGF$\beta$, which can be produced by infiltrating cells such as neutrophils and macrophages, has been shown to have both atherogenic and atheroprotective properties.Toma and McCaffrey (2012); Grainger (2004) The preponderance of these neutrophil-related proteins in the model best predicting recurrent ASCVD risk in the low CRP group corresponds to our findings in LoDoCo2, where proteins related to neutrophil-activation such as MPO were reduced following colchicine treatment.Opstal et al. (2020) Collectively, these findings imply a residual inflammatory risk also in secondary prevention patients with low CRP, with preliminary evidence pointing to the potential involvement of neutrophil-related pathways.

### 3.5.3 Strengths and limitations

The use of samples of two large, well-defined secondary prevention cohorts has supported a robust proteomic analysis. The use of state-of-the-art machine learning technology allows the discovery of non-linear relationships and interactions between proteins, which would not have been identified with traditional statistical methodology.

Several limitations to our study merit discussion. First, by using targeted proteomics, proteins not included in these panels which also predict recurrent ASCVD events may have been missed. However, the goal of this study was to evaluate the feasibility of a high-throughput, protein-based risk score for clinical use, rather than novel protein discovery. Nevertheless, we cannot exclude that the predictive value of a larger protein panel may be even better. Second, the derivation and validation cohort had selective and different enrolment criteria as well as different event risk distribution, which could complicate extrapolation to other risk groups. In the derivation cohort (SMART), patients were included following a myocardial infarction, ischaemic stroke, or transient ischaemic attack, whereas the patients from the validation cohort (Athero-Express) were included after carotid endarterectomy following an ischaemic stroke or transient ischaemic attack. Remarkably, while included after carotid endarterectomy, the relative proportion of patients with a myocardial infarction was higher in the validation cohort compared with the derivation cohort (30.0% vs. 18.3%). Despite these differences between the cohorts, the protein model performance in the validation cohort was comparable to the derivation cohort after recalibration, suggesting suitability for use in different populations. Yet, both cohorts primarily consisted of subjects from European ancestry, so extrapolation to other ethnicities remains to be determined. Lastly, in the derivation cohort, samples were collected after overnight fasting, in contrast to the validation cohort in which the samples were collected non-fasting.

### 3.5.4 Clinical relevance

Single plasma risk markers have failed to robustly improve ASCVD risk scores to date.Force et al. (2018); Mortensen et al. (2018) Using a panel of 50 proteins, we show a significant improvement in discrimination and clinical value attested by the NRI and IDI in secondary prevention. The introduction of expensive novel therapeutics combined with the large variation in ASCVD recurrence risk in secondary prevention underscores the importance of reliable ASCVD risk stratification, which is essential when adhering to the ‘highest risk—highest benefit' principle determining cost-efficacy of expensive novel medication.Annemans et al. (2018) Routine implementation of a dedicated protein panel on top of clinical risk factors may therefore hold a promise to improve therapeutic decisions in secondary prevention.

C-reactive protein has been validated as a reliable marker of residual inflammatory risk,Ridker (2018) as well as a biomarker predicting therapeutic benefit from anti-inflammatory therapies.Ridker et al. (2017a) Conversely, colchicine treatment was recently reported to markedly reduce the residual ASCVD event rate in post-acute coronary syndrome patients, not selected for CRP elevation,Nidorf et al. (2020) whereas colchicine lowered CRP by only 10%.Opstal et al. (2020) In the present study, we observe a preponderance of neutrophil-related proteins contributing to ASCVD risk prediction in patients with low CRP, implying another potential source of residual inflammatory risk independent of the IL6-CRP pathway.Ridker (2018) Collectively, these data lend further support to target specific pathways identified by proteomic analysis. The use of such a pathway-guided strategy instead of a single biomarker approach warrants prospective trials for further validation.

Propelled by expanding proteomic and machine learning technologies, optimal conditions for a high-throughput proteomic assay are approaching. As opposed to clinical risk scores or risk assessment based on genetic candidate genes,Kessler and Schunkert (2021) proteomic scores may more accurately mirror changes in lifestyle.Williams et al. (2019); Lindsey et al. (2015) The major NRI of ASCVD risk in secondary prevention heralds an important further step towards a tailored therapeutic approach in secondary prevention patients, aimed at introducing the use of effective novel medication in the highest-risk patients in a cost-effective manner.Annemans et al. (2018)

## 3.6 Conclusions

We show that a panel of 50 proteins is superior to a clinical risk model in predicting recurrent ASCVD events. In both the derivation and the validation cohort, the proteomic model performed better in terms of discrimination and provided significant NRI whereas calibration was comparable in comparison to the clinical risk model. In addition, we found involvement of neutrophil-related pathways in the subset of low CRP patients, indicating a residual inflammatory ASCVD risk beyond the traditional NLRP3 pathways. Further, large prospective studies will have to confirm the value of proteome-based risk scores in secondary prevention before routine clinical implementation can be advocated.

## 3.7 Supplementary Material

We refer the reader to the online supplementary material at https://academic.oup.com/eurheartj/advance-article/doi/10.1093/eurheartj/ehac055/6525629

## Acknowledgements

For the SMART cohort, we gratefully acknowledge the contribution of the research nurses; R. van Petersen (data-manager) and the members of the Utrecht Cardiovascular Cohort-Second Manifestations of ARTerial disease-Studygroup (UCC-SMART-Studygroup): F.W. Asselbergs and H.M. Nathoe, Department of Cardiology; G.J. de Borst, Department of Vascular Surgery; M.L. Bots and M.I. Geerlings, Julius Center for Health Sciences and Primary Care; M.H. Emmelot, Department of Geriatrics; P.A. de Jong and T. Leiner, Department of Radiology; A.T. Lely, Department of Obstetrics & Gynecology; N.P. van der Kaaij, Department of Cardiothoracic Surgery; L.J. Kappelle and Y.M. Ruigrok, Department of Neurology; M.C. Verhaar, Department of Nephrology & Hypertension, F.L.J. Visseren (chair) and J. Westerink, Department of Vascular Medicine, University Medical Center Utrecht and Utrecht University.

## Funding

This work was supported by an European Research Area Network on Cardiovascular Diseases (ERA-CVD) grant (ERA-CVD JTC2017) and the CVON-Dutch Heart Foundation (2017–20).

## Conflict of interest

N.S.N. is the co-founder of Lipid Tools. A.L.C. reports consulting fees/lecturing fees from Akcea, Amgen, Amryt, Sanofi, Esperion, Kowa, Novartis, Ionis Pharmaceuticals, Mylan, Menarini, Merck, Recordati, Regeneron Daiichi Sankyo, Genzyme, Aegerion, and Sandoz. W.K. reports advisory board/lecturing fees from Novartis, The Medicines Company, DalCor, Kowa, Amgen, Corvidia, Daiichi-Sankyo, Genentech, Novo Nordisk, Esperion, OMEICOS, Sanofi, and Bristol-Myers Squibb, grants and non-financial support from Abbott, Roche Diagnostics, Beckmann, and Singulex, outside the submitted work. E.S.G.S. reports advisory board/lecturing fees paid to the institution of E.S.G.S. by Amgen, Sanofi, Regeneron, Esperion, and IONIS.

# 4

# A systems biology approach to study the pathophysiology behind progression of heart failure

Wouter Ouwerkerk*, **Joao P. Belo Pereira***, Troy Maasland, J. E. Emmens, Sylwia M. Figarska, Iziah E Sama, J. Tromp, Andrea L. Koekemoer, Christopher P. Nelson, Mintu Nath, Simon P. R. Romaine,, Jamie Timmons, John GF Cleland, Faiez Zannad, Dirk J. van Veldhuisen, Chim C Lang, Leong L Ng, Rudolf A. de boer, Natal van Riel, Max Nieuwdorp, Albert K Groen, Erik Stroes, Aeilko H. Zwinderman, Nilesh J. Samani, Carolyn SP Lam, Evgeni Levin*, Adriaan A Voors*
*These authors contributed equally

Submitted

# Abstract

### Background and Aims

We aimed to uncover pathophysiological pathways associated with prognosis in patients with heart failure using a systems biology-based approach. This approach models this complex disease by integrating clinical phenotypic markers to proteomic, transcriptomic and genetic data.

### Methods and Results

We collectively analyzed 54 clinical phenotype markers, 403 circulating plasma proteins, 36,046 transcript expression levels in whole blood, and 6 million genomic markers in 2,516 patients with heart-failure, from the Systems BIOlogy Study to TAilored Treatment in Chronic Heart Failure study. Patients had a median age of 70 (25e-75e percentiles 61-78), 27% were women and 657 (26%) died during a median follow-up of 21 (25e-75e percentiles 15–27) months. We used machine learning methodology based on stacked generalization framework and gradient boosting algorithms to predict all-cause mortality. Results were validated in an independent cohort of 1,738 patients. Biological pathways were identified using enrichment analyses. With this data a multitude of pathways can be extracted. We focussed on the strongest clinical phenotype from our model: Renal (dys-)function, which was a composite of history of renal disease, renal failure, and eGFR. Biological pathways associated with a reduced eGFR were mostly related to cysteine-type peptidase activity and regulation of endopeptidase activity. Key proteins in these pathways are cysteine protease cathepsins. Key markers on pathways associated with a reduced eGFR were (WFDC2 and TRAIL-R2/TNFRSF10B [protein], TRAJ16 [transcriptomic], and LINC00210 [genomic]),

### Conclusion

Using a multi-modal systems biology approach, we found that the strong association between renal dysfunction and mortality in patients with heart failure was linked to cysteine-type peptidase activity and regulation of endopeptidase activity pathways.

## 4.1 Introduction

The pathophysiology of heart failure is complex and involves multiple biological pathways that play a role in disease progression. Understanding the complex pathophysiology of heart failure might identify new treatment targetsShah et al. (2015); Ahmad et al. (2016); Shah, Katz, and Deo (2014). Capturing the entire breadth of heart failure pathophysiology requires integrating genetic, transcriptomic, proteomics and phenotypic markers, using a systems biology approachKitano (2002). System biology is in essence an approach to better understand the complex system of the human body, from its genetic core and proteins to the presentation into phenotypic characteristics and heart failure outcomes. With recent advancements in bioinformaticsKitano (2002), and high-throughput -omics data; integration and subsequent interpretation of multiple high-dimensional -omic datasets is becoming key to revealing novel biological insights. Previous efforts revealed putative markers related to pathological lipid abundanceParker et al. (2019) and cancerYarden and Pines (2012). Such a deep analysis of heart failure requires an enormous repository of data with robust and reproducible observations that can also be validated in an independent populationBayes-Genis et al. (2020). Here, we present a systems biology approach based on integrating multiple high-dimensional -omics modalities using advanced machine learningDeo (2015), with the aim to identify and validate new pathways associated with prognosis in patients with heart failure.

## 4.2 Methods

### 4.2.1 Patient population and study design

The Systems BIOlogy Study to TAilored Treatment in Chronic Heart Failure (BIOSTAT-CHF) was designed to identify pathophysiological pathways related to heart failure progression using a systems biology approach on multi-omics data. The design and baseline characteristics of this study have been previously reportedVoors et al. (2016). Briefly, BIOSTAT-CHF consists of two independent (index and validation) cohorts. Inclusion criteria were similar in both cohorts. The index cohort consisted of 2,516 patients with worsening signs and/or symptoms of heart failure, included from 69 centers in 11 European countries during 2010–2014. The validation cohort consisted of a comparable cohort of 1,738 patients from six centers in Scotland, UK. Patients were enrolled as in- or out-patient, with a median follow-up in each cohort of 21 (25th and 75th percentiles 15-27]) months. The endpoint of interest for the present study was 1-year all-cause mortality. The study complied with the Declaration of Helsinki and was approved by the participating centres' medical ethics committees. All patients provided written informed consent. BIOSTAT-CHF has a large repository of clinical phenotypic data and data encompassing the central dogma: genomics, transcriptomics, and proteomicsCobb (2017), which is needed to get a comprehensive picture of the entire pathway.

### 4.2.2 Phenotypic (clinical) panel

We collected 54 clinical markers in BIOSTAT-CHF (supplementary data: Phenotypic parameters). Phenotypic data consisted of demographic data (e.g., age, sex, medical history, and co-morbidities) and data derived during physical examination (e.g., body mass index, systolic and diastolic blood pressure, and left ventricular ejection fraction [LVEF]). The estimated glomerular filtration rate (eGFR) is calculated by the CKD-EPI formula: 141 * min(Scr/$\kappa$,1)$\alpha$ * max(Scr/$\kappa$, 1)-1.209 * 0.993Age * 1.018 [if female] * 1.159 [if black], where Scr is serum creatinine (mg/dL), $\kappa$ is 0.7 for females and 0.9 for males, $\alpha$ is -0.329 for females and -0.411 for males, min indicates the minimum of Scr/$\kappa$ or 1, and max indicates the maximum of Scr/$\kappa$ or 1.

### 4.2.3 Protein panel from peripheral blood

We measured 403 serum/plasma biomarkers (supplementary data: Protein listings) from several pathophysiological domains, including markers of inflammation, apoptosis, remodelling, myocyte stress/injury, angiogenesis, endothelial function, and several markers of renal function. The protein biomarker data used for this study have been described in recent papersSantema et al. (2018); Tromp et al. (2018); Ouwerkerk et al. (2018). In brief, the biomarkers included standard biochemical blood parameters (e.g., hemoglobin, hematocrit, blood urea nitrogen, and heart failure-related markers [NT-proBNP and BNP]). In addition, four biomarker panels comprising each of 92 protein biomarkers provided by the Olink Bioscience analysis service (Uppsala, Sweden) were measured. These respective panels were Cardiovascular II (CVD II), CVD III, Immune response, and Oncology II panels (https://www.olink.com). The proteins were profiled using Olink Proseek® Multiplex Inflammatory 96x96 platform. The Proseek® kit uses proximity extension assay technology, whereby oligonucleotide-labelled antibody probe pairs bind to their respective targets. Quantification was achieved using a Fluidigm BioMark™real-time PCR platform. The platform provides normalized protein expression (NPX, log2-normalised), rather than an absolute quantification.

### 4.2.4 Transcriptomic panel

Whole blood transcriptomic profiles from 944 patients (626 survivors and 318 non-survivors who died from cardiovascular causes) from the index cohort were obtained using the GeneChip® Human Transcriptomic Array® 2.0 (HTA 2.0) developed by Affymetrix, Inc. (part of Thermo Fisher Scientific). Patients were age- and sex-matched. Details on the protocols and methodology used to assess and confirm the quality of the raw transcriptomic data, the processes used to integrate signals from individual probes on the array to determine the expression levels of each gene and to assess the quality of the summarized RNA expression set data were previously published [Nath et.al 2021]. In total, 36,046 (17,924 protein-coding and 18,122 non-protein-coding) transcripts were analyzed.

### 4.2.5 Genomic panel

Both cohorts were processed, genotyped, quality controlled and imputed independently, using identical protocols. Genotyping of all patients was performed using the Affymetrix Axiom Genome-Wide UKB WCSG genotyping array. Sample level QC was performed for X chromosome homozygosity (sex mismatch) and identity by descent estimates (relatedness and duplicates). Before imputation, variants were removed if their call rate was <95% for variants with minor allele frequency (MAF) ≥5%, or <99% for variants with MAF <5%, or had a Hardy-Weinberg equilibrium $p < 1 \times 10^{-6}$. Imputation was performed using SHAPEIT2Delaneau, Zagury, and Marchini (2013) and IMPUTE2Howie, Donnelly, and Marchini (2009) with the phase 3 release 1000G reference panelSudmant et al. (2015).

### 4.2.6 Statistical Analyses

We used machine learning methods, particularly gradient boosting (with tailored loss functions), with stacked regularizationWolpert (1992). This method can handle multiple data sources in a non-linear manner by learning how to combine the predictions given by models trained on these different data sources into a single coherent output. Furthermore, it is specifically designed, in contrast to standard modern statistical methods (Supplementary data: Benefits of Machine Learning for Multi-omics Analysis), not only to deal with high-dimensional -omics data, where the number of patients is significantly smaller than the number of variables (n<<p), but also when different data sources are collectively used to estimate the 'core mechanism' present in all data sources. In brief, we used a combination of stacking generalization framework with multiple gradient boosting classifiers to improve prediction accuracy. For each individual -omics panel, we built separate level-0 models. These level-0 models were subsequently combined to form the final, level-1, modelPereira et al. (2020). In this approach we are able to use all available data present in each panel (e.g., all phenotype data from all 2,516 patients in that panel to create the level-0 model of the phenotype panel. The level-0 model of the transcriptomic data was estimated using 944 patients). One of the challenges of using this methodology is tuning the various models' hyperparameters. Typically, each model is optimized separately, leading to a local optimum. To achieve a global optimum, we optimized all the models simultaneously using Bayesian OptimizationFrazier (2018). To avoid over-fitting, we used stratified cross-validation over the training partition. Furthermore, we evaluated the model's quality separately in the validation cohort. We conducted stability selection to ensure the feature signatures' reliability and robustnessMeinshausen and Bühlmann (2010). The complete analysis was repeated multiple times (50x). Receiver-Operating-Characteristics Area-Under-Curves (ROC-AUC) were computed each time and averaged over the repeated analyses in both the index and validation cohort. A permutation (randomization test)Marques et al. (2011) was used to evaluate the results' statistical validity. The validation cohort model did not include the transcriptomic panel and its corresponding level-0 model. Nevertheless, our approach is able to validate the phenotype, protein and genomic panels.

### 4.2.7 3D correlation plots

We used Python v. 3.8 (www.python.org), with packages Numpy, Scipy, and Scikits-learn for implementing the stacking model and R (version 4.0, R Foundation for Statistical Computing, Vienna, Austria, www.r-project.org) for visualizations.

### 4.2.8 Pathway enrichment

In complex diseases, like heart failure, there are often a multitude of pathways involved. In order to identify pathways related to mortality, we therefore performed an over-representation analysis of the most important phenotypic marker and its most correlated parameters from the other (protein, transcriptomic, and genomic) panels. We assessed over-representation with ClueGOBindea et al. (2009) in Gene Ontology biological processes, KEGG, and Reactome pathways using the hypergeometric test and the default Bonferroni step down method for multiple testing corrections (family-wise error rate). We used the whole annotation option as a reference set and reported only biological processes with a corrected p-value $< 0.05$ as significant. Data are presented as means ± standard deviations (SD) when normally distributed, as medians (interquartile range) for skewed variables, and as frequencies (percentage) for categorical variables. Differences between patients who died and those who did not in the index and validation cohort were tested using the Students' independent t-test for continuous normally distributed variables. Differences in variables with a skewed distribution were tested using the Mann-Whitney U test. Categorical variables were tested with Chi-Squared tests. We calculated correlations between (non-)normal continuous, ordinal and binary ranked variables from and between the different panels using Pearson's product-moment correlation coefficient, Kendall rank correlation coefficient, and Spearman's rank correlation coefficient, where appropriate. Our data consist of continuous, categorical, ordinal, and binary variables. Correlations between the variables were statistically tested using the Wilcoxon rank-sum testLaVange and Koch (2006).

## 4.3 Results

### 4.3.1 Clinical characteristics

Data was available from 2,516 patients in the phenotypic and protein panels, 944 in the transcriptomic and 2,470 in the genomic panels in the index cohort (Figure 4.3A). The validation cohort had data available for 1,738 patients in the phenotypic and protein panels and for 1,693 in the genomic panel (Figure 4.3B). During a median follow-up of 21 (25th and 75th percentiles 11–32) and 21 (25th and 75th percentiles 15–27) months, 657 (26%) and 501 (32%) patients died in the index and validation cohorts, respectively. Baseline characteristics of the patients who died and those who survived in the index and validation cohorts are presented in Table 4.1. Patients who died in the index cohort were older (73±11 vs 68±12 years; $p < 0.001$), had a higher NYHA class (NYHA class III/IV 74% vs 58%; $p < 0.001$), and more comorbidities. These differences were similar in the validation cohort (Table

4.1).

### 4.3.2 Multi-omics mortality model

Our final mortality model achieved a significant ROC-AUC value of $0.81 \pm 0.02$ in the stratified cross-validated part of the index cohort and $0.85 \pm 0.03$ in the validation cohort (Figure 4.4), both $p < 0.001$ in permutation tests (Supplementary Figure 4.5A). The final level-1 model consisted of 60 markers per panel with a total of 240 phenotypic, proteomic, transcriptomic and genetic markers, all highly associated with mortality and closely related to each other (Supplementary data: Biomarkers included in the final level-1 model; Supplementary Figure 4.5B). The 15 markers most related to mortality, from each panel incorporated in the final model, are presented in Figure 4.1. Here, the relative importance of each marker in the model is visualized for each panel, we also included the overall importance of each individual marker. The direction of the association between each marker and mortality is presented in the spider plot of Supplementary Figure 4.7. For the phenotypic panel, history of renal disease, and renal disease (defined as an estimated glomerular filtration rate (eGFR) <60 mL/min/1.73m2), were the top most pronounced markers. For subsequent analyses we focused on the most predictive phenotypic marker; renal dysfunction. Renal dysfunction was defined by the combination of history of renal disease (#1 phenotype panel), renal failure (eGFR<60; #2 phenotype panel), and eGFR (#4 phenotype panel). Renal dysfunction functions as important landmark for further analyses, because multiple pathways are involved in heart failure.

### 4.3.3 Correlation network and pathway overrepresentation analysis of strongest phenotypic marker

Next, we constructed a correlation matrix of all proteomic, transcriptomic, and genetic markers (Figure 4.2top; Figure 4.8 Figure 4.2center shows renal dysfunction and its strongest associated markers in the protein panel followed by markers from subsequent -omics panels to which these were most strongly correlated. It must be kept in mind that all these markers are present in our final model focussing on 1-year all-cause mortality. This means that each of these separate markers has a relation with mortality. The mean correlation between renal dysfunction and selected proteins was 0.72(±0.04; all $p < 0.001$). WAP four-disulfide core domain protein 2 (WFDC2) and TNF Receptor Superfamily Member 10b (TRAIL-R2/TNFRSF10B) had the highest correlations (both $r = 0.70$, $p < 0.0001$). The mean correlation between the selected proteins and transcripts was 0.27 (±0.05, all $p < 0.01$). IKZF3 had the highest correlation with WFDC2. ($r = 0.35$, $p < 0.01$). Mean correlation coefficients between markers in transcriptomic and genomic panels were lowest, but still apparent given a correlation of LINC00210 with T cell receptor alpha locus joining 20 (TRAJ20) ($r = 0.53$, $p < 0.01$). Finally, we selected all markers that correlated maximally between each panel, starting with renal dysfunction. This resulted in a list of 29 markers (supplementary data Biomarkers included in the final level-1 model). Pathway over-representation analysis of these markers yielded pathways related to peptidase activity regulation (Table 4.2), or more specific cysteine-type endopeptidase inhibitor activity. The proteins involved in these pathways are presented in Figure 4.6. The proteins primarily involved in the cysteine-type and endopeptidase inhibitor activity pathways were cystatin B (CSTB), galectin-9 (LGALS9), TNFRSF10B, Vascular Endothelial Growth Factor A (VEGFA), and WFDC2. A major group of proteins involved in the cysteine-type and endopeptidase inhibitor activity pathways are cysteine protease cathepsins, this included cystatins (e.g. CSTB) and cathepsins.

## 4.4 Discussion

Using a machine learning systems-biology approach, we identified an accurate model to for 1-year mortality, by combining clinical, proteomic, transcriptomic, and genomic markers across four different data modalities. With this many data, a multitude of pathways could be discovered. This paper focusses on the clinical phenotype with the strongest association to mortality; renal dysfunction. Renal disfunction was chosen because 3 of the top 5 from the phenotypes panel were related to renal dysfunction (#1 history of renal disease, #2 renal failure, and #4 eGFR). We found several disease pathways that might explain the strong association between renal dysfunction and mortality in patients with heart failure. Numerous studies have already established the association between renal dysfunction and mortality in patients with heart failureDamman and Testani (2015); Damman et al. (2014). Importantly, eGFR surpasses other prognostically relevant parameters in heart failure, such as New York Heart Association Class and LVEF, in their strength of association with morbidity and mortalityHillege et al. (2000); Heywood et al. (2007). This strong and consistent association between heart failure and renal dysfunction is often explained by the hemodynamic changes in heart failure leading to impaired renal perfusion, where the kidney was considered a sensitive readout of the severity of heart failure. However, the precise mechanisms underlying this strong interaction between renal failure and mortality in patients with heart failure are incompletely understood. The strong relation between CKD and heart failure is also seen in recent pharmacological treatment options that are now available for both CKD and heart failureMembers: et al. (2022); Group (2021). We know, from the FIGARO, FIDELIO and CREDENCE trials, that treating CKD and DM may also reduce heart failure hospitalizationsPitt et al. (2021); Perkovic et al. (2019); Bakris et al. (2020). Similarly, multiple large heart failure trials (EMPEROR-Pooled, PARADIGM and PARAGON) showed an improvement of both heart failure and renal outcomesZannad et al. (2020); Damman et al. (2018); Seferovic et al. (2017).

### 4.4.1 Proteomics markers of heart failure progression

We revealed several pathways across multiple -omics panels linking renal dysfunction to mortality. From the proteomic panel WFDC2 and TRAIL-R2 connected renal dysfunction to mortality. WFDC2 is a known marker for progression of epithelial ovarian cancer and the expression was associated with progressive renal interstitial fibrosis and renal tubular atrophyNakagawa et al. (2015); Montagnana et al. (2011). It has also previously been identified as a potential biomarker for prediction of both renal dysfunction and heart failure severityYuan and Li (2017); de Boer et al. (2013). Levels of WFDC2 been shown to be associated with age, gender, diabetes, smoking, NT-proBNP, kidney function and HF fibrosis biomarkersPiek et al. (2017) TRAIL-R2 is a cell surface receptor of the TNFRSF10B that binds TRAIL and mediates apoptosis and is also associated with progression of renal dysfunctionCarlsson et al. (2017); Rudnicki et al. (2016).

### 4.4.2 Transcriptomics markers of heart failure progression

From the transcriptomic panel, TRAJ family and SLAMF6 were the most important markers linking renal dysfunction to mortality. These markers are presented on cell-surfaces and are related to an increased immune response. The TRAJ family are all part of the J region of the variable domain of T cell receptor (TCR) beta chain that participates in the antigen recognitionLefranc (2014). It appears that there is a potential relation between a reduction in TCR diversity and worsening renal dysfunction stageCrawford et al. (2018); Wong et al. (2017). SLAMF6 is expressed on Natural killer, T, and B lymphocytes. Clustering of SLAMF6, specifically with the TCR, is needed to augment T cell activationDragovich et al. (2019). TRAJ genes have beneficial and protective qualitiesBrincks et al. (2015). The contribution of the genetic markers to the prediction model of all-cause mortality was relatively low. Compared with some rare diseases that are caused by single-locus mutationsFranz, Müller, and Katus (2001); Towbin and Bowles (2002); Bleumink et al. (2004); Morita, Seidman, and Seidman (2005), the genetic component of common polygenic diseases, like renal dysfunction and heart failure, is thought to involve many genetic variantsBayes-Genis et al. (2020); Igarashi and Somlo (2002); Haas et al. (2017). However, the best genetic predictor of mortality, "rs2894240", (ranked 39 overall) is a mutation located between the Notch homologue protein 4 (NOTCH4) and TSBP1-AS1 genes. Both genes are associated with eGFR and kidney functionWuttke et al. (2019); Hellwege et al. (2019).

### 4.4.3 Multi-level pathways of heart failure progression

The pathway over-representation analysis transcending the individual markers revealed pathways primarily related to regulation of peptidase activity. All individual markers were selected in our machine learning models for predicting 1-year all-cause mortality and correlated highly with renal dysfunction. The over-representation analysis yielded 2 more detailed pathways that are part of the peptidase activity pathway: The cysteine-type peptidase activity and regulation of endopeptidase activity pathways. Cysteine-type peptidase pathway is defined as the catalysis of the hydrolysis of peptide bonds in a polypeptide chain by a mechanism in which the sulfhydryl group of a cysteine residue at the active center acts as a nucleophile. The regulation of endopeptidase activity pathways entails any process that modulates the frequency, rate or extent of endopeptidase activity, the endohydrolysis of peptide bonds within proteinsConsortium (2021); Ashburner et al. (2000). Cysteine protease cathepsins are a group of important proteases that regulate numerous physiological processes. They can be found in lysosomes and endosomes are vital for protein breakdown and major histocompatibility complex (MHC) class II-mediated immune responsesTurk, Turk, and Turk (2001); Turk et al. (2012). Cystatin C is the best-studied member in the cardiovascular system and is known to be associated with renal function and HF-severity, independent of renal functionDamman et al. (2012); Verbree-Willemsen et al. (2020); Chen, Tang, and Zhou (2019). Cathepsins have many functions in the arterial wall and heart and there are various factors known that regulate cathepsin expression in the cardiovascular system e.g., angiotensin II, vascular endothelial growth factor (VEGF) and fibroblast growth factor 2 (FGF2)Liu et al. (2018); Cheng et al. (2012). There is increasing evidence that cathepsins (i.e. Cat-B, -K, and -L)Nakagawa et al.

(1998); Ohashi et al. (2003), contribute direct or indirect, by other chronic inflammatory diseases, to cardiovascular diseases by regulating inflammatory molecule production and immune cell activityOhashi et al. (2003); Santamaría et al. (1998); Joseph et al. (1988); Benavides et al. (2001); Sevenich et al. (2010); Yamada et al. (2010); Ridker et al. (2017a,b); Markousis-Mavrogenis et al. (2019). This systems biology approach furthermore taught us that the strong relationship between renal dysfunction and mortality seems to be, according to the strongest emerging proteins and transcriptomic markers, for a large part explained by immune system activation. This provides us with new information about mechanisms underlying the strong interaction between renal failure and mortality in heart failure other than hemodynamic pathways, which only partly explain its pathophysiologyRangaswami et al. (2019). Our study shows that, without focusing on specific predetermined targets, processes related to T-cell activation and T-cell receptor function seem to be of relevance in explaining the relationship between renal dysfunction and mortality in heart failure. This was also supported by our previous publication where we focused on the transcriptomic markers [Nath et.al 2021]

## 4.5 Future perspectives

The role and opportunities of systems biology in unravelling underlying pathology of complex diseases is attracting increasing attention in the field of in cardiologyTrachana et al. (2018); Leopold and Loscalzo (2018); Weng et al. (2017); Joshi et al. (2021). However, as far as we know, there has not yet been a study using this advanced methodology in such a data-rich cohort. Even outside the field of cardiology this data and the ability of validating the results is quite uniqueReel et al. (2021). This comprehensive picture of markers involved in the pathophysiological disease processes underlying all-cause mortality and, more specifically, renal dysfunction in heart failure, might provide potential future therapeutic intervention targets or markers to monitor disease progression. We identified cathepsins as potential new targets for the treatment of heart failure. There are many reports of important roles of cathepsins in cardiovascular diseases, but also tumor progression, cell death, and immune cell signalingTurk et al. (2012); Vasiljeva et al. (2007); Brömme, Panwar, and Turan (2016); Olson and Joyce (2015); Qin and Shi (2011); Kavčič, Pegan, and Turk (2017); Kramer, Turk, and Turk (2017). We have studied the role of Cat-D in heart failure and found that higher levels of circulating Cat-D were independently associated with all-cause mortality and the composite of all-cause mortality and heart failure hospitalizationHoes et al.

(2020). Cat-D was released by stem cells-derived cardiomyocytes following cardiac stretch in correspondence with troponin T release. Silencing Cat-D resulted in elevated levels of troponin T, especially following induced stress. This suggest that intracellular Cat-D is essential for cardiomyocyte survival, while circulating Cat-D are a measure for disease severity. In addition, the current methodological framework is suitable to identify potential underlying pathophysiologic processes of virtually any clinical phenotype or disease, which offers many opportunities for future endeavours to better understand clinical phenotypes or diseases. The advantages of our framework over other approaches like multi-omics factor analysis, canonical correlation analysis, or more classical bivariate protein-protein interactionsSamet (2006); Roweis and Saul (2000); Witten, Tibshirani, and Hastie (2009); Hotelling (1936); WOLD (1975); Argelaguet et al. (2018); Consortium (2013, 2015) is the ability to use non-linear relationships, prevention of overfitting, and a rigorous stability selection procedure.

### 4.5.1 Limitations

The transcriptomic panel consisted of 944 patients selected from the index cohort and matched on age and sex[Nath et.al 2021]. This is an extensive transcriptomic dataset, but unfortunately, data was measured in a pre-selected group of patients from the index cohort and none from the validation cohort. The selection of patients was not random, but skewed towards cardiovascular mortality. Transcriptomic markers are therefore better suited to predict cardiovascular mortality. This might explain the lower contribution of the markers from the transcriptomic panel in our combined systems biology model. The absence of this panel had no impact on the (level-0) model development and validation of the other panels, because our methods are able to handle changes in data-sources. Also, despite the rigorous selection process, the effects of patient selection cannot be determined. Unfortunately, because of the nature of this study, we are not able to draw causal conclusions on the pathways we found. However, it is apparent that when developing the models for predicting mortality so many markers from all panels are independently selected that are associated with kidney function. We assume that, given all these selected markers related to renal function, points us towards the idea that renal dysfunction in heart failure patients is highly associated with mortality. This association is reflected by the underlying pathways found in this study.

| | **Index** | | | **Validation** | | |
|---|---|---|---|---|---|---|
| | **Alive** | **Died** | **p-value** | **Alive** | **Died** | **p-value** |
| **N** | 1859 (74%) | 657 (26%) | | 1214 (75%) | 401 (25%) | |
| **Age (years)** | 68 (11.9) | 73 (11.2) | <0.0001 | 73 (10.5) | 78 (9.7) | <0.0001 |
| **Sex (Males)** | 1370 (74%) | 476 (72%) | 0.57 | 801 (66%) | 270 (67%) | 0.66 |
| **LVEF (%)** | 31 (9.8) | 32 (12.5) | 0.03 | 41 (13) | 41 (13.3) | 0.63 |
| **BMI (kg/m2)** | 28 (5.5) | 27 (5.5) | 0.001 | 29 (6.4) | 28 (6.1) | <0.0001 |
| **Ischemic heart disease** | 946 (51%) | 412 (63%) | <0.0001 | 776 (64%) | 286 (71%) | 0.008 |
| **H.F. hospitalization (<1y)** | 531 (29%) | 263 (40%) | <0.0001 | 301 (2%5) | 130 (32%) | 0.003 |
| **Myocardial Infarction** | 657 (35%) | 306 (47%) | <0.0001 | 575 (48%) | 223 (56%) | 0.006 |
| **DM** | 577 (31%) | 242 (37%) | 0.007 | 367 (30%) | 155 (39%) | 0.002 |
| **COPD** | 279 (15%) | 157 (234%) | <0.0001 | 191 (16) | 104 (26%) | <0.0001 |
| **History of renal disease** | 402 (22%) | 294 (45%) | <0.0001 | 491 (41%) | 241 (61%) | <0.0001 |
| **NYHA I** | 50 (2.76%) | 6 (0.94%) | <0.0001 | 15 (1) | 1 (0) | <0.0001 |
| **NYHA II** | 711 (39.28%) | 157 (24.69%) | | 575 (47) | 92 (23) | |
| **NYHA III** | 853 (47.13%) | 375 (58.96%) | | 516 (43) | 200 (50) | |
| **NYHA VI** | 196 (10.83%) | 98 (15.41%) | | 108 (9) | 107 (27) | |

Table 4.1: **Baseline demographics index and validation cohort**

| **GO Term** | **P-value (Bonferroni corrected)** |
|---|---|
| **Tissue homeostasis** | 0.0013195 |
| **Epidermis development** | 0.0015873 |
| **Cell chemotaxis** | 0.0010968 |
| **Regulation of peptidase activity** | 0.0018942 |
| **Cysteine-type peptidase activity** | 0.0020925 |
| **Regulation of endopeptidase activity** | 0.0020606 |

Table 4.2: **GO term significance**

## 4.6 Conclusions

The present analysis involved multiple -omics modalities - genomics, transcriptomics, proteomics, and clinical measurements - collected for one of the largest data-rich cohorts of patients with heart failure. We found that chronic kidney disease was the phenotype that had the strongest association with mortality and was associated with pathways related to cysteine-type peptidase activity and regulation of endopeptidase activity pathways. These pathways, and Cat-D in particular, might become potential targets for therapy to decrease mortality in patients with heart failure and chronic kidney disease.

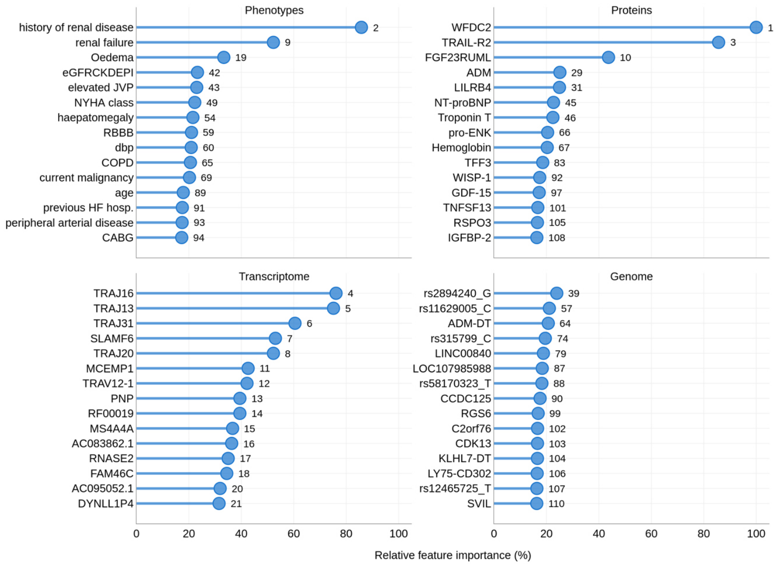


Figure 4.1: **Lollipop plot of the 15 most predictive variables of mortality of each panel** For every panel the top 15 markers are shown in the lollipop plots including their overall ranking. The overall importance ranking was calculated by scaling all relative importance to the importance of each panel.

## 4.7 Tables and Figures

## 4.8 Supplementary Data

### 4.8.1 Benefits of Machine Learning for Multi-Omics Analysis

Modern statistical methods allow building of powerful multivariable predictive models. However, the standard application of these techniques may not be sufficient for estimation of reliable biomarkers in the high-dimensional data. To address this fundamental problem, we followed Stability Selection approach proposed in the seminal paper by Nicolai Meinshausen and Peter

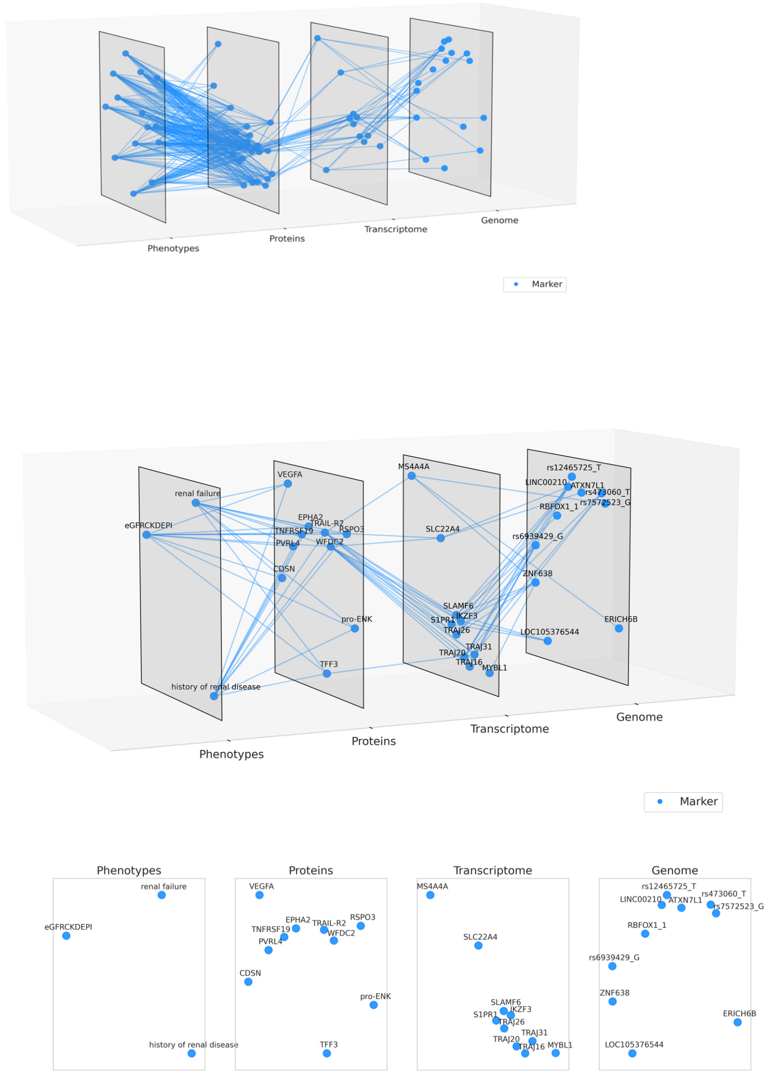


Figure 4.2: **Figure 2 – 3D Correlation plot of the most important clinical markers and its associated markers from other panels.** 3D correlations plot based on our final model, starting with renal dysfunction in the phenotypic panel. The protein biomarkers that correlate maximally with renal dysfunction were selected in the protein panel. Subsequently, the transcriptomic biomarkers that correlate maximally with the protein biomarkers are selected. In turn, we selected genomic markers that correlate maximally with these transcriptomic markers. The distance between markers in a panel is based on multi-dimensional scaling of the distance matrix. A distance matrix is calculated by 1-correlation of each panel.

BühlmannMeinshausen and Bühlmann (2010) to ensure reliability of the findings. We also used a specialized regularization strategy that makes our methodology applicable to the high-dimensional regime. To avoid over-fitting, we used a 10-fold stratified cross-validation over the training partition of the data (80%) while the remaining 20% was used as the testing dataset. For increased confidence, this procedure was repeated multiple times on a completely reshuffled dataset and the average predictive performance reported. For high-dimensional -omics data, application of univariable significance-tests seldom lead to statistically significant results after adjustment for multiple comparisons. These comparisons can easily go into thousands due to large number of variables (phenotypic vs proteins vs transcriptomic vs genes parameters). We applied permutation testing which allowed us to evaluate significance of the "joint panel" of the identified multi-omics markers. We therefore conducted hundreds of re-runs of the model, every time randomly permuting the output variable/clinical heart failure phenotype. This creates a "false" relationship among the multi-omics profiles and outcomes. Performance of the model on large majority of these simulation should be random (AUC 0.5). We evaluated the distribution of all results obtained in these simulations and compared these to our true "joint panel" model. We then computed the statistical significance associated with the "joint panel" model. Our multi-omics machine learning approach provides several advantages over classical techniques:

- Taking into account non-linear relationships
- Extensive prevention of overfitting in different stages (e.g. stratified shuffle split cross-validation and 10-fold cross validation of final model)
- Rigorous Stability Selection procedure with improved prediction
- More reliable biomarkers compared to standard algorithms/univariable methods

### 4.8.2 Correlation analysis

In our analysis, we have encountered both continuous and discrete data types. Consequently, when investigating correlations among various -omics modalities we: 1) continuous-continuous, 2) discrete-discrete and 3) continuous-discrete. For the first two we computed classical statistical correlation coefficients, being Spearman's rank coefficientCorder and Foreman (2014) for continuous-continuous correlations and Kendall's Tau for discrete-discrete correlationsKendall (1938). For every continuous-discrete variable pair we trained a classification model. During the training of every model, the parameters of the model were optimized using a random-search with 5-fold cross-validation. The performance of the models was assessed by ROC AUC scoreBradley (1997), which afterwards was transformed into a score ranging from 0 to 1. Such score forms an absolute pseudo-correlation coefficient which estimates how well the continuous variable can predict the discrete variable.

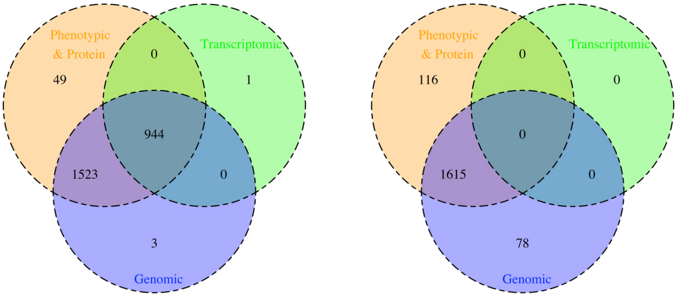


Figure 4.3: **Venn diagram of the number of patients in each panel for left: Index and right: validation cohort**

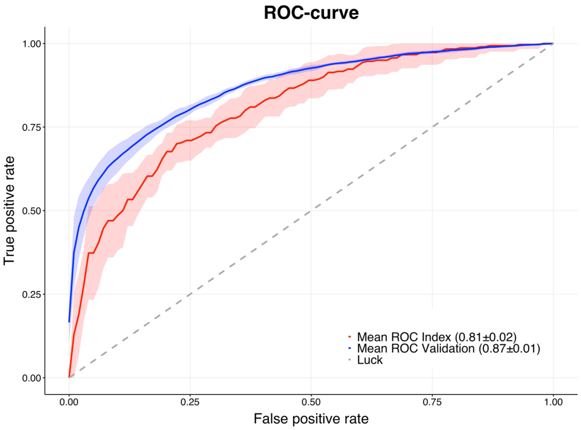


Figure 4.4: **ROC curve with confidence bands** The ROC curves of all cross validated curves are plot with a confidence band for the index and validation cohorts.

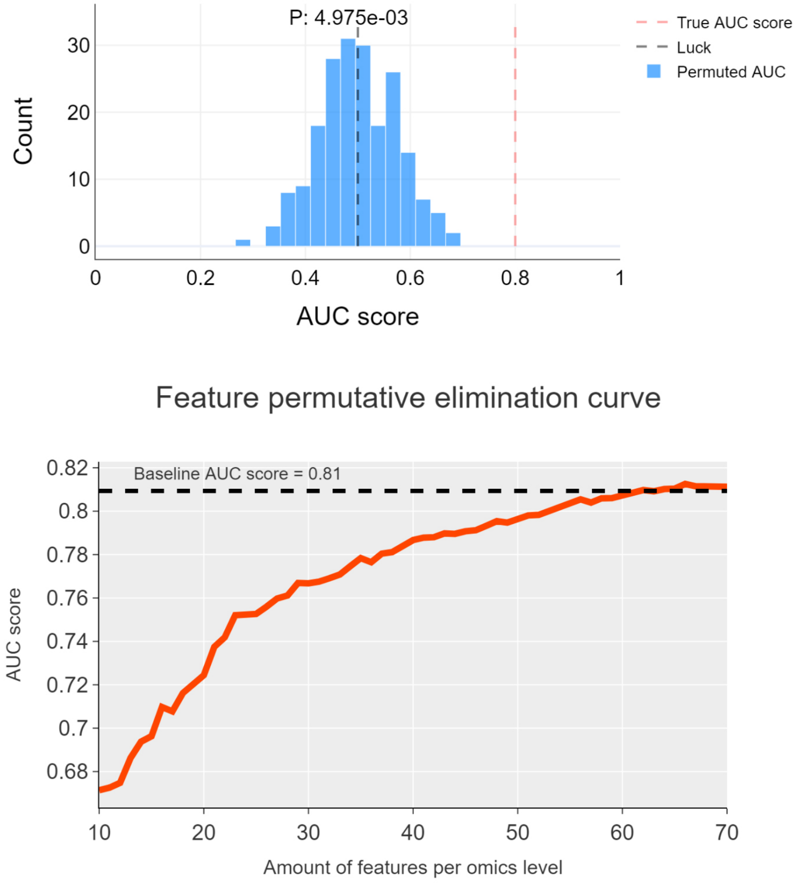


Figure 4.5: **A) Permutation test and B) feature selection** Top: The distribution of ROC AUC values in blue are constructed by permutation in 200 re-runs of the level-1 model. Bottom: The red line is the true ROC AUC of our final level-1 model.

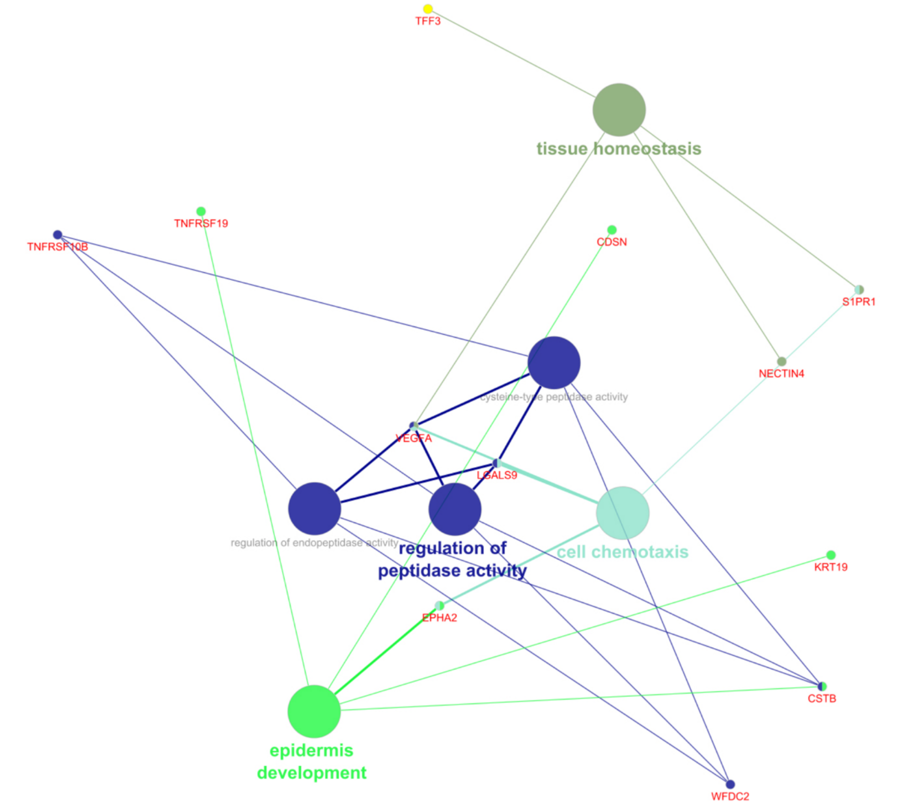


Figure 4.6: **Enrichment network** Cytoscape-ClueGo diagram presenting the renal dysfunction related protein, transcriptomic and genomic markers in red and their relation to the Gene ontology (GO) network. The node size of GO terms represents the enrichment significance. Bonferroni corrected significant terms are in bold.

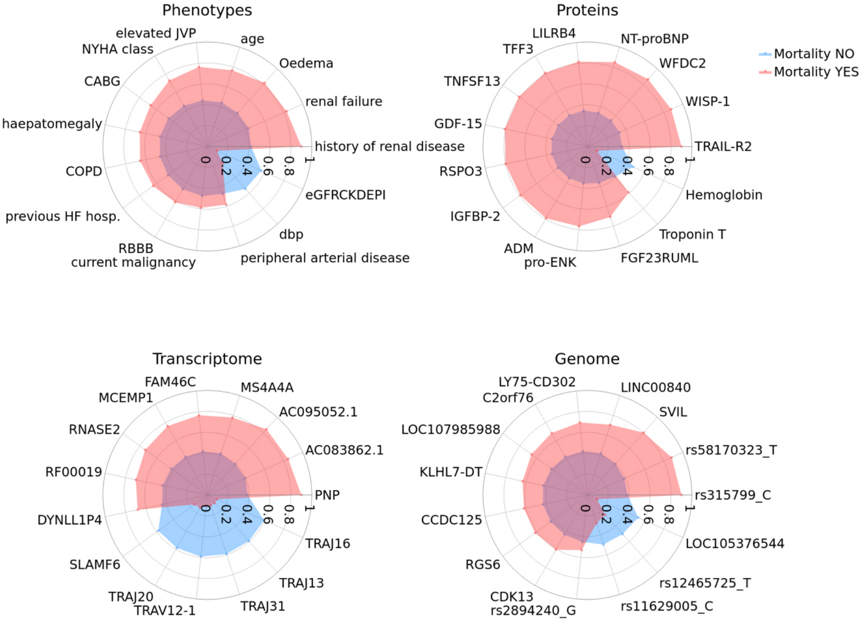


Figure 4.7: **Radar charts** For every panel the differences between the group with fatal outcome (denoted as 'Mortality YES' in the figure) and the group without fatal outcome (denoted as 'Mortality NO') are visualized in a radar chart/spider plot. When a parameter is more frequent in patients who died red is more prevalent or higher, and blue when the parameter is more prevalent or higher in patients who survived.

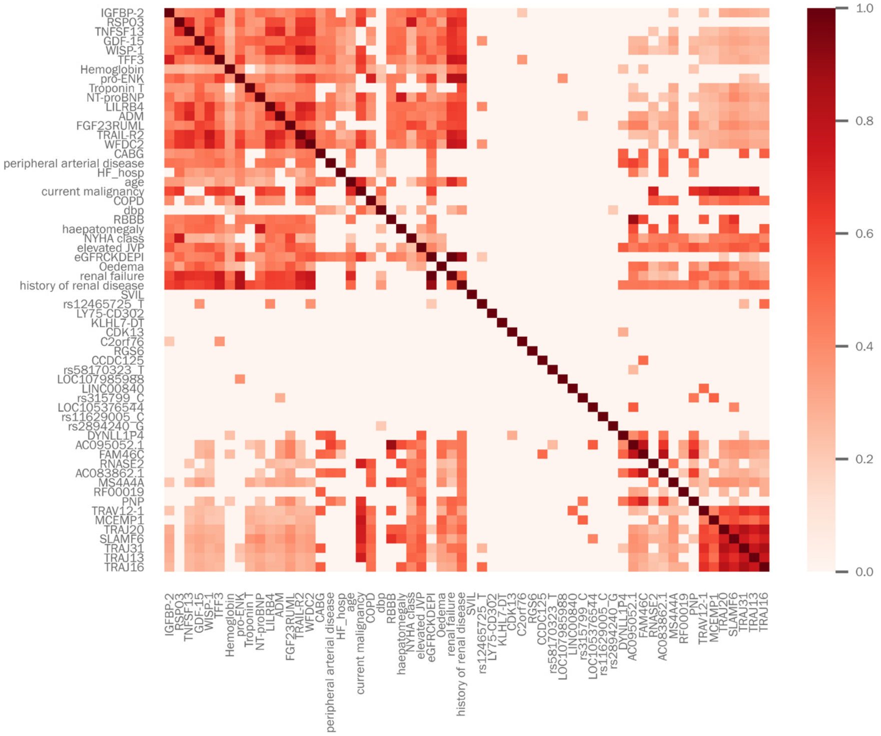


Figure 4.8: **Correlation heatmap** Correlation matrix between the top-60 markers in each of the phenotypic, protein, transcriptomic and genomic panels.

# Part II

# Novel Machine Learning Algorithms for Clinical Research

# 5

# Graph Space Embedding

**João Pereira**, Albert K. Groen, Erik S. G. Stroes, Evgeni Levin

Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19) Main track. Pages 3253-3259.
https://doi.org/10.24963/ijcai.2019/451

## 5.1 Introduction

Learning from interconnected systems can be a particularly difficult task due to the possibly non-linear interaction between the components Linde et al. (2015); Bereau et al. (2018). In some cases, these interactions are known and therefore constitute an important source of prior information Jonschkowski (2015); Zhou et al. (2018). Although prior knowledge can be leveraged in a variety of ways Yu, Simoff, and Jan (2010), most of the research involving interactions, is focused on their discovery. One popular approach to deal with feature interactions, is to cast the interaction network as a graph and then use kernel methods based on graph properties, such as walk-lengths or subgraphs Borgwardt and Kriegel (2005); Shervashidze et al. (2009); Kriege and Mutzel (2012) or, more recently, graph deep convolutional methods Defferrard, Bresson, and Vandergheynst (2016); Fout et al. (2017); Kipf and Welling (2017). In this work however, we focus on the case in which the interactions are feature specific and a universal property of the data instances, which make the pattern search algorithms not suitable for this task. To our knowledge, there is limited research involving this setting, although we suggest many problems can be formulated in the same way (see Figure 5.1). To address this knowledge gap, we present a novel method: *Graph Space Embedding* (GSE), an approach related to the 'random-walk' graph kernel Gärtner, Flach, and Wrobel (2003); Kang, Tong, and Sun (2012) with an important difference: it is not limited to the sum of all walks of a given length, but rather compares similar edges in two different graphs, which results in better expressiveness. Our empirical evaluation demonstrates that GSE leads to an improvement in performance compared to other baseline algorithms when plasma protein measurements and their interactions are used to predict ischaemia in patients with Coronary Artery Disease (CAD) van Nunen et al. (2015); Zimmermann et al. (2015). Moreover, the kernel can be computed in $\mathcal{O}(n^2)$, where $n$ is the number of features, and its hyperparameters efficiently optimized via maximization of the kernel matrix variation.

### 5.1.1 Main Contributions

1. *Graph Space Embedding* function that efficiently maps input into an "interaction-based" space

2. Novel theoretical result on optimal regime for the GSE, namely feasibility region for its parameters

3. *Even Decent Sampling Algorithm*: a strategy to gain insight on which interactions are responsible for the certain prediction

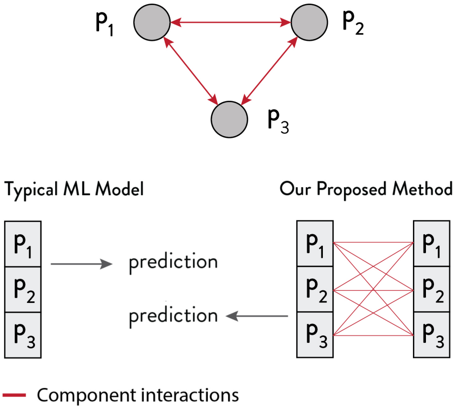


Figure 5.1: A traditional learning algorithm with no structural information will take the feature values and learn to produce a prediction with complete disregard for their interactions (top graph).

## 5.2 Approach

A remark on notation: we will use bold capital letters for matrices, bold letters for arrays and lower case letters for scalars/functions/1-d variables (ex. $\mathbf{X}, \mathbf{x}, x$).

### 5.2.1 Interaction Graphs

Any network can be represented by a graph $\mathcal{G} = \{V, E\}$, where $E$ is a set of edges, $V$ a set of vertices. Denote by $\mathbf{A}_{|V|\times|V|}$ ($|V|$ is equal to the number of features $N$) the adjacency matrix, where $\mathbf{A}_{i,j}$ represents the interaction between feature $i$ and $j$, and whose value is 0 if there is no interaction.

Let $\mathbf{x}_{1\times N}$ be an array with measurements of features 1 to $N$ for a given point in the data. In order to construct an instance-specific matrix, one can weigh the interaction between each pair of features with a function of their values' product:

$$\mathbf{G}_{\mathbf{x}}(\mathbf{A}) = \varphi(\mathbf{A}) \circ \mathbf{x}^{\top}\mathbf{x}, \tag{5.1}$$

where $\varphi(\mathbf{A})$ is some function of the network interaction matrix $\mathbf{A}$, and the operator $\circ$ represents the Hadamard product, i.e. $(\mathbf{A} \circ \mathbf{B})_{i,j} = (\mathbf{A})_{i,j}(\mathbf{B})_{i,j}$.

### 5.2.2 Graph Kernel

Unlike the distance in euclidean geometry, which intuitively represents the length of a line between two points, there is no such tangible metric for graphs. Instead, one has to decide what is a reasonable evaluation for the difference between two graphs in the context of the problem.

A popular approach Gärtner, Flach, and Wrobel (2003) is to compare random walks on both graphs. The $i, j$th entry of the order $k$ power of an adjacency matrix $\mathbf{A}_{|V|\times|V|}$: $\mathbf{A}^k = \underbrace{\mathbf{AA}...\mathbf{A}}_{k\,times}$, corresponds to the number of walks of length $k$ from $i$ to $j$. Any function that maps the data into a feature space $\mathcal{H}$: $\phi : X \to \mathcal{H}$, $k(\mathbf{x}, \mathbf{y}) =< \phi(\mathbf{x}), \phi(\mathbf{y}) >$ is a kernel function. Using the original graph kernel formulation, it is possible to define a kernel that will implicitly map the data into a space where the interactions are incorporated:

$$k_n(\mathbf{G}, \mathbf{G}^{'}) = \sum_{i,j=1}^{n} [\gamma]_{i,j} \left\langle [\mathbf{G}]^i, [\mathbf{G}']^j \right\rangle_F , \tag{5.2}$$

where $\mathbf{G}$ and $\mathbf{G}'$ correspond to $\mathbf{G}_{\mathbf{x}}(\mathbf{A})$ and $\mathbf{G}_{\mathbf{x}'}(\mathbf{A})$ (see eq. 5.1); $\gamma_{i,j}$ is a function that "controls" the mapping $\phi(\cdot)$; and $n$ is the maximum allowed "random walks" length. If $\gamma$ is decomposed into $\mathbf{U\Lambda U}^T$, where $\mathbf{U}$ is a matrix whose columns are the eigenvectors of $\gamma$, and $\mathbf{\Lambda}$ a diagonal matrix with its eigenvalues at each diagonal entry, then equation 5.2 can be re-factored into:

$$k_n(\mathbf{G}, \mathbf{G}^{'}) = \sum_{k,l=1}^{|V|} \sum_{i=1}^{n} \phi_{i,k,l}(\mathbf{G})\phi_{i,k,l}(\mathbf{G}'), \tag{5.3}$$

where $\phi_{i,k,l}(\mathbf{G}) = \sum_{j=1}^{n} [\sqrt{\mathbf{\Lambda}}\mathbf{U}^T]_{i,j}\mathbf{G}^j$. Consequently, different forms of the function $\gamma$ can be chosen, with different interpretations. For the case where $\gamma_{i,j} = \theta^i\theta^j$, which yields:

$$\begin{aligned} k_n(\mathbf{G}, \mathbf{G}') &= \langle \sum_{i=1}^{n} \theta^i [\mathbf{G}]^i, \sum_{j=1}^{n} \theta^j [\mathbf{G}']^j \rangle_F \\ &= \langle \sum_{i=1}^{n} \theta^i [\mathbf{G}]^i, \sum_{i=1}^{n} \theta^i [\mathbf{G}']^i \rangle_F, \end{aligned} \tag{5.4}$$

the kernel entry can be interpreted as an inner product in a space where there is a feature for every node pair $\{k, l\}$, which represents the weighted sum of paths of length 1 to $n$ from $k$ to $l$ ($\phi_{k,l} = \sum_{i=1}^{n} \theta^i \mathbf{G}_{k,l}^i$) Tsivtsivadze et al. (2011). The kernel can then be used with a method that employs the kernel trick, such as support vector machines, kernel PCA or kernel clustering. Another interesting case is when we consider the weighted sum of paths of length 1 to $\infty$. This can be calculated using:

$$k_\infty(\mathbf{G}, \mathbf{G}^{'}) = \langle e^{\beta \mathbf{G}}, e^{\beta \mathbf{G}'} \rangle_F, \tag{5.5}$$

since $e^{\beta\mathbf{G}} = \lim_{n \to +\infty} \sum_{i=0}^{n} \frac{\beta^i}{i!} \mathbf{G}^i$, where $\beta$ is a parameter.

### 5.2.3 Graph Space Embedding

Since we are dealing with a universal interaction matrix for every data point and the interactions are feature specific, it makes sense to compare the same set of edges for every pair of points. As a consequence, we can also avoid solving time-consuming graph structure problems. With these two points in mind, we combined the previous graph kernel methods and the radial basis function (RBF) to develop a new kernel which we will henceforth refer to as Graph Space Embedding (GSE). The radial basis function is defined as:

$$k(\mathbf{x}, \mathbf{y}) = e^{-\frac{||\mathbf{x}-\mathbf{y}||^2}{\sigma^2}} = c\, e^{\frac{2<\mathbf{x},\mathbf{y}>}{\sigma^2}}, \tag{5.6}$$

where $c = e^{-\frac{||\mathbf{x}||^2}{\sigma^2}} e^{-\frac{||\mathbf{y}||^2}{\sigma^2}}$ . The GSE uses the distance $\left\langle \sqrt{\gamma}[\mathbf{G}], \sqrt{\gamma}[\mathbf{G}'] \right\rangle_F$ in the radial basis function:

$$k(\mathbf{G}, \mathbf{G}') = c\, e^{\frac{2<\mathbf{x},\mathbf{y}>}{\sigma^2}} = c \underbrace{\sum_{n=0}^{\infty} \frac{\left(2 \left\langle \sqrt{\gamma}\, \mathbf{G}, \sqrt{\gamma}\, \mathbf{G}' \right\rangle_F \right)^n}{\sigma^{2n}\, n!}}_{r\_w} \tag{5.7}$$

If we then take the upper term of the fraction in $r\_w$ to be $\left[2 \sum_{i=0}^{|E|} \gamma\, \mathbf{G}_i \mathbf{G}_i' \right]^n$, we can use the multinomial theorem to expand each term of the exponential power series, and the expression for the kernel then becomes:

$$k(\mathbf{G}, \mathbf{G}^{'}) = c \sum_{n=0}^{\infty} \underbrace{\left(\frac{2}{\nu}\right)^n}_{\lambda} \underbrace{\sum_{\boldsymbol{\alpha}^n(\cdot)} \frac{\prod_{i=1}^{|E|} [\,\mathbf{G}_i \mathbf{G}_i']^{\alpha_i}}{\prod_{i=1}^{|E|} \Gamma(\alpha_i + 1)}}_{r\_e}, \tag{5.8}$$

where $\Gamma$ is the gamma function, $\mathbf{G}_i \in E$ is the value of edge i in $\mathbf{G}$ and $\nu = \frac{\sigma^2}{\gamma}$. Here, $\boldsymbol{\alpha}^n(\cdot)$ represents a combination of $|E|$ integers: $(\alpha_1, \alpha_2, ..., \alpha_{|E|})$, with $\sum_i^{|E|} \boldsymbol{\alpha}_i^n(\cdot) = n$, and the sum in $r\_e$ is taken over all possible combinations of $\boldsymbol{\alpha}^n(\cdot)$. For instance, for $n = 3$ in a graph with $|E| = 5$, possible examples of $\boldsymbol{\alpha}^3(\cdot)$ include $(0, 1, 1, 1, 0)$ or $(0, 2, 1, 0, 0)$ (see Figure 5.2). We begin by noting

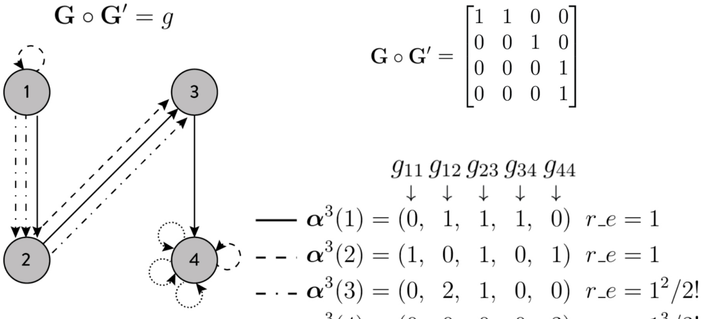

$\mathbf{G} \circ \mathbf{G}' = g$
1
2
3
4
$\mathbf{G} \circ \mathbf{G}' = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$
$g_{11}\ g_{12}\ g_{23}\ g_{34}\ g_{44}$
$\boldsymbol{\alpha}^3(1) = (0, 1, 1, 1, 0)$ $r\_e = 1$
$\boldsymbol{\alpha}^3(2) = (1, 0, 1, 0, 1)$ $r\_e = 1$
$\boldsymbol{\alpha}^3(3) = (0, 2, 1, 0, 0)$ $r\_e = 1^2/2!$
$\boldsymbol{\alpha}^3(4) = (0, 0, 0, 0, 3)$ $r\_e = 1^3/3!$

Figure 5.2: The GSE kernel implicitly compares all edge combinations between $\mathbf{G}$ and $\mathbf{G}'$. In this hypothetical graph, we show a sample of four $\alpha$ combinations for $n = 3$. We denote by $r\_e(\alpha(i))$ the value inside the sum $r\_e$ (see eq. 5.8) corresponding to the combination $\alpha(i)$. Note that while $\alpha(1)$ is a graph walk and $\alpha(2)$ is not, $r\_e(\alpha(1)) = r\_e(\alpha(2))$. However, due to the repetitions in $\alpha(3)$ and $\alpha(4)$, their value is shrunk in relation to the others. The higher the number of repetitions, the more the value shrinks.

that since the sum in $r\_e$ is taken over all combinations $(l, k) \in V \times V$ of size $n$, the GSE then represents a mapping from the input space to a space where all combinations of $n = 0 \to \infty$ edges are compared between $\mathbf{G}$ and $\mathbf{G}^{'}$, walks or otherwise (see fig 5.2). Notice that this is in contrast with the kernel of equation 5.5, where the comparison is between a sum of all possible walks of length $n = 0 \to \infty$ from one node to another in the two graphs.

The GSE also allows repeated edges. However, if the data is normalized so that $\mu(\mathbf{G}_i) \simeq 0, \sigma(\mathbf{G}_i) \simeq 1$, then both the power in the numerator and the denominator of $r\_e$ will effectively dampen most combinations with repeated edges, with a higher dampening factor for higher number of repetitions and/or combinations. Even for outlier values, the gamma function will quickly dominate the numerator of $r\_e$. The $\lambda$ factor serves the purpose of shrinking the combinations with higher number of edges for $\nu > 2$. Finally, $\sigma^2$ now serves a dual purpose: the usual one in RBF to control the influence of points in relation to their distance (see equation 5.6), while at the same time controlling how much combinations of increasing order are penalized.

### 5.2.4 $\nu$ Feasibility Region

As discussed in the above section, the hyperparameter $\nu$ controls the shrinking of the contribution of higher order edge combinations. Intuitively, not all values of $\nu$ will yield a proper kernel matrix since too large of a value will leave out too many edge combinations while one too small will saturate the kernel values. This motivates the search for a $\nu$ value feasible operation region, where the kernel incorporates the necessary information for separability. Informally speaking, the kernel entry $k(\mathbf{G}, \mathbf{G}')$ measures the similarity of $\mathbf{G}$ and $\mathbf{G}'$. In case too few/many edge combinations are considered, the variation of the kernel values will be equal to 1. Therefore, we use the variation of the kernel matrix $\sigma^2(\mathbf{K})$ as a proxy to detect if $\nu$ is within acceptable bounds. We shall refer to the ability of the kernel to map the points in the data into separable images $\phi(\mathbf{x})$ as kernel expressiveness.

To determine this region analytically, we find the $\nu_{max}$ that yields the largest kernel variation, and then use the loss function around this value to determine in which direction the value $\nu$ should take for minimal loss.

**Lemma 5.2.1.** $\max_\nu \sigma^2(\mathbf{K}(\nu))$ *can be numerically estimated and is guaranteed to converge with a learning rate* $\alpha \leq \frac{D}{2(D-1)d_{max}}$, *where* $D$ *is the total number of inter graph combinations and* $d_{max}$ *is the largest combination distance.*

*Proof.* The analytical expression for the variance is:

$$\sigma^2(\mathbf{K}(\nu)) = E[\mathbf{K}(\nu)^2] - \underbrace{E[\mathbf{K}(\nu)]^2}_{b} =$$
$$\left(\frac{D-1}{D}\right)\sum_{d=1}^{D} e^{-2\nu d} - \frac{1}{D^2}\sum_{i \neq j}^{D^2-D} 2e^{-\nu(d_i+d_j)} \ , \tag{5.9}$$

where we used the binomial theorem to expand $b$, and $d = ||\mathbf{G} - \mathbf{G}'||^2$. To guarantee the convergence of numerical methods the function derivative must be Lipschitz continuous:

$$\frac{\|\mathbf{K}'(\nu) - \mathbf{K}'(\nu')\|}{\|\nu - \nu'\|} \leq L(\mathbf{K}') \ : \forall \, \nu, \nu', \tag{5.10}$$

by overloading the notation: $\mathbf{K}'(\nu) = \frac{\partial \sigma^2(\mathbf{K}(\nu))}{\partial \nu}$ to simplify the expression. The left side of equation 5.10 becomes:

$$\begin{aligned} &\frac{\|\top - \Lambda\|}{\|\nu - \nu'\|}, \\ &\Lambda = 2\left(\frac{D-1}{D^2}\right)\left[\sum_{d=1}^{D} d(e^{-2\nu d} - e^{-2\nu' d})\right], \\ &\top = \frac{2}{D^2}\sum_{i\neq j}^{D^2-D}(d_i + d_j)\left(e^{-\nu(d_i+d_j)} - e^{-\nu'(d_i+d_j)}\right). \end{aligned} \tag{5.11}$$

Since $0 \leq e^{-\beta} \leq 1 \, : \forall \beta \in \mathbb{R}$, then:

$$\|\top - \Lambda\| \leq 2\left(\frac{D-1}{D}\right) d_{max}. \tag{5.12}$$

When $\epsilon = \nu - \nu' \to 0$ :

$$e^{-c\nu} - e^{-c\nu'} = \underbrace{\frac{e^{c\nu'} - e^{c\nu}}{e^{c(\nu+\nu')}}}_{\delta} \to 0, \, : \, \nu, \nu' > 0, \tag{5.13}$$

and $\delta$ tends much faster to 0 then $\epsilon$, since the denominator of $\delta$ is the exponential of the sum of $\nu$ and $\nu'$. Thus, the function $k'(\nu)$ is Lipschitz continuous with constant equal to: $L(\mathbf{K}'(\nu)) = 2\left(\frac{D-1}{D}\right) d_{max}$ . □

We shall later demonstrate empirically that $\nu^* = \max_\nu \; \sigma^2(\mathbf{K}(\nu))$ improves the class separability for our dataset.

### 5.2.5 Comparison with Standard Graph Kernels

The original formulation of the graph kernel by Gartner et. al (see eq. 5.2), multiplies sums of random walks of length $i$ from one edge to another ($k \to l$) by sums of random walks $k \to l$ from the other graph being compared of a length not necessarily equal to $i$:

$$\begin{aligned} k_n(\mathbf{G}, \mathbf{G}^{'}) &= \sum_{i,j=1}^{n} [\gamma]_{i,j} \left\langle [\mathbf{G}]_{kl}^{i}, [\mathbf{G}']_{kl}^{j} \right\rangle_F \\ &= \sum_{k,l=1}^{|V|} \sum_{i=1}^{n} [\mathbf{G}]_{kl}^{i} \sum_{j=1}^{n} [\gamma]_{i,j} [\mathbf{G}']_{kl}^{j} \end{aligned} . \tag{5.14}$$

The infinite length random walk formulation (see eq. 5.5) behaves in a similar way. Our method though, always compares the same set of edges in the two graphs.

Another important difference is the complexity of our method versus the random-walk graph kernel. For an $m \times m$ kernel and $n \times n$ graph, the worst-case complexity for a length $k'$ random walk kernel is $\mathcal{O}(m^2 k' n^4)$ and $\mathcal{O}(m^2 k' n^2)$ for dense and sparse graphs, respectively Vishwanathan et al. (2010). The GSE, on the other hand, is always $\mathcal{O}\left(m^2 n^2\right)$ since the heaviest operation is the Frobenius inner product in order to compute the distance between $\mathbf{G}$ and $\mathbf{G}'$. Moreover, once this distance is computed, evaluating the kernel for different values of $\nu$ is $\mathcal{O}(1)$, which combined with the fact that the variance of this kernel is Lipschitz continuous, allows for efficient searching of optimal hyperparameters (see section 5.2.4).

### 5.2.6 Interpretability

How could we better understand what the GSE is doing, when it maps points into an infinite-dimensional space? A successful recent development in explaining black-box models is that of Local Interpretable Model-agnostic Explanations (LIME) Ribeiro, Singh, and Guestrin (2016), where a model is interpreted locally by making slight perturbations in the input and building an interpretable model around the new predictions. We too shall monitor our model's response to changes in the input, but instead of making random perturbations, we will perturb the input in the direction of maximum output change.

Given an instance from the dataset $\mathbf{x}_{1\times N}$, where $N$ is the number of features, and the function that will incorporate the feature connection network $\mathbf{G}_{\mathbf{x}}(\mathbf{A})$ (e.g. $\mathbf{G}_{\mathbf{x}}(\mathbf{A}) = \mathbf{A} \circ \mathbf{x}^{\top}\mathbf{x}$), we will find the direction to which the model is the most sensitive (positive and negative). Unlike optimization, where the goal is to converge as fast as possible, here we are interested in the intermediate steps of the descent. This is because we shall use the set $\mathbf{G} = \{\mathbf{G}_{\mathbf{x}_1}, \mathbf{G}_{\mathbf{x}_2}, ..., \mathbf{G}_{\mathbf{x}_M}\}$ and the black-box model's predictions $\mathbf{f} = \{f(\mathbf{x}_1), f(\mathbf{x}_2), ..., f(\mathbf{x}_M)\}$ to fit our interpretable model $h(\mathbf{G}) \in \mathcal{H}$ (where $\mathbf{x}_i$ is a variation of the original sample $\mathbf{x}_0$, and $\mathcal{H}$ represents the space of all possible interpretable functions $h$). This way, we will indirectly unveil the interactions that our model is most sensitive to, and show how these impact the predictions. To penalize complex models over simpler ones, we will introduce a function $\Omega(h)$ that measures model complexity. To scale the model complexity term appropriately, we can find a scalar $\theta$ so that the expected value of $\Omega(h)$ is equal to a fraction $\varepsilon$ of the expected value of the loss:

$$\mathbf{E}[\theta\Omega(h)] = \varepsilon\mathbf{E}[\mathcal{L}] \leftrightarrow \theta = \frac{\varepsilon\mathbf{E}[\mathcal{L}]}{\mathbf{E}[\Omega(h)]}. \tag{5.15}$$

Lastly, for highly non-linear models, the larger the input space the more complex the output explanations are likely to be, so we will weigh the sample deviations the same as the original sample $\mathbf{x}_0$ using the model's own similarity measure $k(\mathbf{G}_{\mathbf{x}_i}, \mathbf{G}_{\mathbf{x}_0})$. Putting it all together:

$$\xi(\mathbf{x}_0) = \min_{h \in \mathcal{H}} \mathcal{L}\Big(h, f, k(\mathbf{G}_{\mathbf{x}_i}, \mathbf{G}_{\mathbf{x}_0})\Big) + \theta\Omega(h). \tag{5.16}$$

where $\mathcal{L}\Big(h, f, k(\mathbf{G}_{\mathbf{x}_i}, \mathbf{G}_{\mathbf{x}_0})\Big)$ is the loss of $h$ when using $\mathbf{G}_{\mathbf{x}_i}$ to predict the black-box model output $f(\mathbf{x}_i)$, weighted by the kernel distance to the original sample $k(\mathbf{G}_{\mathbf{x}_i}, \mathbf{G}_{\mathbf{x}_0})$.

### Even Descent Sampling Method

In order to adequately cover the most sensitive regions, we need to take steps with equidistant output values. Thus, we developed a novel adaptive method to sample more in steeper regions and less in flatter ones. The intuition is that we would like to approximate the function values in unexplored regions, so that we choose an appropriate sampling step while considering the uncertainty of the approximation. Due to the stochastic nature of the method, it is able to escape local extremes. Consider the value of function $f$ at a point $\mathbf{x}_0$ and its first order Taylor approximation at an arbitrary point $\mathbf{x}$:

$$f(\mathbf{x}) \approx \hat{f}(\mathbf{x}) = f(\mathbf{x}_0) + \nabla_{\mathbf{x}} f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0). \tag{5.17}$$

The larger the difference $\delta = \mathbf{x} - \mathbf{x}_0$, the less likely it is that the approximation error $f(\mathbf{x}) - \hat{f}(\mathbf{x})$ is small. Assume we would like to model the random variable $F$, which takes the value of 1 if the approximation error is small ($\delta = |\hat{f}(\mathbf{x}) - f(\mathbf{x})| \approx 0$), and 0 otherwise. We will model the probability density function of $F$ as being:

$$p_F(f = 1|\delta) = \lambda e^{-\lambda\delta}. \tag{5.18}$$

Consider also the random variable $T$ which takes the value of 1 if the absolute difference in the output for a point $\mathbf{x}$ exceeds an arbitrary threshold ($|f(\mathbf{x}) - f(\mathbf{x}_0)| > \tau$), and 0 otherwise. Assume there is zero probability this event occurs for sufficiently small steps: $\delta < a(\tau)$, for some value $a(\tau)$. Let us further assume that our confidence that $|f(\mathbf{x}) - f(\mathbf{x}_0)| > \tau$ increases linearly after the value $\delta = a(\tau)$, until the maximum confidence level is reached at $\delta = b$. After some value $\delta = c$, we decide not to make any further assumptions about this event, so we attribute zero probability from that point on. This can be modeled as:

$$p_T(t = 1|\delta) = \begin{cases} \frac{2}{v}\frac{\delta - a(\tau)}{u}\,, & a(\tau) < \delta \le b \\ \frac{2}{v}\,, & b < \delta \le c \\ 0\,, & \text{otherwise} \end{cases}, \tag{5.19}$$

where $v = 2c - a(\tau) - b$, $u = b - a(\tau)$ and $T = 1$, if $|f(\mathbf{x}) - f(\mathbf{x}_0| > \tau$ and 0 otherwise. The distribution of interest is then $p_S = p(f = 1 \cap t = 1|\delta)$. To simplify the calculations, we impose the uncertainty about our approximation (expressed by $F$) and the likelihood of a sufficiently large output difference (expressed by $T$) to be independent given $\delta$: $p(f = 1 \cap t = 1|\delta) = p(t = 1|\delta)p(f = 1|\delta)$, and since the goal is to sample steps from this distribution, we will divide it by the normalization constant: $Z = p(f = 1 \cap t = 1) = \int_{-\infty}^{+\infty} p(f = 1 \cap t = 1|\delta)d\delta$. See Figure 5.3 for an illustration of the method.

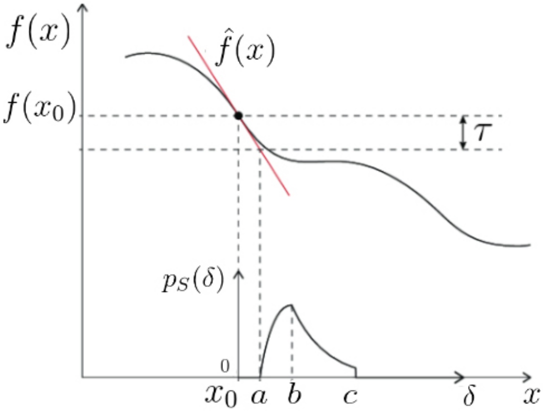


Figure 5.3: Illustration of the even descent sampling. $\hat{f}(x)$ approximates the function $f(x)$ and an estimation of how much $\delta = |x - x_0|$ is required to achieve $|f(x) - f(x_0)| \le \tau$, is computed. Then a sample of $x$ is drawn according to $p_S = p(f = 1 \cap t = 1|\delta)$

There are a couple of properties that can be manipulated for a successful sampling of the output space:

### Controlled Termination

To force the algorithm to terminate after a minimum number of samples $M_{min}$ have been sampled, one can decrease the value of $a(\tau)$ with each iteration so that it becomes increasingly more likely that a value of $\delta$ will be picked such that $|f(\mathbf{x}) - f(\mathbf{x}_0)| < \tau$, terminating the routine. For this purpose, one can compute the estimated threshold value $\tau_0$ that will keep the routine running.

$$|\hat{f}(\mathbf{x}) - f(\mathbf{x}_0)| \geq \tau \Leftrightarrow \sum_{i=1}^{N} \nabla_{\mathbf{x}} f(\mathbf{x}_0)[i]\delta[i] \geq \tau, \tag{5.20}$$

where $N$ is the number of features. This is an underdetermined equation, but one possible trivial solution is to set:

$$\delta[i] \geq \frac{\tau}{N' \ \nabla_{\mathbf{x}} f(\mathbf{x}_0)[i]} \equiv \tau_0, \tag{5.21}$$

where $N'$ is the number of non-zero gradient values, then let $a(\tau)$ decay with time so that it will reach this limit value after $M_{min}$ iterations:

$$a(\tau)_i = \tau_0 \left(1 + \frac{\theta_a(M_{min} - i)}{M_{min}}\right). \tag{5.22}$$

### Escaping Local Extrema

To make it more likely to escape local extrema, one possibility is to set the cut-off value $c$ larger when the norm of $\tau_0$ (eq. 5.21) is larger than its expected value, and smaller otherwise:

$$c = b\left(c_l + \frac{\mathbf{E}\left[||\tau_0||_2\right] - ||\tau_0||_2}{\mathbf{E}\left[||\tau_0||_2\right] + ||\tau_0||_2}\right) \ , \ c_l \in \ ]2, +\infty[\,. \tag{5.23}$$

This formulation allows jumping out from zones where the gradient is locally small, while taking smaller steps where the gradient is larger than expected.

### Termination When Too Far from Original Sample

Since we are trying to explain the model locally, the sampling should terminate when the algorithm is exploring too far from the original sample. For that purpose, one can set $\lambda$ to increase with increasing distance $d$ to the original sample, pushing the probability density towards the left: $\lambda(d) = e^{-\frac{d}{\sigma^2}}$.

Putting all of the above design considerations together, you can find the complete routine in algorithm 1.

**Algorithm 1** Even Descent algorithm

```
Input: $f, \mathbf{x}_0, \mathbf{A}$
Parameter: $\tau, \lambda, \theta_a, b, c_l, M_{min}$
Output: $\mathbf{X}', \mathbf{f}$
$i \leftarrow 0$, $f_i \leftarrow f(\mathbf{x}_0)$, $\mathbf{f} \leftarrow [f_i]$, converged $\leftarrow$ False
$\mathbf{E}[||\tau_0||] = 0$, $\mathbf{X}' \leftarrow [\mathbf{x}_0]$
while converged $\neq$ True do
    $i \leftarrow i + 1$
    $\nabla f \leftarrow$ ComputePartialDers$(\mathbf{x}_0, \mathbf{A}, f)$
    $\tau_0 \leftarrow \tau / |N' * \nabla f|$
    $a, b, c \leftarrow$ UpdatepS$(i, \theta_a, M_{min}, \mathbf{E}[||\tau_0||], c_l)$
    $\mathbf{E}[||\tau_0||] \leftarrow (\mathbf{E}[||\tau_0||](i-1) + ||\tau_0||)/i$
    $\delta \leftarrow$ EvenSample$(\lambda, a, b, c)$
    $\mathbf{x}_i \leftarrow \mathbf{x}_i \pm \ \delta * \nabla f$
    Append$(\mathbf{f}, f(\mathbf{x}_i))$, Append$(\mathbf{X}', \mathbf{x}_i)$
    if $|f_i - f_{i-1}| < \tau$ then
        converged $\leftarrow$ True
    end if
end while
return $\mathbf{X}', \mathbf{f}$
```

# 5.3 Experiments

## 5.3.1 Materials

For all our analysis, we used plasma protein levels of patients with suspected coronary artery disease who were diagnosed for the presence of ischaemia Bom et al. (2019). A total of 332 protein levels were measured using proximity extension arrays Assarsson et al. (2014), and of the 196 patients, 108 were diagnosed with ischaemia. The protein-protein interactions data is available for download at StringDB Jensen, Kuhn et al. (2009). We implemented the GSE and the random walk kernel in python and used sci-kit learn implementation Pedregosa and et al. (2011) for the other algorithms in the comparison.

## 5.3.2 Ischaemia Classification Performance

We benchmarked the GSE performance and running time when predicting ischaemia against the random-walk graph kernel, RBF, and random forests.

Additionally, in order to test the hypothesis that the protein-interaction information is improving the analysis, we also tested GSE using a constant matrix full of ones as the interaction matrix. For this benchmark, we performed a 10-cycle stratified shuffle cross-validation split on the normalized protein data and recorded the average ROC area under the curve (AUC). To speed up the analysis, we used a training set of 90 pre-selected proteins using univariate feature selection with the F-statistic Hira and Gillies (2015). The results are shown in table 5.1. The GSE outperformed all the other compared methods, and the fact that the GSE with a constant matrix (GSE*) had a lower performance increases our confidence that the prior interaction knowledge is beneficial for the analysis. The GSE is also considerably faster than the Random-Walk kernel, as expected. To test how both scale increasing feature size, we compared the running time of both for different pre-selected numbers of proteins. The results are depicted in Figure 5.4.

| **Method** | **AUC std** | **AUC avg** | **Run time avg(s)** |
|---|---|---|---|
| **GSE** | 0.055890 | **0.814141** | 7.63 |
| **RWGK** | 0.051704 | 0.808838 | 1720 |
| **RF** | 0.066036 | 0.764141 | 17.99 |
| **GSE*** | 0.082309 | 0.787879 | 6.59 |
| **RBF** | 0.095247 | 0.779293 | 1.16 |

Table 5.1: The GSE benchmark against random-walk graph kernel (RWGK), random forests (RF), the GSE with constant interaction matrix (GSE*), and radial basis function (RBF). For all kernels, SVM was used as the learning algorithm.

### 5.3.3 Performance for Different $\nu$ Values

Recall from section 5.2.4 that a feasible operating region for the $\nu$ values in the GSE kernel was analytically determined. We wanted to investigate how the loss function performs within this region, and whether it is possible to draw conclusions regarding the GSE kernel behaviour with respect to the interactions. To test this, the $\nu^* = \max_\nu \sigma^2[k(\nu)]$ was found using a gradient descent (ADAM Kingma and Ba (2015)) on the training set over 20 stratified shuffle splits (same preprocessing as in 5.3.2). We then measured the ROC AUC on the validation set using 12 multiples of $\nu^*$. The results can be seen in Figure 5.5. It is quite interesting that our proxy for measuring kernel expressiveness turns out to be a convex function peaking at $\nu^*$.

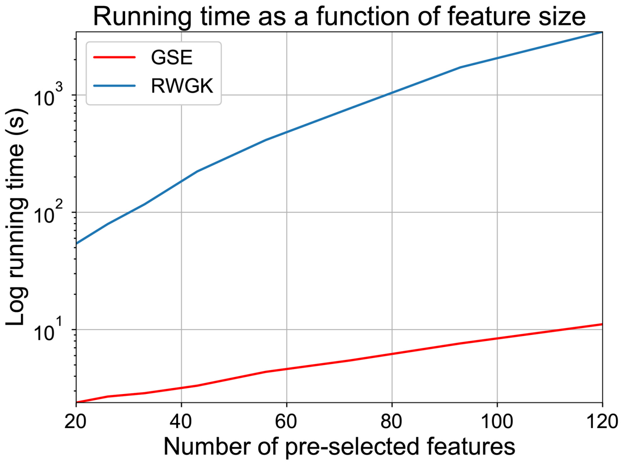


Figure 5.4: Average running time of the GSE and the Random-walk graph kernel (RWGK), per number of pre-selected features

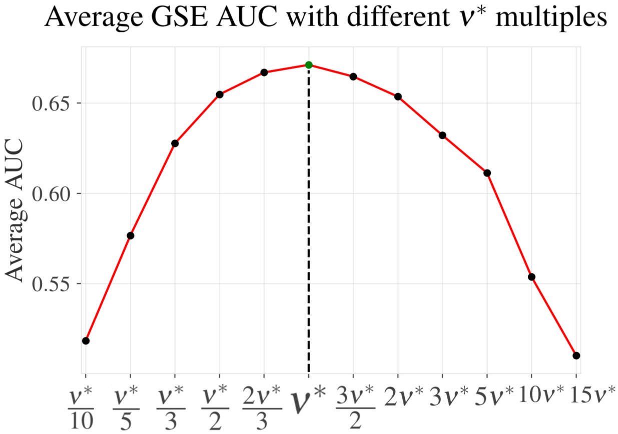


Figure 5.5: Average ROC AUC on validation set using GSE with different $\nu$ values over 20 stratified shuffle splits. Horizontal axis - Multiples of $\max_\nu \sigma^2[k(\nu)]$ here denoted by $\nu^*$. The AUC as function of the $\nu$ values looks convex and peaks exactly at $\nu^*$

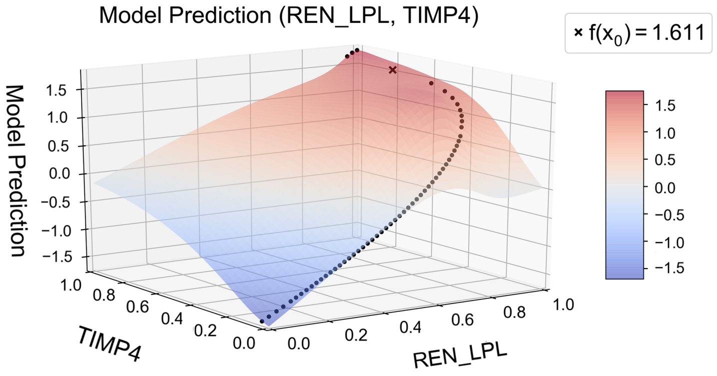


Figure 5.6: Even Descent Sampling for a random patient in our dataset. This analysis reveals our model ”predicts” this patient could be treated by lowering protein ”TIMP4” and the interaction between ”REN” and ”LPL”.

### 5.3.4 Interpretability Test

To test how interpretable our model's predictions are, first we trained the model on a random subset of our data and used the trained model to predict the rest of the data. Then we employed the method described in section 5.2.6 on a random patient in the test set, using decision trees as the interpretable models $h(\mathbf{G}) \in \mathcal{H}$, and a linear weighted combination of *max depth* and *min samples per split* as the complexity penalization term $\Omega(h)$. We then picked the two most important features and made a 3d plot using an interpolation of the prediction space. The result is depicted in Figure 5.6.

The Even Descent Sampling tests instances which are approximately equidistant in the output values. For this patient, our model 'predicts' its ischaemia risk could be mitigated by lowering protein TIMP metallopeptidase inhibitor 4 (”TIMP4”) and the interaction between lipoprotein lipase (”LPL”) and renin (”REN”).

## 5.4 Conclusions

In this paper, we address the problem of analyzing interconnected systems and leveraging the often-known information about how the components interact. To tackle this task, we developed the *Graph Space Embedding* algorithm and compared it to other established methods using a dataset of proteins and their interactions from a clinical cohort to predict ischaemia. The GSE results outperformed the other algorithms in running time and average AUC. Moreover, we presented an optimal regime for the GSE in terms of a feasibility region for its parameters, which vastly decreases the optimization time. Finally, we developed a new technique for interpreting black-box models' decisions, thus making it possible to inspect which features and/or interactions are the most relevant.

# 6

# Manifold Mixing for Stacked Regularization

**João Pereira**, Erik S. G. Stroes, Albert K. Groen, Aeilko H. Zwinderman, Evgeni Levin

In: Cellier, P., Driessens, K. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019. Communications in Computer and Information Science, vol 1167. Springer, Cham. https://doi.org/10.1007/978-3-030-43823-4_36

## 6.1 Introduction

Exponential increase in multi-modal data, stemming from different instruments and measurements presents both an opportunity and a challenge. With a larger sheer volume of information, there is potentially more we can learn for a given process, but coherently combining different data sources with the goal of improving analysis remains a challenging and underdeveloped task. In the medical field for instance, multiple omics data such as proteins or lipids encode somewhat related biological information. Therefore, one might expect that health or disease state is reflected in both of these modalities, despite their different format. Learning frameworks such as manifold alignment Wang and Mahadevan (2011); Cui et al. (2014) and domain adaptation Hajiramezanali et al. (2018); Kumar et al. (2018) may not be directly applicable as they try to find a common latent manifold and learn to transfer knowledge from a source to a target domain, respectively. Orthogonally to existing methods, we present a way of "mixing" information from multiple domains, without imposing hard similarity between them. The motivation is that for a given outcome, the "core mechanism" (e.g. health or disease state) is reflected in all of these modalities and so this commonality can become more evident when the source domain (e.g. proteins) can accordingly transform the local geometry of the target (e.g. lipids).

## 6.2 Approach

We use a stacked regularization setting Wolpert (1992) where each level-one model is trained using "mixed manifolds" of various data modalities. In the next subsections we briefly discuss classical stacked regularization, domain alignment, adaption, and finally propose our mixing algorithm. Regarding notation, we will use capital bold, bold and no formatting for matrices, vectors and scalars or functions, respectively (e.g. $\mathbf{X}, \mathbf{x}, f/W$). We will also use calligraphic font to denote spaces (e.g. $\mathcal{X}$).

### 6.2.1 Stacking

Let $\mathbf{X}$ be a dataset of $N$ samples whose values are sampled from an input space $\mathcal{X} = \{\mathcal{X}^1, \mathcal{X}^2, ..., \mathcal{X}^M\}$ where $\mathcal{X}^1$ to $\mathcal{X}^M$ are subspaces corresponding to different "sources" or "views" $1 \to M$ which we will refer to as "domains". Denote by $y$ the output sampled from an output space $\mathcal{Y}$. In a supervised setting, the goal is to compute $p\left(y|\mathbf{x}^1, ..., \mathbf{x}^M\right)$, where $\mathbf{x}^i$ are the coordinates of an instance from $\mathbf{X}$ in the domain $\mathcal{X}^i$. In stacked regularization, or stacking, the input is passed to a first layer of $W_0$ predictors $g_1^0(\mathbf{x})$, ..., $g_{W_0}^0(\mathbf{x})$, with:

$$g_i^0(\mathbf{x}) = p\left(y|\mathbf{x}^1,\, \ldots,\, \mathbf{x}^M,\, \boldsymbol{\theta}_i^0\right) \ , \tag{6.1}$$

where $\boldsymbol{\theta}_i^0$ are the hyperparameters of the $i$th model. For our task, we suggest to pass one data source per model: $g_i^0(\mathbf{x}) = p\left(y|\mathbf{x}^i,\, \boldsymbol{\theta}_i^0\right)$, so that the width of the first layer $W_0$ is equal to the number of domains $M$. The output from this layer is then passed to one or more layers of $W_k$ models $g_1^k$, ... , $g_{W_k}^k$ which blend the outputs of the previous ones:

$$g_i^k(\mathbf{x}) = p\left(y|g_1^{k-1}(\cdot),\, \ldots,\, g_{W_k}^{k-1}(\cdot)),\, \boldsymbol{\theta}_i^k\right),\, k \in [1, L] \ , \tag{6.2}$$

where $L$ is the total number of blending layers and $\boldsymbol{\theta}_i^k$ the hyperparameters of $i$th model from the $k$th layer. The last blending layer is then passed to a final model $f$ that produces the output $f(\mathbf{x}) = p\left(y|g_1^L(\mathbf{x}),\, \ldots,\, g_{W_L}^L(\mathbf{x}),\, \boldsymbol{\theta}^{L+1}\right)$, where $\boldsymbol{\theta}^{L+1}$ are the hyperparameters of $f$. You can visualize the stacked model general architecture in figure 6.1. From a frequentist point of view, the goal of stacking is then to find:

$$\underset{\boldsymbol{\theta}}{argmin}\ \mathcal{L}(\mathbf{y}; f(\mathbf{X}), \boldsymbol{\theta}) \ , \tag{6.3}$$

where $\boldsymbol{\theta}$ is the set of hyperparameter values from all of the stack models, $\mathbf{y}$ is the output for all of the data, and $\mathcal{L}$ is the loss function when using $f(\mathbf{x})$ to predict $\mathbf{y}$. For a fully Bayesian approach, one should compute the posterior probability of each model by integrating out the hyperparameter values:

$$\begin{gathered} p\left(g_i^k|Z\right) \propto p\left(Z|g_i^k\right) p\left(g_i^k\right) \propto \\ p\left(g_i^k\right) \int p\left(Z|\boldsymbol{\theta}_i^k, g_i^k\right) p\left(\boldsymbol{\theta}_i^k|g_i^k\right) d\boldsymbol{\theta}_i^k \ , \end{gathered} \tag{6.4}$$

where $Z$ is the complete dataset $(\mathbf{X}, \mathbf{y})$.

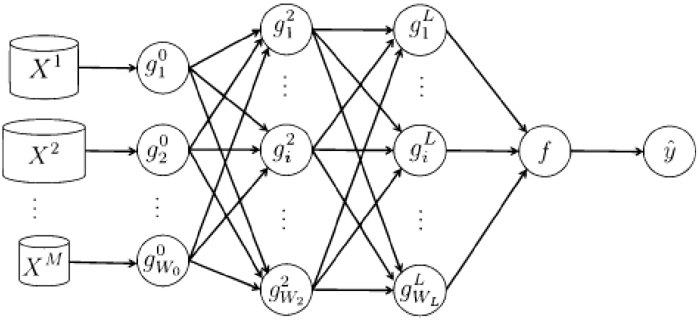


Figure 6.1: Our proposed stacked setting

Although this approach is attractive because it considers the uncertainty of the model, it also incurs high computational cost for large $\boldsymbol{\theta}_i^k$.
Two important aspects are: a) optimizing each model independently does not guarantee finding the global optimal stacked model and b) there is an implicit assumption that each model $g_i^k$ can learn/handle data from different sources (possibly with different formats) effectively.

### 6.2.2 Stacking Optimization

Finding an optimal stacked model can be done by optimizing each sub-model individually or by jointly optimizing all the sub-models. Optimizing each model individually has an important complexity advantage because the number of possible $\boldsymbol{\theta}$ combinations increases exponentially with the number of parameters: $k^{|\boldsymbol{\theta}|}$, where $k$ is the number of values considered for each parameter.

**Lemma 6.2.1.** *For a given dataset* $\mathbf{X}, \mathbf{y}$*, stacked model* $f(x)$ *and parameters* $\boldsymbol{\theta}$*, the following relation is true:* $\mathcal{L}\left(\mathbf{y}, f(x), \boldsymbol{\theta}^*\right) \leq \mathcal{L}\left(\mathbf{y}, f(x), \boldsymbol{\theta}'\right)$*, where*

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{argmin}\ \mathcal{L}\left(\mathbf{y}, f(x), \boldsymbol{\theta}\right),\ \boldsymbol{\theta}' = \varphi_{k=1}^{L}\left(\varphi_{i=0}^{W_k}\left(\underset{\boldsymbol{\theta}_i^k}{argmin}\ \mathcal{L}\left(\mathbf{y}, g_i^k(x), \boldsymbol{\theta}_i^k\right)\right)\right),$$

*and* $\varphi_{j=1}^{L}\left(\underset{\boldsymbol{\theta}_j}{argmin}\, f(\boldsymbol{\theta_j})\right)$ *is a sequential composition of minimizations with respect to index j:* $\underset{\boldsymbol{\theta}_{j=L}}{argmin}\left(\underset{\boldsymbol{\theta}_{j=L-1}}{argmin}\left(...\underset{\boldsymbol{\theta}_{j=1}}{argmin}\left(f(\boldsymbol{\theta_j})\right)\right)\right)$.

*Proof.* Let $\mu$ be a measure on the measurable space $(\Theta, \boldsymbol{\theta})$. Since $\boldsymbol{\theta}$ is a disjoint set, its measure is just:

$$\mu(\boldsymbol{\theta}) = \prod_{k=0}^{L+1} \prod_{i=1}^{W_k} \mu(\boldsymbol{\theta}_i^k) \tag{6.5}$$

Denote by $\{\boldsymbol{\theta}_i^k\}^*$ the set of values that satisfy $\underset{\boldsymbol{\theta}_i^k}{argmin}\ \mathcal{L}\left(\mathbf{y}, g_i^k(x), \boldsymbol{\theta}_i^k\right)$, and $\{\boldsymbol{\theta}_i^k\}$ the set of values $\boldsymbol{\theta}_i^k$ can take. Since $\{\boldsymbol{\theta}_i^k\}^* \subset \{\boldsymbol{\theta}_i^k\}$, then $\mu\left(\{\boldsymbol{\theta}_i^k\}^*\right) \leq \mu\left(\{\boldsymbol{\theta}_i^k\}\right)$, $\forall\, i, k$, and so:

$$\mu\left(\boldsymbol{\theta}'\right) = \prod_{k=0}^{L+1} \prod_{i=1}^{W_k} \mu\left(\{\boldsymbol{\theta}_i^k\}^*\right) \leq \prod_{k=0}^{L+1} \prod_{i=1}^{W_k} \mu\left(\{\boldsymbol{\theta}_i^k\}\right) \tag{6.6}$$

$\mathcal{L}\left(\mathbf{y}, f(x), \boldsymbol{\theta}^*\right)$ is thus optimizing over a larger domain than $\mathcal{L}\left(\mathbf{y}, f(x), \boldsymbol{\theta}'\right)$ is, yielding $\mathcal{L}\left(\mathbf{y}, f(x), \boldsymbol{\theta}^*\right) \leq \mathcal{L}\left(\mathbf{y}, f(x), \boldsymbol{\theta}'\right)$

□

There is a trade-off between complexity and performance when it comes to optimizing the model. If performance is the goal, then the next step is to decide what form is the optimization going to take. A grid-search would quickly become unfeasible for models with multiple hyperparameters, so an attractive solution is to instead use Bayesian optimization Acerbi and Ji (2017).

### 6.2.3 Domain Alignment

Note that at this point there is no information sharing between the first layer's models. However, in many situations it may be desirable that some information *is* shared across these models since they are build using different modalities of the same sample set. The motivation is that even though the samples come from different distributions, the generating processes should be similar and thus they should lie in a similar low-dimensional manifold. This is the central problem of *Manifold Alignment* Hun, D., and K. (2003). Our *Manifold Mixing* is based on similar motivation, with crucial difference - we consider that each domain has a contribution of its own, and therefore we will not enforce an exact match between the manifolds but merely a transformation of the local inter-sample geometry using all the domains, indirectly linking the stacked first layer models together.

### 6.2.4 Domain Adaptation

Given two domains $\mathcal{X}^s$ and $\mathcal{X}^t$, that are different but related, the goal of *Domain Adaptation* is that of learning to transfer knowledge acquired from the source $\mathcal{X}^s$ to the target $\mathcal{X}^t$. The most common setting is when there are many labeled examples in the source, but not in the target, and therefore one tries to learn an estimator $h$ such that it minimizes the error on both the source and target distribution prediction Ben-David et al. (2010). In our setting, the source and target domains represent different modalities of the same sample. For clarity, we will use source and target domain definitions as well, and use the former to transform the latter.

### 6.2.5 Manifold Mixing

We would like to address the question: how to combine data from different domains with a similar relation to the output? Our approach consists in creating a map between each pair of domains $\mathcal{X}^s \to \mathcal{X}^t$, while deforming the local geometry of the two to become more similar. We drew inspiration from LLE Roweis and Saul (2000) in that we will also use the neighbours of a point to predict its position. In our case, we will use the neighbours of this point in the other domains to predict its position in the original domain. Consider a set of points $S$ and two mappings taking the points in $S$ to two coordinate systems of domains $\mathcal{X}^t$ and $\mathcal{X}^s$: $\varphi : S \to \mathbb{R}^{|t|}$, $\psi : S \to \mathbb{R}^{|s|}$, and suppose the subsets $\mathbf{X}^t$, $\mathbf{X}^s$ of the dataset $\mathbf{X}$ are measured in these coordinate systems. Let us introduce an approximation $\mathbf{L}^t_s$ to the mapping $\varphi \circ \psi^{-1} : \mathbb{R}^{|s|} \to \mathbb{R}^{|t|}$ from the coordinates of domain $\mathcal{X}^s$ to the coordinates of domain $\mathcal{X}^t$: $\min\limits_{\mathbf{L}^t_s} \sum_{i=1}^N ||\mathbf{x}^t_i - \mathbf{L}^t_s \mathbf{x}^s_i||^2$, with $\mathbf{x}^t_i, \mathbf{x}^s_i$ corresponding to the $i$th entry of $\mathbf{X}^t$ and $\mathbf{X}^s$, respectively. The optimal solution is then given by:

$$\begin{gathered}\frac{\partial}{\partial \mathbf{L}^t_s} \sum_{i=1}^N ||\mathbf{x}^t_i - \mathbf{L}^t_s \mathbf{x}^s_i||^2 = 0 \Leftrightarrow \sum_{i=1}^N \mathbf{L}^t_s \left(\mathbf{x}^s_i \mathbf{x}^{s\intercal}_i\right) = \sum_{i=1}^N \mathbf{x}^t_i \mathbf{x}^{s\intercal}_i \\ \mathbf{L}^t_s \left(\mathbf{X}^s \mathbf{X}^{s\intercal}\right) = \mathbf{X}^t \mathbf{X}^{s\intercal} \Leftrightarrow \mathbf{L}^t_s = \mathbf{X}^t \mathbf{X}^{s\intercal} \left(\mathbf{X}^s \mathbf{X}^{s\intercal}\right)^{-1} \quad .\end{gathered} \tag{6.7}$$

Denote by $\mathbf{n}^t_i[j]$ the $j$th neighbour of instance $\mathbf{x}_i$ in the domain $\mathcal{X}^t$. Let the array of the points in $\mathcal{X}^s$ which are the neighbours of instance $\mathbf{x}^t_i$ in the domain $\mathcal{X}^t$ be: $\mathbf{N}^{s \leftarrow t}_i = \left[\mathbf{x}^s_{\mathbf{n}^t_i[1]}, \mathbf{x}^s_{\mathbf{n}^t_i[2]}, \dots, \mathbf{x}^s_{\mathbf{n}^t_i[k]}\right]$. Our goal is to 'mix' information from different manifolds. This is accomplished by projecting the neighbours of $\mathbf{x}^t_i$ from the source to the target domain and then finding the linear combination of the points that best reconstructs $\mathbf{x}^t_i$ in the original domain:

$$\min_{\mathbf{w}_i} \sum_i ||\mathbf{x}_i^t - \tilde{\mathbf{x}}_i^{t\leftarrow s}||^2 = \min_{\mathbf{w}_i} \sum_i ||\mathbf{x}_i^t - L_s^t \mathbf{N}_i^{s\leftarrow t} \mathbf{w}_i||^2, \tag{6.8}$$

where $\tilde{\mathbf{x}}_i^{t\leftarrow s}$ is the reconstruction of $\mathbf{x}_i^t$ using domain $\mathcal{X}^s$. We visualize how substituting $\mathbf{x}_i$ by $\tilde{\mathbf{x}}_i^{t\leftarrow s}$ might affect the target manifold in figure 6.2. After setting the derivative w.r.t. $\mathbf{w}_i$ to zero, the optimal solution corresponds to:

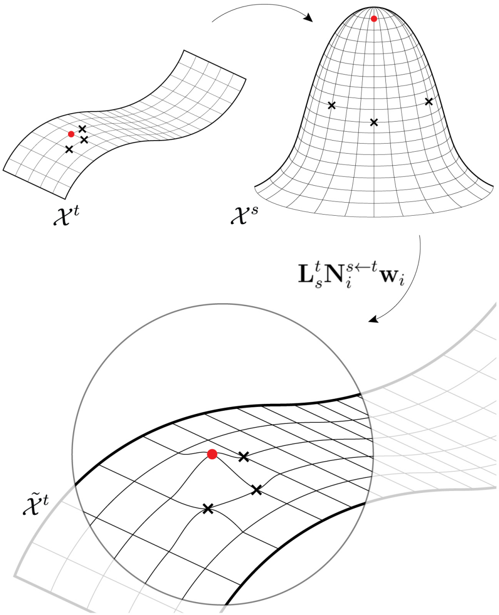


Figure 6.2: Target manifold being deformed by the source manifold using the manifold mixing method. The crosses are the neighbours of point $\mathbf{x}_i$ (point in red) in the target domain. These neighbours are mapped from the source to the target domain and then used to locate $\mathbf{x}_i$. This causes the target manifold to be locally deformed by the source manifold.

$$\mathbf{w}_i = \left( \left( \tilde{\mathbf{N}}_i^{t\leftarrow s} \right)^\intercal \tilde{\mathbf{N}}_i^{t\leftarrow s} \right)^{-1} \left( \tilde{\mathbf{N}}_i^{t\leftarrow s} \right)^\intercal \mathbf{x}_i^t \ , \tag{6.9}$$

where $\tilde{\mathbf{N}}_i^{t\leftarrow s} = \mathbf{L}_s^t \mathbf{N}^{s\leftarrow t}$, the neighbors of $\mathbf{x}_i$ in $\mathcal{X}_t$ projected from their coordinates in $\mathcal{X}_s$ back to the coordinates in $\mathcal{X}_t$. We can now transform the

**Algorithm 2** Manifold Mixing Algorithm

```
Input: data $\mathbf{X} = [\mathbf{X}^1, \ldots, \mathbf{X}^M]$, domain weights $\beta$
Output: transformed data $\tilde{\mathbf{X}}$
for $t = 1$ to $M$ do
    $\mathbf{n}^t \leftarrow \text{NearestNeighbours}(\mathbf{X}^t, k)$
    $\tilde{\mathbf{X}}^t \leftarrow \beta_t \mathbf{X}^t$
    for $s$ in $m \in [1, M] \setminus t$ do
        $\mathbf{L}_s^t \leftarrow \mathbf{X}^t \mathbf{X}^{s\intercal} \left(\mathbf{X}^s \mathbf{X}^{s\intercal}\right)^{-1}$
        for $\mathbf{x}_i = 1$ to $N$ do
            $\mathbf{N}_i^{s\leftarrow t} \leftarrow \mathbf{X}^s[\mathbf{n}_i^t]$, $\tilde{\mathbf{N}}_i^{t\leftarrow s} \leftarrow \mathbf{L}_s^t \mathbf{N}_i^{s\leftarrow t}$
            $\mathbf{w}_i \leftarrow \left(\left(\tilde{\mathbf{N}}_i^{t\leftarrow s}\right)^\intercal \tilde{\mathbf{N}}_i^{t\leftarrow s}\right)^{-1} \left(\tilde{\mathbf{N}}_i^{t\leftarrow s}\right)^\intercal \mathbf{x}_i^t$
            $\tilde{\mathbf{X}}_i^t \mathrel{+}= \beta_s \tilde{\mathbf{N}}_i^{t\leftarrow s} \mathbf{w}_i$
        end for
    end for
end for
$\tilde{\mathbf{X}} \leftarrow [\tilde{\mathbf{X}}_1, , ..., , \tilde{\mathbf{X}}_M]$
return: $\tilde{\mathbf{X}}$
```

original space $\mathcal{X}^t$ into the space reconstructed from the other domains $\tilde{\mathcal{X}}^t$ by computing for each instance the weighted mean of its reconstructions:

$$\tilde{\mathbf{x}}_i^t = \beta_t \mathbf{x}_i^t + \sum_{s \neq d} \beta_s \tilde{\mathbf{x}}_i^{t\leftarrow s} \quad , \tag{6.10}$$

where $\beta_j$ can be seen as the prior of domain $\mathcal{X}^j$'s relevance, and $\sum_j \beta_j = 1$. When evaluating a new point $\mathbf{x}_{new}$, first the nearest neighbours from the training set are found, and then the reconstruction is given by $\tilde{\mathbf{N}}_i^{s\leftarrow t} \mathbf{w}_{new}$. The complexity of the algorithm is bounded by the matrix inversion of the coordinate mapping in equation 6.7, and therefore the algorithm complexity is $\mathcal{O}(d^3)$, where $d$ is the maximum number of features among all the domains.

## 6.3 Experimental Section

### 6.3.1 Methods

To test our method we used a recent clinical cohort Bom et al. (2019) containing data on patients with cardiovascular disease. There are 440 subjects in the dataset of which 56 suffered from an early cardiovascular event. For each patient, 359 protein levels and 9 clinical parameters are measured. We evaluate performance our method (stacked regularization with manifold mixing) for predicting a cardiovascular event. We compare proposed approach to that of using a standard stacked model with joint Bayesian optimization of the hyper-parameters, as well as with using random forest on the merged/ feature concatenated datasets (protein levels and clinical parameters). For both our method and the standard stacking, the architecture consisted of two larger random forest models in the first layer and a smaller one in the output.

### 6.3.2 Data selection and preprocessing

We perform random shuffles with 90% train size and even class distribution in the train/test set. We use remaining 10% to test the model. Since the dataset is imbalanced (much larger number of negative than positive subjects), we took a random sample from the negative class of size equal to the total number of positive class subjects, prior to the split at each shuffle. The protein were measured using a technology that uses standard panels for different proteins, meaning some of the proteins might have no relation to the outcome at all. For this reason, for each run we pre-selected 50 proteins using Univariate Feature Selection on the training set. Then, we normalized the train and test data independently, and measured the average ROC for each of the methods. We perform 5-fold cross validation for optimal hyper-parameter estimation on the train set for the random forests, and bayesian optimization for the stacked models. Once that is accomplished, we retrain the model with the optimal parameters on the complete training set and test on the remaining 10%. We repeat the this procedure multiple times and report the average ROC-AUC as well as the standard deviation. Described strategy is frequently referred to as stability selection procedure  Meinshausen and Bühlmann (2010) The proteins were measured using OLINK technology that records expression levels of proteins via targeted and customised analysis  Bom et al. (2019).

### 6.3.3 Results

The results are presented in figure 6.3. Proposed approach (MM stacked) outperformed both the regular stacked model and the random forests (RF) using the merged data. Both stacked regularized techniques outperformed standard RF.

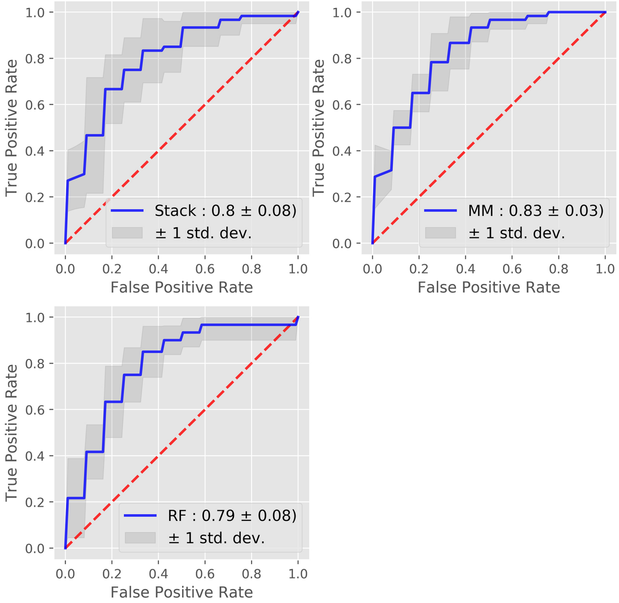


Figure 6.3: Average AUC for the three methods compared. The highest performance is that of the Manifold Mixing stacked model (MM), and both stacked models outperformed using Random Forests (RF) on the merged data

## 6.4 Conclusions and Future work

In this paper we propose the manifold mixing framework to improve the analysis of multi-modal data stemming from different sources. In our preliminary experiments, the obtained results support efficacy of our method. We outperform both standard stacked regularization and the model built on feature concatenated data. In the near future, we plan on performing further tests with larger number of shuffles, and testing on different datasets and heterogeneous domains. One pitfall of the current algorithm is the linearity of the map between manifolds which might fail in highly curved regions. A possible solution is to kernelize the method using graph kernels. Another interesting direction is to subdivide the manifold into multiple subregions based on the local curvature and create a mapping per subregion.

## Acknowledgments

Supported by a European Research Area Network on Cardiovascular Diseases (ERA-CVD) grant (ERA CVD JTC2017, OPERATION). Thanks to Cláudia Pinhão for supporting with the design of the figures.

# 7

# Interpretable Models via Pairwise permutations algorithm

Troy Maasland*, **João Pereira***, Diogo Bastos, Marcus de Goffau, Max Nieuwdorp, Aeilko H. Zwinderman, Evgeni Levin

*Equal contribution to this work

Submitted and accepted by AIMLAI, ECML PKDD 2021: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases

# Abstract

One of the most common pitfalls often found in high dimensional biological data sets are correlations between the features. This may lead to statistical and machine learning methodologies overvaluing or undervaluing these correlated predictors, while the truly relevant ones are ignored. In this paper, we will define a new method called *pairwise permutation algorithm* (PPA) with the aim of mitigating the correlation bias in feature importance values. Firstly, we provide a theoretical foundation, which builds upon previous work on permutation importance. PPA is then applied to a toy data set, where we demonstrate its ability to correct the correlation effect. We further test PPA on a microbiome shotgun dataset, to show that the PPA is already able to obtain biological relevant biomarkers.

## 7.1 Introduction

Measuring feature importance has often been plagued by high feature correlations. One important drawback is the lack of a theoretical definition for variable importance, in case variables are correlated Grömping (2009) Gregorutti, Michel, and Saint-Pierre (2017), even in linear models Grömping (2009). From a clinical perspective, correlated biomarkers are of high interest because they both may play a role in a shared biological pathway identified by the model and yet exhibit different behaviour in other circumstances. The method proposed in this paper, which will be referred to as *pairwise permutation algorithm* (PPA), allows us to calculate the importance of features without having to rely on the previously mentioned selection approaches. Highly correlated features, which have a similar relation with the output value, should have close importance ranks since they explain the same variability in the data. The *pairwise permutation algorithm* aims to provide feature importance values while avoiding the use of aggressive pre-selection techniques, since these techniques might remove relevant information from the data. It also manages to retain model interpretability by generating an importance value per feature, even when applied to black box models. Moreover, when working with highly dimensional biological data sets, it is simply not feasible to try and address each of the correlations in the data individually.

### Notation

We will refer to a single instance of the data-set as instance or point interchangeably throughout the paper. We denote matrices, 1-dimensional arrays and scalars with capital bold and regular text, respectively (e.g. $\mathbf{X}$, $\mathbf{x}$, $\alpha$). Matrices' columns and rows will be denoted by $\mathbf{X}[:, i]$ and $\mathbf{X}[j, :]$, respectively. The expected loss of a function given by: $\frac{1}{N}\sum_{i=1}^{N} l\left[y, f(\mathbf{x}_i)\right]$ will be denoted by $E_l\left[f(\mathbf{X})\right]$.

## 7.2 Related Work

In this work, we focus on model-agnostic procedures which can be divided into local and global methods. Local-based methods such as LIME (Local Interpretable Model agnostic Explanations) and its variants Ribeiro, Singh, and Guestrin (2016); Pereira et al. (2019) attempt to explain predictions on single data points by perturbing it and building a simple, yet interpretable model on the perturbed predictions. Similarly, SHAP (SHapley Additive exPlanations)Lipovetsky and Conklin (2001), offers a local explanation based on the additional prediction value each feature has when adding it to all possible feature subsets. Unlike local-based methods, global methods are concerned with determining the overall model behaviour and what features it values for its prediction. For example, in clinical research, the goal is to determine biomarkers that can identify a condition in the general population, or potential targets for novel drug development. Therefore, in this setting, we are mainly concerned with a more holistic view of feature importance i.e. global. A notable example is that of permutation importance which was first introduced by Breiman Breiman (2001) in random forests as a way to understand the interaction of variables that is providing the predictive accuracy. Suppose that for a certain feature $i$ in data-set $\mathbf{X}$, we randomly permute the instances' values, and denote the resultant data-set by $\mathbf{X}_i^{\pi}$. Permutation importance is defined as the difference in the expected model loss on the original dataset and the original one:

$$PI_{\{i\}}(f) := E_l\left[f(\mathbf{X}_i^{\pi})\right] - E_l\left[f(\mathbf{X})\right] \tag{7.1}$$

For random forests, there is already available work that analyzes the behaviour of this permutation importance, including the cases when high correlations are present. Gregorutti et al Gregorutti, Michel, and Saint-Pierre (2017) provided a theoretical description of the effect of correlations on the permutation importance, a phenomenon already observed by Toloși and Lengauer Gregorutti, Michel, and Saint-Pierre (2017) Strobl et al. (2007). Furthermore, a feature selection procedure was introduced, which was more efficient in selecting important, highly correlated variablesGregorutti, Michel, and Saint-Pierre (2017). Strobl et al showed that the larger feature importance values for correlated predictors in random forests were due to the preference for such predictors in the early splits of the trees. A new conditional permutation-based feature importance calculation was suggested, in order to circumvent this inflation, as well as the depreciation for its correlated predictor Strobl, Boulesteix, and et al (2008). Furthermore, Hooker and Mentch proposed the 'permute and relearn' approach Hooker and Mentch (2019). Based on this approach we define the relearned permutation importance as

$$PI_j^{\pi L} = E_l\left[f^{\pi j}(\mathbf{X_t})\right] - E_l\left[f(\mathbf{X_t})\right] \tag{7.2}$$

In which $f^{\pi j}$ is the model trained on the train dataset $\mathbf{X^{\pi j}}$, in which feature j is permuted, $f$ the model trained on the original train dataset $\mathbf{X}$ and $\mathbf{X_t}$ the test dataset. One drawback of this approach was also mentioned in the context of correlated features, as this resulted in the compensation effect, in which the importance of the correlated features was reduced Hooker and Mentch (2019). Local based methods, such as the ones introduced earlier, are focused on the contribution of each feature towards individual predictions, whereas permutation importance gives us a more broad estimation, since it is based on the overall accuracy of the model. While the former approach provides a higher degree of interpretability, the latter is usually more appropriate in a research environment, in which the aim would be to discover new leads which could help researchers to investigate the underlying biological mechanisms.

## 7.3 Pairwise Permutations Algorithm

### 7.3.1 Intuition

Features that are equally important for the output value should have similar feature importance ranks, and these should not be affected by feature correlation. In an attempt to prevent the compensation effect for correlated features mentioned by Hooker and Mentch, we have chosen to permute all the feature pairs and calculate the corresponding permutation importance of the pair. A key assumption in our method is that the higher the correlations, the larger should be the correction to that feature individual importance.

### 7.3.2 Definition

In this section, we define Pairwise Permutation Importance (PPI) as the weighted average of the permutation importance values, computed using the 'permute and relearn' approach defined in Equation 7.2. The correlations between the feature pairs will act as the weights. Let $\mathbf{R}$ be the correlations matrix between all the features and $\mathbf{R}_{i,j}$ the correlation value between features i and j. Let $PI_{i,j}$ define the relearn permutation importance (see Equation 7.2) when both the feature $i$ and $j$ have been permuted together, and $PI_{i,i}$ the relearn permutation importance, when only feature $i$ is permuted.

$$PPI_i = \underbrace{\frac{1}{\sum_{j=1}^{M} |\mathbf{R}_{i,j}|}}_{q} \underbrace{\left( PI_{i,i} + \sum_{\substack{j=1 \\ j \neq i}}^{M} |\mathbf{R}_{i,j}| \cdot PI_{i,j} \right)}_{p} \tag{7.3}$$

Note that when a feature has no correlations in the data, according to the previous equation, the PPI will actually follow the relearn permutation importance. Since for complex data sets with thousands of features the computational time can become infeasible ($\mathcal{O}(N^2)$), one possible simplification is to set a threshold and consider only the permutation pairs with a correlation above it. We define this procedure in algorithm 7.3.3.

### 7.3.3 Expected Difference

It might be tempting to compute the expected loss of the model, perform the permutation analysis and then compute the difference of the expected losses. This is actually how Fisher et al. Fisher, Rudin, and Dominici (2018) defined the permutation importance. However, we note that this procedure is sub-par as we illustrate in the following theorem:

**Theorem 7.3.1.** *For a given function $f : \mathbb{R}^M \to \mathbb{R}$, let $\mathbf{X}$ and $\mathbf{x}$ be a sample and an instance from the domain of $f$, respectively, $\mathbf{X}_i^\epsilon$ be $\mathbf{X}$ with permuted values for the r.v. $X_i$ and $\tilde{\mathbf{x}}$ an instance from $\mathbf{X}_i^\epsilon$. Then, for any loss function $l\left[y, f(\mathbf{x})\right]$ and norm function $||\cdot|| : \mathbb{R}^M \to \mathbb{R}$ it holds that:*
$\mathbf{E}\left[||l\left[y, f(\mathbf{x})\right] - l\left[y, f(\tilde{\mathbf{x}})\right]||\right] \geq ||\mathbf{E}\left[l\left[y, f(\mathbf{x})\right]\right] - \mathbf{E}\left[l\left[y, f(\tilde{\mathbf{x}})\right]\right]||$

*Proof.* Consider the following convex function $\varphi(x) = \|x\|$ for $x = l\left[y, f(\mathbf{x})\right] -$

$l\left[y, f(\tilde{\mathbf{x}})\right]$. Then, by Jensen's inequality:

$$\begin{aligned}\mathbf{E}\left[\varphi(x)\right] \geq \varphi\left(\mathbf{E}\left[x\right]\right) \Leftrightarrow & \mathbf{E}\left[\left\|l\left[y, f(\mathbf{x})\right] - l\left[y, f(\tilde{\mathbf{x}})\right]\right\|\right] \geq \left\|\mathbf{E}\left[l\left[y, f(\mathbf{x})\right] - l\left[y, f(\tilde{\mathbf{x}})\right]\right]\right\| \\ & = \left\|\mathbf{E}\left[l\left[y, f(\mathbf{x})\right]\right] - \mathbf{E}\left[l\left[y, f(\tilde{\mathbf{x}})\right]\right]\right\|\end{aligned}$$

□

This means that computing the expected value of the normed difference of individual loss values is more robust to non-linear relationships between the input variables then computing the difference of the normed expected loss values.

**Algorithm 3** Pairwise permutations algorithm

**Input:** $\mathbf{X}$, $\mathbf{X_t}$, $\mathbf{y_{test}}$, $E_l\left[f(\mathbf{X_t})\right]$, $\mathbf{R}$, $\alpha$

**Return:** $\mathbf{v}$

```
for feature $i$ in $\mathbf{X}$ do
    $p \leftarrow 0$, $q \leftarrow 0$ (equation 7.3)
    for feature $j$ in $\mathbf{X}$ do
        if $|\mathbf{R}_{i,j}| > \alpha$ then
            Permute the feature pair ($i$, $j$) together in $\mathbf{X}$
            Retrain the model with the permuted input data $\mathbf{X^{\pi}_{i,j}}$
            Calculate the model's error, $E_l\left[f^{\pi,i,j}(\mathbf{X_t})\right]$, on the test data
            Calculate $PI_{i,j}$ through the relearn formula $E_l\left[f^{\pi,i,j}(\mathbf{X_t})\right] - E_l\left[f(\mathbf{X_t})\right]$
            if i=j then
                $p \leftarrow p + PI_{i,i}$
            else
                $p \leftarrow p + |\mathbf{R}_{i,j}| \cdot PI_{i,j}$
            end if
            $q \leftarrow q + |\mathbf{R}_{i,j}|$
        end if
    end for
    $PPI_i \leftarrow p/q$
    $\mathbf{v}$.append($PPI_i$)
end for
```

## 7.4 Simulations with toy dataset

To see how our new PPA would behave for correlated features, we generated a toy dataset, based on the one used by Hooker and Mentch Hooker and Mentch (2019). The data was created by assuming a linear regression model:

$$y_i = x_{i1} + x_{i2} + x_{i3} + x_{i4} + x_{i5} + 0x_{i6} + 0.5x_{i7} + 0.8x_{i8} + 1.2x_{i9} + 1.5x_{i10}. \tag{7.4}$$

This was then turned into a classification model, by generating the binary outcome y with the classification rule :

$$y_i = \begin{cases} 1, & \text{for } y_i + \varepsilon_i \geq \overline{y} \\ 0, & \text{otherwise} \end{cases}, \tag{7.5}$$

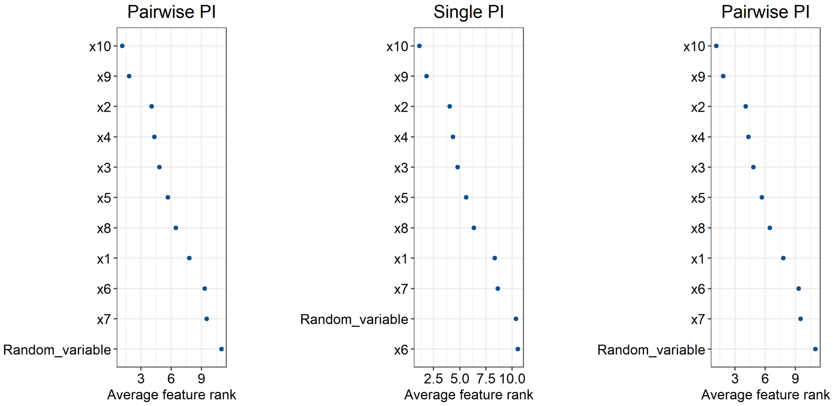
with $\varepsilon \sim N(0, 0.1)$. All features were generated from a multivariate normal distribution $N(0, \Sigma)$ with $\Sigma$ equal to the identity matrix, except that $\Sigma_{12} = \Sigma_{21} = \rho = 0.9$. All features were then transformed into a uniform distribution, mimicking the data generation procedure of Hooker and Mentch Hooker and Mentch (2019). In total, 1000 samples were generated.

In case the features are in the same scale, the coefficients in the linear model can be seen as the conditional importance of the feature on all other variables Strobl, Boulesteix, and et al (2008) Hooker and Mentch (2019). Therefore, based on the magnitude of the coefficients, we can rank the features on their importances, where features with the same coefficients should be equally important, while a feature with a higher coefficient should get higher importance than a feature with a lower coefficient. The order of the features should not be affected by any correlations between the features.

Using XGBoost with the logistic loss function as the classification algorithm Caruana et al. (2004); Chen and Guestrin (2016), we performed 50 stratified shuffle splits (70%train/30%test) and measured the ROC AUC after adding a noise feature to the dataset and standard scaling it. We found the XGBoost optimal hyperparameters using a 5-fold cross validation grid search. To compute the PPIs, a correlation threshold of 0.3 was used. Also, the single Permutation Importance (SPI) for each feature was obtained based on the 'permute and relearn' procedure, see Equation 7.2.

### 7.4.1 Results

The classification model obtained an average AUC of 0.97 ± 0.01. As shown by the average feature ranks in Figure 7.1(a), our new PPA is able to retrieve the



(a) $x_1$ and $x_2$ with $\rho = 0.9$ (b) $x_1$ and $x_2$ with $\rho = 0.9$ (c) $x_1$ and $x_6$ with $\rho = 0.9$

Figure 7.1: Average rank ± standard error for each feature based on the Pairwise Permutation Importance Algorithm for (a) and (c) and the Single Permutation Importance Algorithm for (b).

right order of feature importances, in which $x_{10}$ is clearly the most important one, followed by $x_9$. As expected, $x_6$ and the random variable are identified as the least important features. The results for the SPI are shown in Figure 7.1(b). It is clearly shown that the PPA outperforms this approach, as the SPI decreased the importance of the correlated features $x_1$ and $x_2$ and was not able to retrieve the right order of feature importances. This was also observed for the random forest algorithm by Hooker and Mentch Hooker and Mentch (2019). The toy dataset showed that in case two features have the same coefficient in the linear model and are correlated, the PPA is able to retrieve the right order for the feature importances. We also analysed the effect of a correlation of $\rho = 0.9$ between $x_1$ and $x_6$, by changing the covariance matrix $\Sigma$ to $\Sigma_{16} = \Sigma_{61} = 0.9$ and setting the correlation between $x_1$ and $x_2$ to 0. This represents a case in which an important feature is correlated to an irrelevant feature. However, we saw in this case that the importance of $x_1$ was decreased by $x_6$, while the importance of $x_6$ was increased by $x_1$, as shown in Figure 7.1(c). This could be expected as the grouped importance is shared equally between both features, while in the case of features with different importances, this might not be the right assumption. In this case, the PPA may not be the appropriate choice

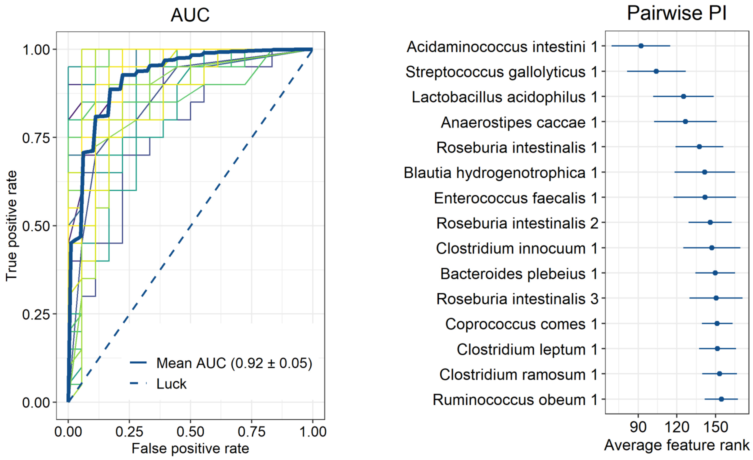
## 7.5 Microbial biomarkers for Type 2 Diabetes Mellitus

In this section we test the PPA on a real-world dataset, specifically microbiome data. The goal is to obtain biologically relevant markers. Therefore, we downloaded the Qin 2012 microbiome dataset from MLRepo Vangay, Hillmann, and Knights (2019), Qin et al. (2012). This curated classification dataset contained shotgun data for 124 samples, representing Chinese healthy controls (n = 59) and Type 2 Diabetes Mellitus (T2D) patients (n = 65). For full details of the preprocessing of the raw sequence reads for datasets in MLRepo, see Vangay, Hillmann, and Knights (2019). We used the same procedure as in the previous section with some additional preprocessing. First, the read counts were rarefied to 28 358 reads per sample, which was the lowest observed number of reads in a sample. After that, features with less than 6 reads per sample on average, representing a relative abundance of 0.02%, were removed. The final dataset consisted then of 124 samples with 377 OTUs.

### 7.5.1 Results

The classification model was able to achieve an average roc-auc score of 0.92 ±0.05, as depicted in figure 7.2(a).

Figure 7.2(b) represents the top 15 most predictive microbial OTUs in the classification model. Analyzing these OTUs (and several more beyond the top 15) primarily highlights 2 main patterns. The strongest pattern observed in the data, most likely represents an effect that T2D has on the dietary behavior of these Chinese T2D patients. *Lactobacillus acidophilus, Acidaminococcus intestini* and *Anaerostipes caccae* are strongly associated with T2D and with each other in this dataset. A regular dose of *L. acidophilus* is commonly recommended in Chinese Medicine Cohen (2015). Fermented soybean products are popular in China (i.a.) and various of these products commonly contain *L. acidophilus* Chang, Kim, and Han (2010), Bedani, Rossi, and Saad (2013), Kanda et al. (1976). Indeed, there is evidence that supports the beneficial claims regarding these fermented products and T2D Kwon et al. (2010), Mueller et al. (2012). Trans-aconitic acid in the urine is a biomarker for the consumption of soy products Münger et al. (2017) and *Acidaminococcus* is known to be able to oxidise trans-aconitate Cook,



(a) The roc-auc plot.

(b) Top 15 ranked feature importances.

Figure 7.2: (a) Individual ROC AUC curves for each shuffle and average ROC AUC plot for all shuffles. (b) Average rank ± standard error for the top 15 ranked features based on the Pairwise Permutation Importance Algorithm.

Wells, and Russell (1994) converting it to acetate. *A. caccae*, is an acetate and lactate consuming butyrate producer. Cross-feeding interactions between *L. acidophilus* and *A. caccae* have been analyzed in detail in vitro Moens, Verce, and Vuyst (2017). Other butyrate producing species, like *Roseburia intestinalis*, can have similar cross-feeding interactions Saulnier et al. (2009) but were not part of this specific pattern, but with the 2nd main pattern (see below), suggesting that *A. caccae* was part of same fermented soybean product popular with, or given to, these Chinese T2D patients that likely also contained *L. acidophilus* and *A. intestini.*

The second pattern involves several butyrate producers (the *Roseburia, Faecalibacterium, Coprococcus* genera, several *Eubacterium* species and *Anaerostipes hadrus*) in a cross-feeding relationship with various acetate producing dietary fibre degrading species (*Blautia* and *Ruminococcus* representatives). This cluster of species is generally found to be negatively associated with T2D, not just in this study throughout the diabetes microbiome field Qin et al. (2012); de Goffau et al. (2013); Hur and Lee

(2015); Hartstra et al. (2015); Murri et al. (2013). Insufficient butyrate production has been associated with both T1D and T2D development both in rats, mice and in humans Noureldein et al. (2020), Endesfelder et al. (2016), Jia et al. (2017), Khan and Jena (2016). Besides being used by colonocytes as a primary energy source Donohoe et al. (2012) butyrate is a powerful inhibitor of histone deacetylase, which has emerged as a target in the control of insulin resistance Sharma and Taliyan (2016), Dirice et al. (2017), Khan and Jena (2015). Animal and in vitro studies have generally found a beneficial effect of butyrate and acetate on glucose homeostasis and insulin sensitivity Canfora, Jocken, and Blaak (2015).

## 7.6 Conclusions

In this paper, we have set a first step in correcting the compensation effect, observed for 'permute and relearn' permutation importances in case correlated features are present. Our new PPA is able to obtain the right ranking of features, when two features are highly correlated and have the same importance, stated by the magnitude of their coefficient, in linear models. Furthermore, while not yet optimal for correlations between more than 2 features or correlated features with unequal importance related to the output variable, our PPA is already able to obtain relevant biological insights in a Chinese Type 2 Diabetes microbiome dataset.

## 7.7 Acknowledgments

We would like to thank Manon Balvers for helping with computational experiments.

# 8

# Covered Information Disentanglement: Model Transparency via Unbiased Permutation Importance

**João P. B. Pereira**, Erik S.G. Stroes, Aeilko H. Zwinderman, Evgeni Levin

Proceedings of the AAAI Conference on Artificial Intelligence, 36(7), 7984-7992. https://doi.org/10.1609/aaai.v36i7.20769

# Abstract

Model transparency is a prerequisite in many domains and an increasingly popular area in machine learning research. In the medical domain, for instance, unveiling the mechanisms behind a disease often has higher priority than the diagnostic itself since it might dictate or guide potential treatments and research directions. One of the most popular approaches to explain model global predictions is the *permutation importance* where the performance on permuted data is benchmarked against the baseline. However, this method and other related approaches will undervalue the importance of a feature in the presence of covariates since these cover part of its provided information. To address this issue, we propose Covered Information Disentanglement (*CID*), a framework that considers all feature information overlap to correct the values provided by *permutation importance*. We further show how to compute *CID* efficiently when coupled with *Markov random fields*. We demonstrate its efficacy in adjusting *permutation importance* first on a controlled toy dataset and discuss its effect on real-world medical data.

## 8.1 Introduction

Understanding the biological underpinnings of disease is at the core of medical research. Model transparency and feature relevance are thus a top priority to discover new potential treatments or research directions. One of the current most popular methods to explain local model predictions is *SHAP* Lipovetsky and Conklin (2001); Štrumbelj and Kononenko (2014); Lundberg and Lee (2017), a game-theoretic approach that considers the features as “players” and measures their marginal contributions to all possible feature subset combinations. *SHAP* has also been generalized in *SAGE* Covert, Lundberg, and Lee (2020) to compute global feature importance. However, recent work by Kumar et al. Kumar et al. (2020) exposes some mathematical issues with *SHAP* and concludes that this framework is ill-suited as a general solution to quantifying feature importance. Other local-based methods such as *LIME* Ribeiro, Singh, and Guestrin (2016) and its variants (see e.g. Singh, Ribeiro, and Guestrin (2016); Ribeiro, Singh, and Guestrin (2018.); Guidotti et al. (2018); Pereira et al. (2019)) build weak yet explainable models on the neighborhood of each instance. While this achieves higher prediction transparency for each data point, in this work, we are mainly concerned with a more holistic view of importance, which may be more appropriate to guide new research directions and unravel disease mechanisms. Tree-based methods are very commonly selected for this purpose because they compute the impurity or *Gini importance* Breiman (2001). The impurity importance is biased in favor of variables with many possible split points; i.e. categorical variables with many categories or continuous variables Strobl et al. (2007). A generally accepted alternative to computing the *Gini importance* is the *permutation importance* Breiman (2001), which benchmarks the baseline performance against permuted data. There is, however, the issue of multicollinearity. When features are highly correlated, feature permutation will underestimate the individual importance of at least one of the features, since a great deal of the information provided by this feature is "covered" by its covariates. One option is to permute correlated features together Toloşi and Lengauer (2011). However, this implies choosing an arbitrary correlation grouping threshold. Most importantly, it misses the differentiation between each feature's contribution to the final prediction. Motivated by the idea that there is an information overlap between different features, we develop Covered Information Disentanglement (*CID*),[1] an information-theoretic approach to disentangle the shared information and scale the *permutation importance* values accordingly. We demonstrate how *CID* can recover the right importance ranking on artificial data and discuss its efficacy on the Cardiovascular Risk Prediction dataset Hoogeveen et al. (2020).

## 8.2 Methodology

**Notation**

We denote matrices, 1-dimensional arrays, and scalars/functions with capital bold, bold, and regular text, respectively (e.g. $\mathbf{X}$, $\mathbf{x}$, $\alpha/f$). Given a dataset $\mathbf{X}_{M\times N}$, we will denote its random variables by capital regular text with a subscript and the values using lowercase (e.g. $X_i$ and $x_i$), while the joint density/mass will be represented as $p(x)$. The expected loss of a function given by: $\frac{1}{M}\sum_{i=1}^{M} l\left[y,\, f(\mathbf{x}_i)\right]$ will be denoted by $\mathcal{L}\left[f\left(\mathbf{X}\right)\right]$.

### 8.2.1 Information Theory background

*Information theory* (IT) is a useful tool used in quantifying relations between random variables. The basic building block in IT is the *entropy* of an r.v. $X_i$, which is defined as: $H(X_i) \equiv -\sum_{x_i} p(x_i)\log p(x_i)$. The *joint entropy* between

[1] We make an implementation of *CID* publicly available at: https://github.com/JBPereira/CID.

r.v.s $X_i$ and $X_j$ is defined as: $H(X_i, X_j) \equiv -\sum_{x_i}\sum_{x_j} p(x_i, x_j)\log p(x_i, x_j)$. The *mutual information* between r.v.s $X_i$ and $X_j$ is the relative entropy between the joint entropy and the product distribution $p(x_i)p(x_j)$: $I(X_i, X_j) \equiv \sum_{x_i}\sum_{x_j} p(x_i, x_j)\log\frac{p(x_i,x_j)}{p(x_i)p(x_j)}$. For a more thorough exposition to IT, the reader can refer to Cover and Thomas (2012).
Using the definitions above, one can derive properties that resemble those of set theory, where joint entropy and mutual information are the information-theoretic counterparts to union and intersection, respectively Ting (2008). In order to keep this intuition when generalizing to higher dimensions, one can define the entropy of the union of $N$ features as:

**Definition 8.2.1.** ***Multivariate Union Entropy***

$$H\left(\cup_{i=1}^{N} X_i\right) \equiv -\sum_{x_i} p(x_1,\, ...,\, x_N)\,log\, p(x_1,\, ...,\, x_N)$$

and using the Inclusion-Exclusion principle, we can define the intersection as:

**Definition 8.2.2.** ***Multivariate Intersection Entropy***

$$H\left(\cap_{i=1}^{N} X_i\right) \equiv \sum_{x_1,\ldots,x_N} p(x_1,\, ...,x_N)h_{ci}(x_1,\, ...,x_N),$$

$$h_{ci}(x_1,\, ...,x_N) = \sum_{k=1}^{N}(-1)^{k-1} \sum_{\substack{I\subseteq\{1,\,...,\,N\};\\ |I|=k}} h(x_{I_1},\, ...\, ,\, x_{I_k}),$$

*$h(\mathbf{x}) = -log\, p(\mathbf{x})$ and $h_{ci}$ is the local co-information.*

This definition of multivariate intersection is also called co-information and it may yield negative values. This can happen for instance if $X_i$ has no correlation with $X_I$ but knowing $X_I$ introduces a correlation between the two (what is commonly known as ‘explaining away’). This motivated Williams and Beer to draw the distinction between redundant and synergistic information and propose *partial information decomposition* (PID) Williams and Beer (2010). Ince (Ince, 2017) thoroughly analyzed the multivariate properties of PID directly applied to multivariate entropy and suggested to divide the individual terms in definition 8.2.2, so that positive local entropy terms correspond to redundant entropy, while the negative ones correspond to synergistic entropy.

### 8.2.2 Permutation Feature Importance

Feature importance is a subjective notion that may vary with application. Consider a supervised learning task where a model $f$ is trained/tested on dataset $\mathbf{X}$, $\mathbf{y}$ and its performance is measured by a function $\mathcal{L}$. In this work, we will refer to feature importance as the extent to which a feature $X_i$ affects $\mathcal{L}[f(\mathbf{X})]$, on its own and through its interactions with $X_{\setminus\{i\}}$. Permutation importance was first introduced by Breiman Breiman (2001) in random forests as a way to understand the interaction of variables that is providing the predictive accuracy.
Consider a dataset $\mathbf{X}_{M\times N}$ and denote the $j$th instance of the $i$th feature by $\mathbf{X}_i^j$. Suppose the set $\{1, ..., M\}$ is sampled and denote the subsample by $\mathbf{s}$, $\mathbf{s} \subseteq \{1, ..., M\}$. Consider further a random permutation of this subset which we denote by $\boldsymbol{\pi}\left(\mathbf{s}\right)$ and its $j$th element by $\boldsymbol{\pi}_j\left(\mathbf{s}\right)$. The *permutation importance*, is given by:

$$e_i(f,\mathbf{s}) = \sum_{j\in\mathbf{s}}^{|\mathbf{s}|}\left(\mathbf{E}_{\sim p(\boldsymbol{\pi})}\left[\mathcal{L}\left(f\left(\mathbf{X}_1^j, ..., \mathbf{X}_i^{\boldsymbol{\pi}_j(\mathbf{s})}, ..., \mathbf{X}_N^j\right)\right)\right]\right.$$
$$\left. - \mathcal{L}\left(f\left(\mathbf{X}_1^j, ..., \mathbf{X}_N^j\right)\right)\right) \tag{8.1}$$

$$e_i(f) = \mathbf{E}_{\sim p(\mathbf{s})}\ \left[e_i(f,\mathbf{s})\right] \tag{8.2}$$

### 8.2.3 Covered Information Disentanglement

In the presence of covariates, the *permutation importance* measures the performance dip caused by removing the non-mutual information between the feature and the remaining data. That is:

$$e_i\left(f\right) = \mathcal{I}_i\left(f\right) - e_i^{\cup}(f), \tag{8.3}$$

where $\mathcal{I}_i\left(f\right) = \mathbf{E}_{\sim p(\mathbf{s})}\left[\mathcal{I}_i(f,\mathbf{s})\right]$ is the expected total importance of feature $i$ under model $f$ (the quantity we are interested in) and $e_i^{\cup}\left(f\right) = \mathbf{E}_{\sim p(\mathbf{s})}\left[e_i^{\cup}(f,\mathbf{s})\right]$ is the expected performance dip covered by all other variables. To compute $e_i^{\cup}\left(f\right)$ would require applying the Inclusion-Exclusion principle and measuring the performance dip for all possible feature combinations of size 1 to the number of features. Instead, we note that $e_i^{\cup}\left(f\right)$ intuitively measures the model performance dip when the model is deprived of the information covered by the r.v.s that are correlated with $X_i$. For an intuitive depiction of the problem, see figure 8.1.

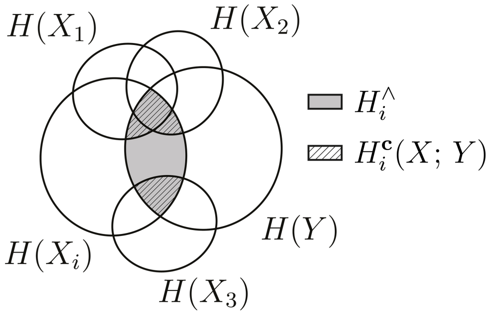


Figure 8.1: An illustration of the *permutation importance* bias in the presence of covariates and the measures needed to correct it. The mutual information between random variable $X_i$ and $Y$ (represented in gray) is covered by the information provided by r.v.s $X_1$, $X_2$ and $X_3$. Permutation importance only measures the non-covered part (non-shaded gray), and to correct its value, we suggest computing $H_i^{\mathbf{c}}(X;Y)$.

Motivated by the analogy between set-theory and information measures, we define the joint information between an r.v. and the target variable that is "covered" by the other r.v.s as:

**Definition 8.2.3.** ***Covered information (CI)*** *Given an r.v. $X_i$ and a set of distinct r.v.s $X_{\mathbf{i}^-}$, $\mathbf{i}^- = \{1, \dots N\}\backslash\{i\}$, the information of $X_i$ w.r.t. $Y$ covered by $X_{\mathbf{i}^-}$ is defined as:*

$$H_i^{\mathbf{c}}(X;Y) = H\left(X_i \cap Y \cap \left\{\cup_{j\in\mathbf{i}^-} X_j\right\}\right).$$

When it is clear from the context what $Y$ and $X_{\mathbf{i}^-}$ are, we will abbreviate $H_i^{\mathbf{c}}(X;Y)$ into $H_i^{\mathbf{c}}$, denote the mutual information with $Y$ by $H_i^\wedge$, and the respective local co-information terms for the $k$th row in the dataset with $h_{ik}^{\mathbf{c}} \equiv h_i^{\mathbf{c}}(\mathbf{X}_i^k, Y^k)$ and $h_{ik}^\wedge \equiv h_i^\wedge(\mathbf{X}_i^k, \mathbf{y}^k)$. We further divide $H_i^{\mathbf{c}}$ and $H_i^\wedge$ into its redundant and synergistic counterparts, which for a specific sample $\mathbf{s}$ are given by:

$$\textbf{Redundant MI}: H_i^{\wedge^+}(\mathbf{s}) = \frac{1}{|\mathbf{s}|}\sum_{k\in\mathbf{s}} \max\left(0, h_{ik}^{\wedge}\right)$$
$$\textbf{Synergistic MI}: H_i^{\wedge^-}(\mathbf{s}) = \frac{1}{|\mathbf{s}|}\sum_{k\in\mathbf{s}} \left|\min\left(0, h_{ik}^{\wedge}\right)\right|$$
$$\textbf{Redundant CI}: H_i^{\mathbf{c}^+}(\mathbf{s}) = \frac{1}{|\mathbf{s}|}\sum_{k\in\mathbf{s}} \max\left(0, h_{ik}^{\mathbf{c}}\right)$$
$$\textbf{Synergistic CI}: H_i^{\mathbf{c}^-}(\mathbf{s}) = \frac{1}{|\mathbf{s}|}\sum_{k\in\mathbf{s}} \left|\min\left(0, h_{ik}^{\mathbf{c}}\right)\right|$$

**Assumption 8.2.1.** *Permutation importance and entropy terms are related through a map $\phi_f : \mathbb{R}^4 \to \mathbb{R}$, such that $e_i(f, \mathbf{s}) = \phi_f\left(H_i^{\mathbf{c}^+}(\mathbf{s}), H_i^{\mathbf{c}^-}(\mathbf{s}), H_i^{\wedge^+}(\mathbf{s}), H_i^{\wedge^-}(\mathbf{s})\right) + \epsilon$, where $\epsilon$ is an error term.*

Thus, if assumption 8.2.1 holds, we can use the information of $X_i$ w.r.t. $Y$ by $X_{\mathbf{i}^-}$ and approximate equation 8.3 with:

$$\begin{aligned} e_i^{\cup}(f, \mathbf{s}) \approx &\phi_f\left(0, H_i^{\mathbf{c}^-}(\mathbf{s}), H_i^{\wedge^+}(\mathbf{s}), H_i^{\wedge^-}(\mathbf{s})\right) - \\ &\phi_f\left(H_i^{\mathbf{c}^+}(\mathbf{s}), H_i^{\mathbf{c}^-}(\mathbf{s}), H_i^{\wedge^+}(\mathbf{s}), H_i^{\wedge^-}(\mathbf{s})\right). \end{aligned} \tag{8.4}$$

This means we can approximate the result of permuting all possible combinations of features by computing only the single-feature permutation loss and the covered information of r.v. $X_i$ by all the others. Here, we are implicitly defining: $\mathcal{I}_i(f, \mathbf{s}) \equiv \phi_f\left(0, H_i^{\mathbf{c}^-}(\mathbf{s}), H_i^{\wedge^+}(\mathbf{s}), H_i^{\wedge^-}(\mathbf{s})\right)$, and thus the true importance in the performance difference scale is given by mapping the entropy values when there is no redundant entropy to the space of performance differences.

Since we are predicting the feature importance using a map between entropy terms (which measure model-agnostic importance) and *permutation importance* values, the end result depends only on how learnable is the model behavior w.r.t to entropy. Moreover, since the entropy values are computed for the different subsample sets $\mathbf{s}$, the overall importance variability is also estimated.

For two datasets where $I(X_i, Y) > I(X_i, Y')$ but the covered info of $(X_i, Y) > (X_i, Y')$, *CID* would correctly value $\mathcal{I}_i(f) > \mathcal{I}_i(f')$ which is not guaranteed using Shapley based methods since the contributions to subsets

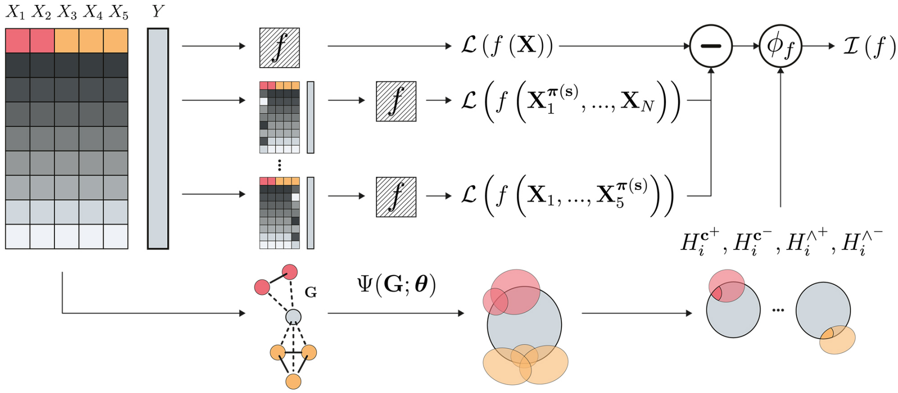

$X_1$ $X_2$ $X_3$ $X_4$ $X_5$
$Y$
$f$
$\mathcal{L}(f(\mathbf{X}))$
$\mathcal{L}\left(f\left(\mathbf{X}_1^{\pi(\mathbf{s})}, ..., \mathbf{X}_N\right)\right)$
$\mathcal{L}\left(f\left(\mathbf{X}_1, ..., \mathbf{X}_5^{\pi(\mathbf{s})}\right)\right)$
$\phi_f$
$\mathcal{I}(f)$
$H_i^{\mathbf{c}^+}, H_i^{\mathbf{c}^-}, H_i^{\wedge^+}, H_i^{\wedge^-}$
G
$\Psi(\mathbf{G}; \boldsymbol{\theta})$

Figure 8.2: *CID* importance diagram. The permutation feature importance is computed by first calculating the expected loss of the model $f$ $(\mathcal{L}(f(\mathbf{X})))$. Then, each feature's values are permuted and the expected loss of $f$ computed. Subtracting each permuted dataset loss to the original one yields the *permutation importance*. *CID* starts by inferring the network $\mathbf{G}$ for the *Markov random field* $\Psi$ (alternatively, a prior network is given), then the *MRF* parameters $\boldsymbol{\theta}$ are inferred, and finally, $H_i^{\mathbf{c}}/H_i^{\wedge}$ are computed for each feature, which are then used to train the entropy/PI model $\phi_f$ and predict the true importance $\mathcal{I}(f)$.

of features correlated with $X_i$ are biased. The Shapley efficiency+symmetry properties also imply that correlated features' scores are scaled down. To see this, consider $X_i = X_j$, then symmetry→ $\phi_i(v_f) = \phi_j(v_f)$ and efficiency→ $\phi_i(v_f) = \phi_j(v_f) = (v_f(D) - \sum_{k \neq i,j} \phi_k(v_f))/2$. In contrast, *CID* values do not sum to the complete data performance, but rather are meaningful individually.

There is still the issue of computing $H_i^{\mathbf{c}}$, since it involves computing $p(X)$. Since directionality is irrelevant for the purpose of computing overlapping information, we suggest to model $p(X)$ using an undirected graphical model (UGM). Let $G = (V, E)$ denote a graph with $N$ nodes, corresponding to the $\{X_1, ..., X_N\}$ features, and let $\mathcal{C}$ be a set of cliques (fully-connected subgraphs) of the graph $G$. Denoting a set of clique-potential functions by $\{\psi_{\mathcal{C}} : \mathcal{X}^{|\mathcal{C}|} \to \mathbb{R}\}$, the distribution of a *Markov random field* (*MRF*) Koller and Friedman (2009) is given by: $p(x) = \prod_{c \in \mathcal{C}} \psi_c(x_c)/\mathbf{Z}$, where $\mathbf{Z} = \int \prod_{c \in \mathcal{C}} \psi_c(x_c) dx$ is the partition function. By the Hammersley-Clifford theorem, any distribution that can be represented in this way satisfies: $X_i \perp X_j | X_{\mathcal{N}(X_i)}$ for any $X_j \notin \mathcal{N}(X_i)$, where $\mathcal{N}(X_i)$ is the set $\{X_k : (i,k) \in E\}$. This allows to significantly simplify the expression of covered information yielding the main result of this paper:

**Theorem 8.2.1.** *Consider an r.v. $X_i$ and set of r.v.s $X_{\mathbf{i}^-}$, $\mathbf{i}^- = \{1, ..., N\}\backslash\{i\}$, a response r.v. $Y$, as well as the set of r.v.s that are neighbors to both $X_i$ and $Y$: $X_{\mathcal{N}(i,y)}, \mathcal{N}(i,y) \in \cup\{\mathcal{N}(X_i), \mathcal{N}(Y)\}$. For a Markov random field, the covered information of $X_i$ by $X_{\mathbf{i}^-}$ w.r.t. $Y$ is given by:*

$$H_i^{\mathbf{c}} = H_i^{\wedge} - \mathbf{E}_{\sim p(x_{\mathcal{N}(i,y)})}\left[log\left(f\frac{\mathbf{d}^T\mathbf{F}\,\mathbf{e}}{\mathbf{d}^T\mathbf{F}_y\mathbf{F}_{x_i}^T\mathbf{e}}\right)\right],$$

*where $\mathbf{F}$ is a matrix with the product of joint potential values $\psi_{\mathcal{C}_F}$ for set of cliques $F : \{c\,|\,X_i, Y \in c\}$; $f$, $\mathbf{F}_y$ and $\mathbf{F}_{x_i}$ are an entry, column, and row of $\mathbf{F}$, respectively, while $\mathbf{d}$ and $\mathbf{e}$ are arrays with the product of potential values $\psi_{\mathcal{C}_D}$, $\psi_{\mathcal{C}_E}$ for set of cliques $D : \{c\,|\,X_i \in c, Y \notin c\}$ and $E : \{c\,|\,X_i \notin c, Y \in c\}$ with fixed $X_{\mathbf{i}^-}$.*

*Proof.* Using definition 8.2.1, 8.2.2 and 8.2.3:

$$H_i^{\mathbf{c}} = H_i^{\wedge} + \overbrace{H(X_i \cup Y \cup X_{\mathbf{i}^-})}^{\textcircled{1}} - \overbrace{H(X_{\mathbf{i}^-} \cup Y)}^{\textcircled{2}} + \overbrace{H(X_{\mathbf{i}^-})}^{\textcircled{3}} - \overbrace{H(X_i \cup X_{\mathbf{i}^-})}^{\textcircled{4}}.$$

The probability density for Markov Random fields is equal to $p(x) = \prod_{c\in\mathcal{C}} \psi_c(x_c)/\mathbf{Z}$, where $\mathbf{Z}$ is the partition function and $\mathcal{C}$ is the set of cliques in the Markov network. Define two sets of cliques: $A : \{c\,|\,X_i \in c\}$ and $B : \{c\,|\,X_i \notin c\}$. In that case (ignoring the partition function term because it cancels out):

$$\textcircled{1} = -\sum_x p(x)\left[log\prod_{b\in B}\psi_b(x_b) + log\prod_{a\in A}\psi_a(x_a)\right],$$

$$\textcircled{2} = -\sum_x p(x)\left[log\prod_{b\in B}\psi_b(x_b) + log\sum_{x_i}\prod_{a\in A}\psi_a(x_a)\right],$$

$$\textcircled{1} - \textcircled{2} = -\sum_x p(x)log\left(\frac{\prod_{a\in A}\psi_a(x_a)}{\sum_{x_i}\prod_{a\in A}\psi_a(x_a)}\right).$$

To compute $\textcircled{3} - \textcircled{4}$, define four sets of cliques: $C : \{c\,|\,X_i \notin c, Y \notin c\}$, $D : \{c\,|\,X_i \in c, Y \notin c\}$, $E : \{c\,|\,X_i \notin c, Y \in c\}$, and $F : \{c\,|\,X_i \in c, Y \in c\}$. In order to reduce the clutter, we will introduce the following functions:

$d(x_i, x_{\mathbf{i}^-}) = \prod_{j \in \mathbf{i}^-, j \sim i} \psi(x_i, x_j)$, $e(y, x_{\mathbf{i}^-}) = \prod_{j \in \mathbf{i}^-, j \sim y} \psi(y, x_j)$, $f(x_i, y) = \psi(x_i, y)$, where we will abbreviate $d(x_i, x_{\mathbf{i}^-})$ into $d(x_i)$ and $e(y, x_{\mathbf{i}^-})$ into $e(y)$ when the value for random variable $X_{\mathbf{i}^-}$ is fixed. Then (again, ignoring the partition function):

$$
\begin{aligned}
③ &= -\sum_x p(x) \left[ log \prod_{c \in C} \psi_c(x_c) + log \sum_{x_i} \sum_y d(x_i) e(y) f(x_i, y) \right], \\
④ &= -\sum_x p(x) \left[ log \prod_{c \in C} \psi_c(x_c) + log \sum_y d(x_i) e(y) f(x_i, y) \right], \\
③ - ④ &= -\sum_x p(x) log \left( \frac{\sum_{x_i} \sum_y d(x_i) e(y) f(x_i, y)}{\sum_y d(x_i) e(y) f(x_i = X_i, y)} \right),
\end{aligned}
$$

where $f(x_i = X_i, y)$ is the function $f$ for a fixed value of the r.v. $X_i$. Since the set of cliques $A = \{D \cup F\}$, and denoting by $d(X_i)$, $f(X_i, Y)$ the functions $d$ and $f$ for fixed values of $X_i$ and $Y$, then:

$$
\begin{aligned}
&(① - ②) + (③ - ④) = \\
&-\sum_x p(x) log \left( \frac{\sum_{x_i} \sum_y d(X_i) d(x_i) f(X_i, Y) e(y) f(x_i, y)}{\sum_{x_i} \sum_y d(X_i) d(x_i) f(x_i, Y) e(y) f(X_i, y)} \right) \\
&= -\mathbf{E}_{\sim p(x_{\mathcal{N}(i,y)})} \left[ log\, f(X_i, Y) + log \left( \frac{\mathbf{d}^T \mathbf{F}\, \mathbf{e}}{\mathbf{d}^T \mathbf{F}_y \mathbf{F}_{x_i} \mathbf{e}} \right) \right],
\end{aligned}
$$

where $x_{\mathcal{N}(i,y)}$ is an instance of the set of r.v.s that are neighbors to either $X_i$ or $Y$, $\mathbf{d}$ and $\mathbf{e}$ are column arrays with the different values of $d(x_i)$ and $e(y)$ for fixed $X_{\mathbf{i}^-}$, $\mathbf{F}$ is a matrix with all the values $f(x_i, y)$ with varying values of $X_i$ in the rows and $Y$ in the columns, and $\mathbf{F}_y$ and $\mathbf{F}_{x_i}$ are row and column vectors of $\mathbf{F}$ corresponding to fixed $Y$ and fixed $X_i$, respectively. This yields the result of the theorem. □

### Considerations and simplifications

If a 2-clique *MRF* is chosen, then $\mathbf{F}$ depends only on $X_i$ and $Y$, and can be computed before the expectation.
***Gaussian MRF*:** Learning an *MRF*'s network structure is expensive. One popular approach is to use *graphical lasso* Friedman, Hastie, and Tibshirani (2008) which learns the entries of a Gaussian precision matrix by

finding: $\min_{\mathbf{\Lambda} \in \mathbb{S}^n_+} -log\ det(\mathbf{\Lambda}) + tr(\mathbf{S\Lambda}) + \rho ||\mathbf{\Lambda}||_1$, where $\mathbf{\Lambda}$ is the precision matrix (constrained to belong to $\mathbb{S}^n_+$, the set of positive semi-definite $n \times n$ matrices), $\mathbf{S}$ is the empirical covariance matrix and $\rho$ acts in analogy to Lasso regularization by penalizing a large number of non-zero precision entries. We can model the potentials using *Gaussian Markov random fields* whose potentials are $\psi_{s,t}(x_s, x_t) = \exp\left[-\frac{1}{2} x_s \Lambda_{st} x_t\right]$, $\psi_s(x_s) = \exp\left[-\frac{1}{2}\left(x_s^2 \Lambda_{ss} - 2\eta_s x_s\right)\right]$, where $\boldsymbol{\eta} = \mathbf{\Lambda}\boldsymbol{\mu}$ ($\boldsymbol{\mu}$ is the mean vector).

**Discrete Approximation:** Continuous *MRF* such as Gaussian Markov Random fields depend on a continuous multivariate distribution and thus the entropy must be replaced by differential entropy, which violates many of the desired properties of discrete entropy. Therefore, we will approximate a continuous distribution with a discrete one $p(x_i) \approx \delta_i p(\overline{x_i})$, where $\delta_i$ is the $i$th feature bin size and $\overline{x_i}$ is the mean value of the bin, and then carry on with our computations as specified in theorem 8.2.1. For the case where all bins have the same size per feature, all the $\delta$s cancel out.

**Complexity:** If we approximate the expectation in theorem 8.2.1 with the empirical expectation, then the asymptotic complexity becomes $\mathcal{O}(SB^2)$, where $S$ is the number of samples and $B$ is the maximum between the number of bins used to discretise continuous values and the maximum number of values the discrete features take (typically, $B \ll S$). This can be computed in parallel for each feature.

**Baseline and maximum importance** The *permutation importance* of the whole feature set: $e_X(f, \mathbf{s}) = \mathcal{I}_X(f, \mathbf{s}) = \phi_f \left(0, 0, H_X^{\wedge^+}(\mathbf{s}), H_X^{\wedge^-}(\mathbf{s})\right)$ and/or the empty set: $e_\varnothing(f, \mathbf{s}) = \mathcal{I}_\varnothing(f, \mathbf{s}) = \phi_f\,(0, 0, 0, 0)$ can be added to the info-PI set to improve the model map $\phi_f$.

**Out-of-distribution problem** In PI, models are evaluated in regions outside the training distribution domain. For CID, substituting PI for permute and retrain or feature ablation solves this issue.

## 8.3 Experimental Section

To test the *CID* ranking adjustment, we first tested it on a toy dataset where the real importances are known, and a real-world medical dataset. We implemented *CID* in Python using scikit-learn's *graphical lasso* Pedregosa and et al. (2011). For the toy dataset, we used scikit-learn's *Extremely Randomized Trees* and *Bayesian regression* implementations, and for the medical dataset we used a *Gradient Boosting Survival model* Pölsterl (2020).

### 8.3.1 Multivariate Generated Data Test

In order to test if *CID* adjusts the permutation ranking into the correct one, we took 2000 samples from a multivariate distribution with the following marginal distributions: $X_1 \sim Uni(0,1)$, $X_3 \sim$ Gamma(1.5,2), $X_4 \sim \text{Beta}(0.5, 0.5)$, $X_2 \sim X_3 \cdot X_4$, $X_5 \sim -\text{Exponential}(0.2)$, $X_6 \sim \sin(X_4)$ and $X_7 \sim X_8 \cdot X_9 + (1 - X_8) \cdot X_{10}$ with $X_8 \sim \text{Bin}(1, 0.7)$ and $X_9 \sim \mathcal{N}(-5, 1)$, $X_{10} \sim \mathcal{N}(5, 1)$. Consider also the binning values: $\mathbf{b} = [0, 0.375, 0.5, 0.575, 0.625, 0.7, 0.775, 0.85, 0.975]$. We then defined the outcome variable as: $y_j = \sum_{i=1}^{7} x_i \cdot \mathbb{I}\left(b_i \leq u_j < b_{i+1}\right) + \left(\sum_{k=2}^{4} x_k\right) \cdot \mathbb{I}\left(b_8 \leq u_j < b_9\right) + \left(\sum_{l=5}^{6} x_l\right) \cdot \mathbb{I}\left(u_j \geq b_9\right)$, where $u$ is an observation of $U \sim \text{Uni}(0, 1)$. The true importances are thus: $\mathcal{I}_1 \geq \mathcal{I}_2 \geq \mathcal{I}_3 \geq \mathcal{I}_4 \geq \mathcal{I}_5 = \mathcal{I}_6 \geq \mathcal{I}_7$. We transformed the data into Gaussian using quantile information and chosen gaussian markov random fields to pair with *CID*. The graph was inferred using graphical lasso with a grid-search cross-validation to determine the optimal $l_1$ penalization parameter. To test the *CID* correction, we performed 200 Shuffle Splits with Extremely Randomized Trees and computed the *Gini importance* for each feature, as well as the *permutation importance*(PI). We then adjusted the feature importances using the *CID* algorithm and Bayesian Regression as $\phi$ (see assumption 8.2.1). You can compare the rankings in figure 8.3. As can be seen from the swarmplot in figure 8.3, with the exception of $X_1$, PI placed a nearly equal weight on all features, centered around zero, presumably due to the high feature covariance. The *CID* was able to rectify this and ranked the features in the right order. It also placed every feature importance at non-zero with a gap between unequally important features and similar importance for $X_5/X_6$, matching well the true importances. Moreover, notice how the *Gini importance* underestimated $X_3/X_1$, presumably because $X_2$ offers many quality splitting points due to the overlap and similarity with $X_3/X_4$.

### 8.3.2 Cardiovascular Event Prediction with Proteomics

#### Problem Introduction

Cardiovascular diseases (CVDs) are the number one cause of death globally. Identifying asymptomatic people with the highest cardiovascular (CV) risk remains a crucial challenge in preventing their first cardiac event. Clinically used risk algorithms offer limited accuracy Piepoli et al. (2016). Consequently, a substantial proportion of the general population at risk remains unidentified until their first clinical event. Hoogeveen and Belo Pereira et al. recently

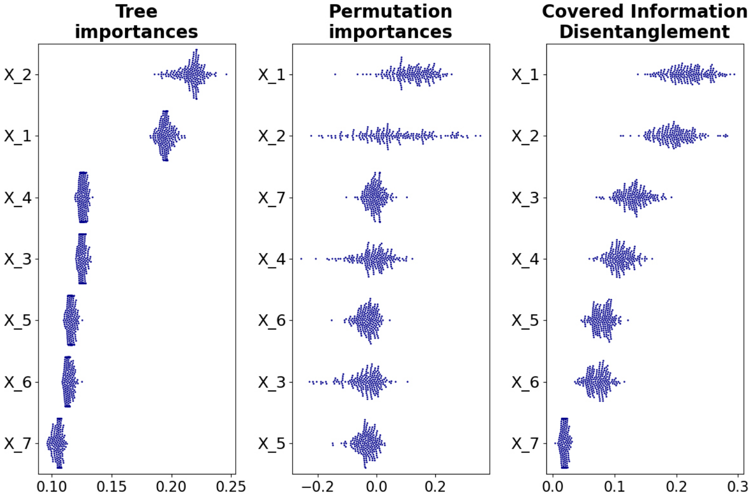


Figure 8.3: Comparison of the importance ranking on the multivariate gaussian dataset given by from left to right: *Tree importance* ( *Gini importance* ), *permutation importance*, *CID* importance. The feature order is given by the importance median. The ground truth is $\mathcal{I}_1 \geq \mathcal{I}_2 \geq \mathcal{I}_3 \geq \mathcal{I}_4 \geq \mathcal{I}_5 = \mathcal{I}_6 \geq \mathcal{I}_7$.

demonstrated increased efficacy in predicting primary events using protein-based models Hoogeveen et al. (2020). Since technical advances now allow for cheap and reproducible high-throughput proteomic analysis Assarsson et al. (2014), the field is prime for identifying new diagnostic markers or therapeutic targets, as well as developing new targeted protein panels to quickly and cheaply assess the risk of various diseases. The success of this endeavour is, of course, dependent on reliable feature importance identification.

The reason this dataset is a good candidate to test *CID*, is the "biological robustness" of living systems Kitano (2004); Stelling et al. (2004). Biological robustness describes a property of living systems whereby specific functions of the system are maintained despite external and internal perturbations. In proteomics, robustness is achieved in two ways: since protein structure is intimately related to function Schermann (2008), proteins with similar structure can exhibit similar functions, and proteins can be synthesized through different pathways in the metabolic network. This means two proteins located upstream the network relative to a third causing disease will have redundant information, and so do two proteins whose structure is similar

**Algorithm 4** *CID* Importance

```
Input: $\mathbf{X}_{M \times N}$, $\mathbf{y}$, $f$, $\Psi$, $\mathbf{G}$(optional)
Return: $\mathcal{I}(f)$
$\mathbf{S} \leftarrow$ SampleSubsets($\{1, ..., M\}$)
$\mathbf{e}(f) \leftarrow$ PermutationImportance($\mathbf{X}$, $\mathbf{y}$, $f$, $\mathbf{S}$)
$\mathbf{G} \leftarrow$ InferGraph([$\mathbf{X}$, $\mathbf{y}$])                ▷ Infer graph if not povided
$\Psi_\theta \leftarrow$Infer$MRF$Params($\Psi$, $\mathbf{X}$, $\mathbf{y}$)
$\mathbf{H}^\wedge \leftarrow$ ComputeMutualInfo($\mathbf{X}$, $\mathbf{y}$), $\mathbf{H}^\mathbf{c} \leftarrow \mathbf{0}$
$\mathcal{N} \leftarrow$ GetNeighbors($[\mathbf{X}, y]$, $\mathbf{G}$)
for $i$ in $[1, \ldots, N]$ do                                ▷ can be parallelized
    for $j$ in $[1, \ldots, M]$ do
        $\mathbf{d}, \mathbf{e}, \mathbf{F} \leftarrow$ Potentials($\Psi_\theta$, $\mathbf{X}, \mathbf{y}$, $i, j, \mathcal{N}_i, \mathcal{N}_y$)
        $\mathbf{H}^\mathbf{c}_i\,[j] \leftarrow \mathbf{H}^\wedge_i\,[j] - log\left(f \frac{\mathbf{d}^T \mathbf{F}\,\mathbf{e}}{\mathbf{d}^T \mathbf{F}_y \mathbf{F}_{x_i} \mathbf{e}}\right)$
    end for
end for
$H^{\mathbf{c}^+}, H^{\mathbf{c}^-}, H^{\wedge^+}, H^{\wedge^-} \leftarrow$ RedundSyn($\mathbf{H}^\wedge, \mathbf{H}^\mathbf{c}, \mathbf{S}$)
$\phi \leftarrow$ FitEntropyPI$\Big(H^{\mathbf{c}^+}, H^{\mathbf{c}^-}, H^{\wedge^+}, H^{\wedge^-}, \mathbf{e}(f)\Big)$
$\mathcal{I}(f) \leftarrow \mathbf{E}_{\sim p(\mathbf{s})}\left[\phi\left(0, H^{\mathbf{c}^-}, H^{\wedge^+}, H^{\wedge^-}\right)\right]$
```

(this is depicted in figure 8.4).

### 8.3.3 Dataset Description

The dataset consists of a selection of 822 seemingly healthy individuals in a nested case-control sample from the EPIC-Norfolk study Day et al. (1999). Seemingly healthy individuals were defined as study participants who did not report a history of CV disease. A total of 411 individuals who developed an acute myocardial infarction (either hospitalization or death) between baseline and follow-up through 2016 were selected, together with 411 seemingly healthy individuals who remained free of any CV disease during follow-up. In the original study, the authors demonstrate how predicting short-term events leads to a significant accuracy improvement Hoogeveen et al. (2020), presumably because the proteomic profile will change over time. We used the early-event prediction dataset, where we only included patients who suffered from an event earlier than 1500 days from measurement (total of 100 patients). We do not make the code for this analysis available due to data confidentiality.

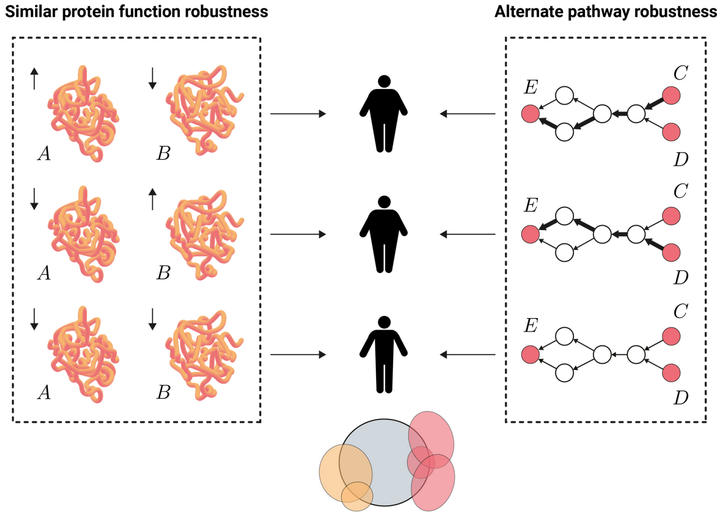

Similar protein function robustness
Alternate pathway robustness
A
B
A
B
A
B
E
C
D
E
C
D
E
C
D

Figure 8.4: Illustration of biological robustness for the event prediction with proteomics problem. On the left square, it is shown how the levels of two different proteins with similar structure (and hence, similar function) impact the outcome (obesity); on the right square, it is shown how two different proteins can influence the levels of a third outcome-related one through different pathways in the metabolic network; on the bottom, there is a Venn diagram representing the information overlap of the outcome (in gray) and the other proteins considered.

### 8.3.4 Importance Ranking Experiment Details

To evaluate the models' performance on days-to-event regression, we performed 100 shuffle splits and measured the mean square error on the test set. We used 5-fold cross-validation to select the optimal hyper-parameters of a Survival Gradient Boosting regressor Pölsterl (2020). To prevent overfitting, we pre-selected 50 proteins using univariate selection. We then compared the *CID* with permutation importance, *Univariate importance* , *SAGE* Covert, Lundberg, and Lee (2020), and *Tree importance* (*Gini importance*). We used GraphicalLasso (GL) for network inference in all our experiments and selected the $l_1$ regularization term using grid-search cross-validation. For the cardiovascular event survival analysis, we discretized the data into 10 bins.

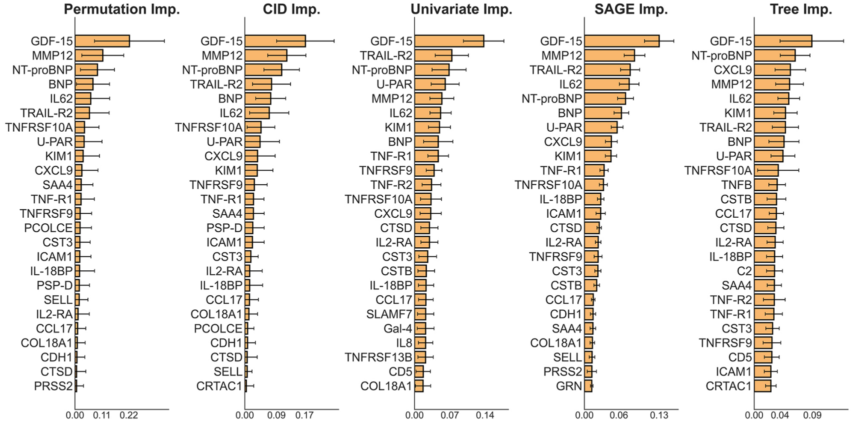


Figure 8.5: Importance rankings for cardiovascular event prediction using proteomics given by *permutation importance*, *CID*, *univariate importance*, *SAGE* and *tree importance* ( *Gini importance* )

For this experiment we used:

$$
\begin{aligned}
e_i(f, \mathbf{s}) \quad &= \phi_f \left( H^{\mathbf{c}^+}_{X_i}(\mathbf{s}), H^{\mathbf{c}^-}_{X_i}(\mathbf{s}), H_i^{\wedge^+}(\mathbf{s}), H_i^{\wedge^-}(\mathbf{s}) \right) \\
&= \mathcal{I}_i(f, \mathbf{s}) g \left( H^{\mathbf{c}^+}_{X_i}(\mathbf{s}) \right) \left( 1 - \frac{H^{\mathbf{c}^+}_{X_i}(\mathbf{s})}{H_i^{\wedge^+}(\mathbf{s})} \right), \\
g \left( H^{\mathbf{c}^+}_{X_i}(\mathbf{s}) \right) &= \begin{cases} c, & \text{if } H^{\mathbf{c}^+}_{X_i}(\mathbf{s}) > 0, \quad c \in [1, +\infty[ \\ 1, & \text{otherwise} \end{cases},
\end{aligned}
$$

that is, the *permutation importance* is modelled as the true importance weighted by the fraction of uncovered information (disregarding synergy) scaled by $c$. We then found $c$ using grid-search on the values: $1/c = [1.2, 1.4, 1.6, 1.8, 2, 2.2, 2.4]$. We removed data instances that contained values exceeding 4 times the standard deviation to achieve better discretization.

### Results

Overall, *CID* spreads the importance more evenly than Perm. imp. and aligns better with the Univariate ranking. Thus, this corroborates the hypothesis that Perm. imp. underrates correlated features. *CID* ranked TRAIL-R2, PSP-D, and IL2-RA two or more places higher, while it ranked SELL and

| Method | Corr. | MSE top feats | Avg. cycle time(s) |
|---|---|---|---|
| Perm. Imp. | 0.8697 | 0.2824 ± 0.0107 | 0.7756 ± 0.2173 |
| *CID* | **0.8787** | **0.2801** ± 0.0098 | 18.76 ± 7.7256 |
| Univar. Imp. | 0.8185 | 0.2947 ± 0.0090 | 0.0008 ± 0.0003 |
| *SAGE* | 0.8499 | 0.2858± 0.0064 | 42179 ± 3835 |
| Tree imp. | 0.7219 | 0.2900± 0.0087 | - |

Table 8.1: Correlation between subset model performance and the subset's sum of importances for each method (higher is better) and the mean squared error on top 10 to 35 features for each method (lower is better), as well as the average running time per cycle in seconds.

PCOLCE five and seven places lower, respectively.
**Gold Standard:** To establish a gold-standard analysis of the ranking, we asked world-renowned cardiovascular experts who commented on the comparison. TRAIL-R2 and GDF-15 were identified as the highest predictors of long-term mortality in patients with acute myocardial infarction in Skau et al. (2017). PSP-D has been identified as a strong clinical predictor of future adverse clinical outcome in stable patients with chronic heart failure in Brankovic et al. (2019). Il2-RA has been positively associated with all-cause mortality, CVD mortality, incident CVD, stroke, and heart failure in Durda et al. (2015). To date, SELL and PCOLCE have not been associated as major players in the development of cardiovascular disease.
**Quantitative measure:** In order to establish a quantitative measure of the ranking quality, we followed an approach similar to what is described in Covert, Lundberg, and Lee (2020), where multiple subsets of the data were selected, the models were re-trained for each subset and then for each subset and importance method we measured the correlation between the performance and the subset's sum of importances. We also computed the model performance when trained on the top 10 to 35 proteins of each method. We also report the average running time per cycle conducted on an 8-core Intel(R) Core(TM) i7-7700HQ CPU @ 2.81Ghz. The results are displayed in table 8.1 which shows *CID* outperformed the other methods on this dataset.

## 8.4 Discussion and Conclusion

*Permutation importance* is a popular algorithm used to equip black-box models with global explanations. It has the advantage of being easy to understand, but its validity suffers in the presence of covariates. We propose a novel framework (CID) to disentangle the shared information between covariates and show how using *Markov random field*s leads to tractability, making *permutation importance* competitive against methodologies where all marginal contributions of a feature are considered, such as *SHAP*. Due to network inference's complexity, we have only explored *graphical lasso* in conjunction with *Gaussian Markov random field*s. Although this particular implementation is attractive for its scalability and intuitiveness, it might lack sufficient expressive power to model more complex relationships between features.

Recently, A. Fisher proposed *model class reliance* (MCR), a method to estimate the range of variable importance for a pre-specified model class and shown how it can be computed as a series of convex optimization problems for model classes whose empirical loss is convex, although general computation procedures are still an open area of research (Fisher, Rudin, and Dominici, 2018). By learning a map between *permutation importance* and entropy terms, the importances retrieved by *CID* are less dependent on the specific fitted model than *permutation importance* or *SHAP*, but the map quality still relies on a consistent model behavior with regards to redundant entropy, as well as a good *MRF* approximation to the data distribution. The former might depend on the groups of features and thus future work includes modeling this map using graph methods on the inferred network, where the node features are the entropy terms. The latter could be improved by using a class of non-parametric *MRF*s with higher flexibility. Should these two problems be solved, then *CID* provides a truly model-agnostic feature importance framework while retaining the intuitiveness of *permutation importance*.

## 8.5 Ethical Statement

With an increasing reliance on using machine learning methods to research impactful domains such as biology and medicine, it is more important than ever to achieve model transparency and accurately determine feature relevance. In this work, we develop an efficient way to incorporate interactions when ranking variables. In the biomedical domain with thousands or millions of complex interactions among proteins, metabolites, genes, and so on, speed and correctness in determining the elements governing a given process are critical because they could significantly mitigate time, resources, and human lives lost. On the other hand, model transparency can also be exploited to develop adversarial examples or gain unwarranted access to protected systems/data.

## 8.6 Acknowledgments

We would like to thank Nick Nurmohamed for helping with the interpretation of the medical dataset results as well as Cláudia Pinhão for the graphics' design.

# 9

# Summary and discussion

Cardiovascular diseases are a group of disorders with known and unknown causes. Beyond the established causal factors, there are poorly understood mechanisms raising the risk in patients with low traditional risk factors. To develop new risk algorithms that can generalize well to the general population, we need models capable of capturing complex interactions between different physiological systems and thus encompass more than one event-causing mechanism. Models from the Machine Learning (ML) field are promising candidates to fill this gap due to their increased flexibility over traditional statistical approaches. The rapid expansion in feature size presents a challenge even for powerful ML models urging the need for biology-tailored algorithms to take advantage of all the available data. In **part I** of this thesis, we first demonstrate the improved prediction capabilities of combining ML models with high-throughput technology data compared to traditional methods. In **part II** we discuss novel ML algorithms to incorporate biological knowledge, integrate multiple domains, and mitigate feature discovery bias.

## 9.1 Part I: Applied Machine Learning in Clinical Research

Identifying cardiovascular risk in asymptomatic people remains a critical challenge in primary prevention with clinically used algorithms falling short of predictive accuracy. In chapter 2, we hypothesized that since plasma proteomics data capture a snapshot of the patient current physiological state, it should include several markers that are on their own and through interactions indicative of disease. Even though individual causal proteins have previously been incorporated in risk prediction algorithms, these resulted in modest improvements Piepoli et al. (2016), likely due to the high specificity of such proteins and the lack of interactions in the model. The ability of ML models to incorporate complex non-linear interactions should mitigate this issue. To test our hypothesis, we measured 368 proteins in 1524 subjects from independent prospective primary prevention cohorts used as derivation and validation sets, respectively. We trained three models: a proteins-only model, a model using clinical markers, and a combined model. The protein model outperformed the refitted clinical model in the derivation and the validation cohorts. The difference in predictive accuracy was the highest for 3-year event prediction, which is consistent with the idea that both lifestyle and interventions influence proteomics making this data type better suited for short-time prediction. The most important features (retrieved by the model Gini-importance) are consistent with previous findings. Thus, improvement is likely due to the methodology interactions' modeling capacity. This work is a stepping stone toward using multiplex markers coupled with ML to determine patient-specific risk.

Beyond primary prevention, patients with established cardiovascular disease have a high risk of subsequent CVD events, so in chapter 3 we set out to test whether the use of proteomics would yield similar prediction improvements in secondary prevention. We tested this hypothesis first on 870 patients who entered the SMART cohort for myocardial infarction, stroke, or transient ischaemic attack with a 10-year SMART risk score above 15% and validated the results on 700 subjects who underwent a carotid endarterectomy following a stroke or transient ischaemic attack from the Athero-Express cohort. As in primary prevention, we found the protein model significantly improved risk prediction compared to the clinical model. C-reactive protein (CRP) is used in clinical practice as stratifying marker to identify 'residual inflammatory risk', but it is unclear whether CRP reflects the entirety of inflammatory responses involved in atherogenesis. We tested this hypothesis by dividing the patients into below and above median CRP and repeated the event prediction analysis. The low CRP model retrieved four neutrophil-related proteins not present neither in the initial model nor in the high CRP model, suggesting a different inflammatory mechanism. Overall, the successful secondary event prediction using proteomics is further evidence to justify using targeted proteomics for risk prediction.

After establishing in the earlier chapters that proteomics improves risk prediction, the next step is to include more data types and perform pathway analysis to provide a more general overview of CVD's pathophysiology. In chapter 4, we used a multi-modal model jointly trained on genomics, transcriptomics, proteomics, and clinical measurements to predict all-cause mortality in the BIOSTAT-CHF cohort Voors et al. (2016). The cohort features an index cohort composed of 2,516 patients with worsening signs or symptoms of heart failure and a comparable validation cohort of 1,738 patients. After showing the model can predict with high accuracy, we inspected the top most important clinical features of which renal dysfunction-related markers were the most represented. Thus, since many pathways are involved in heart failure, we focused the pathway analysis on a combination of renal disease history, renal failure, and eGFR. We then computed correlations between the most predictive markers for all domains and selected those with the maximum correlation between each panel, starting with renal dysfunction. This process yielded a list of 29 features. Performing pathway over-representation analysis on these markers retrieved pathways related to cysteine-type peptidase activity and regulation of endopeptidase activity pathways. These pathways might become potential targets for therapy to decrease mortality in patients with heart failure and chronic kidney disease.

## 9.2 Part II: Novel Machine Learning Algorithms for Clinical Research

Despite the successful application of ML methods to predict CV risk in part I, we found some challenges that motivated the development of novel tailored algorithms. The significant improvement in risk prediction using proteomics raised the question: "Can we gain additional predictive performance by incorporating protein-protein interaction (PPI) networks?" Domain knowledge integration is particularly relevant in the Medical domain

since the data typically contains higher number of features compared to the number of examples. In chapter 5, we develop a novel graph kernel to map the data into a higher dimensional space where the dimensions are weighted protein interactions' combinations of different sizes, which we call *Graph Space Embedding*. What makes this graph kernel unique is that, for all dimensions, it compares the same edges on both graphs. This design is particular to our use case because the PPI networks are universal in humans. GSE achieves improved ichaemia prediction over the existing graph kernels (which, in turn, outperform the baseline method without PPI information) using a fraction of the computational resources.
Even though we tailored this algorithm to our specific needs, it should still be advantageous in any application where a universal feature interaction network is present for all instances in the data. There are many such systems in healthcare, though we have not tested the algorithm performance in other settings.

After demonstrating the predictive power of targeted proteomics in chapters 2 and 3, we adopted a multi-domain approach in chapter 4 to maximize heart failure prediction performance. The different domains provide perspectives on the patient's physiology, and therefore there is a larger opportunity to identify the patient's hidden states (like inflammation). However, we used a stacked framework to combine the different data modalities. In this setting, each dataset is passed to an independent model and the predictions are then combined by a meta-model to produce the final prediction. The datasets are therefore indirectly tied by the meta-model. However, these modalities form a system of tightly connected layers, so we hypothesized if sharing information across modalities would improve the final model. To this end, in chapter 6 we develop a novel multi-domain pre-processing technique called *Manifold Mixing* which uses the topology of each dataset to "deform" the others, effectively sharing information across all modalities. We demonstrate how this yields improved ichaemia predictive performance when using proteomics and clinical parameters compared to the standard stacking technique and a single model trained on the concatenated data. These results are promising but preliminary. Several improvements are possible such as using non-linear inter-domain maps and partitioning the data topologies into non-overlapping regions based on curvature level.

All the algorithmic efforts above focused on improving prediction, but perhaps an even more relevant ML contribution to Medical research is reliable feature discovery. The models capable of distilling useful information of large, complex, and heterogeneous data are becoming increasingly opaque. With the rapid expansion of both the Medical and ML fields, it is unlikely that a particular model will remain relevant over the years, underscoring the need for model-agnostic interpretation methodology. Model explanation techniques fall under the umbrella of local and global explanations, the former serving the purpose of improving trustworthiness in clinical practice while the latter can guide general research directions or inspire novel drug targets. LIME Ribeiro, Singh, and Guestrin (2016) is a popular local explainability method that first perturbs data points in random directions and then trains a simple interpretable model on the original complex model's predictions. Evaluating the model for random perturbations is a wasteful process since not all changes will significantly change the model's output, so in chapter 5, we extend this method by taking statistically even steps in the direction of maximum output change. The evenly spaced coverage of the output's support allows us to use interpolation methods and accurately reconstruct the output surface. This method could be particularly useful in models with expensive predictions since model evaluation is limited to the critical model behavior zones.

Global feature importance is what drove the pathophysiology discussions in part I. Because we used a tree-based model in all our analyses, we could retrieve the Gini-importance, which measures importance based on the total decrease in impurity for all nodes that include a split on the feature. Although this is a crucial advantage of tree-based models, this metric is known to be biased in favor of variables with many possible split points as well as correlated ones. A popular solution to the former is to use *permutation importance* (PI), a model-agnostic method where each feature's column is first independently randomized, and then the model performance on this modified dataset is compared to baseline. Intuitively, the more the performance suffers, the more the model relies on that particular feature for prediction. Nevertheless, this approach is still biased in the case of correlated features because the model still has partial access to the permuted feature's information. In chapter 7, we tackle this issue by computing a feature correlation matrix and then permute together all pairs of features exceeding a threshold correlation. Then, we define the importance as the sum of *permutation importance* values weighted by the correlations. We call this method *Pairwise Permutation Algorithm* (PPA). We show how this reduces the PI bias in a controlled toy dataset and then discuss the validity of the biological implications stemming from the PPA ranking.

PPA is a stepping stone toward reducing PI bias, but it is an incomplete solution since it only considers pairs of features. However, there are many higher-order feature interactions in biological data. The number of combinations for tuples larger than two rapidly explodes, so permuting groups of features for variable group size is not a viable option. In chapter 8, we treat this problem from an information-theoretic perspective. The problem with PI is that it only measures the joint information between the feature and the output that is not "overlapping" with the other features in the data. Since information theory provides metrics with properties analogous to set theory, we first quantify the amount of overlapping information between the feature of interest and the rest with respect to the output. We then build a map from information to the PI values' space and recover the true importance by setting the covered information of each point to zero and finding its map's image. Finally, because the information theory metrics are intractable, we prove how using Markov Random Fields significantly simplifies their computation leading to a scalable, truly unbiased, model agnostic feature importance method. To test its efficacy, we designed a dataset with groups of correlated non-normal random variables and demonstrate how CID can recover exactly the true importance values while both PI and Gini importance yield biased results. We call this framework *Covered Information Disentanglement* (CID). For real datasets, the true importance is usually unknown, and estimating it is precisely the goal, so we adopted two importance ranking metrics discussed in Covert, Lundberg, and Lee (2020):

1. Correlation between model performance trained on random feature subsets and sum of importances for said subsets
2. Average performance on top k important features

The second is not favorable for CID since there will likely be a high degree of overlapping information among the top k features, but it is still an indication of overall ranking quality. Nevertheless, it outperformed the other considered methods on both performance metrics when tested on a real proteomics dataset. We further discussed the validity of this ranking from a biological perspective and argued these are consistent with previous findings.

CID is perhaps the development with the highest potential impact on medical research in this thesis. However, there is a need for flexible Markov Random Fields capable of modeling heterogeneous data before it can get widespread adoption.

## 9.3 Concluding remarks

Predicting cardiovascular risk is a challenging task because it involves multiple complex biological pathways. In this thesis, we demonstrate how

Machine Learning (ML) models can distill relevant information for accurate risk prediction out of large, heterogeneous datasets. The rapid expansion of molecule measuring technology provides a unique opportunity to detect disease progression, but the growing feature to sample size ratio makes the Medical field especially problematic from a modeling perspective. We can mitigate this difficulty by designing algorithms tailored to the biological systems' characteristics or by integrating the vast domain knowledge available. In the future, the interplay between ML and Biology will accelerate Medical research and ML adoption in clinical practice, although reliable model explainability is a hard requirement for such developments.

# A
# Supplementary Material

| | **EPIC-case ($n = 411$)** | **EPIC-control ($n = 411$)** | **PLIC-case ($n = 351$)** | **PLIC-control ($n = 351$)** |
|---|---|---|---|---|
| Age (years) | $66 \pm 7.8$ | $62 \pm 7.7$ | $55 \pm 8.1$ | $54 \pm 8.2$ |
| Male gender | 282(68.6) | 254(61.8) | 117(33.3) | 116(33.0) |
| BMI (kg/m$^2$) | $26.8 \pm 3.7$ | $26.6 \pm 3.6$ | $26.9 \pm 4.2$ | $26.4 \pm 3.2$ |
| Systolic blood pressure (mmHg) | $144 \pm 19$ | $136 \pm 17$ | $134 \pm 17$ | $130 \pm 16$ |
| Diastolic blood pressure (mmHg) | $86 \pm 12$ | $83 \pm 11$ | $84 \pm 9$ | $82 \pm 9$ |
| Current smoker | 61(15) | 22(5.4) | 78(22.2) | 56(16.0) |
| Total cholesterol (mg/dL)[a] | $250 \pm 47$ | $243 \pm 43$ | $225 \pm 39$ | $220 \pm 38$ |
| HDL cholesterol (mg/dL)[a] | $50 \pm 14$ | $53 \pm 15$ | $55 \pm 15$ | $58 \pm 15$ |
| LDL cholesterol (mg/dL)[a] | $164 \pm 41$ | $157 \pm 39$ | $147 \pm 37$ | $142 \pm 35$ |
| Triglycerides (mg/dL)[b] | $168(115 - 239)$ | $151(106 - 222)$ | $102(66 - 143)$ | $86(61 - 119)$ |
| hsCRP (mg/L) | $2.1(1.1 - 5.0)$ | $1.3(0.7 - 2.9)$ | – | – |
| HbA1c (%) | $5.77 \pm 1.28$ | $5.38 \pm 0.79$ | – | – |
| Antidiabetic drug use baseline | – | – | 3(0.9) | 2(0.6) |
| Lipid lowering drug use baseline | 9(2.2) | 6(1.5) | – | – |
| Antihypertensive drug use baseline | 150(36.5) | 75(18.2) | 92(26.2) | 68(19.4) |
| Median time of follow-up (years) | $15.1(7.7 - 19.6)$ | $20.5(19.6 - 21.2)$ | $11.1(10.9 - 11.3)$ | $11.1(11.0 - 11.3)$ |
| Lipid lowering drug use baseline | 9(2.2) | 6(1.5) | – | – |
| Antihypertensive drug use baseline | 150(36.5) | 75(18.2) | 92(26.2) | 68(19.4) |
| Median time of follow-up (years) | $15.1(7.7 - 19.6)$ | $20.5(19.6 - 21.2)$ | $11.1(10.9 - 11.3)$ | $11.1(11.0 - 11.3)$ |

Table A.1: Values are n (%), mean ± standard deviation, or median (IQR) for skewed data.
BMI, body mass index; EPIC, European Prospective Investigation; HDL, high-density lipoprotein; hsCRP, high sensitivity C-reactive protein; IQR, inter-quartile range; LDL, low-density lipoprotein; PLIC, Progressione della Lesione Intimale Carotidea.

[a] To convert to mmol/L, divide with 38.7.

[b] To convert to mmol/L, divide with 88.6.

| Characteristic | Derivation cohort (SMART) | Validation cohort (AE) |
|---|---|---|
| Number of patients | 870 | 700 |
| Age (years) | 65(9) | 70(9) |
| Male sex | 657(75.5) | 479(68.4) |
| BMI (kg/m2) | 26.9 ± 3.9 | 26.2 ± 3.8 |
| Systolic blood pressure (mmHg) | 146 ± 22 | 152 ± 25 |
| Diastolic blood pressure (mmHg) | 82 ± 12 | 82 ± 31 |
| Active smoking | 299(34.4) | 81(20.2) |
| Total cholesterol (mmol/L) | 4.95 ± 1.22 | 4.31 ± 1.12 |
| HDL cholesterol (mmol/L) | 1.22 ± 0.36 | 1.10 ± 0.36 |
| LDL cholesterol (mmol/L) | 2.98 ± 1.07 | 2.43 ± 0.91 |
| Triglycerides (mmol/L) | 1.42(1.00 − 2.10) | 1.49(1.08 − 2.04) |
| C-reactive protein (mg/L) | 2.5(1.2 − 5.2) | 2.0(1.0 − 4.5) |
| Diabetes mellitus | 178(20.5) | 163(23.3) |
| Lipid-lowering therapy | 546(62.8) | 541(77.5) |
| Antihypertensive therapy | 578(66.4) | 509(72.9) |
| Follow-up time (years) | 7.98(4.61 − 12.16) | 3.00(2.17 − 3.10) |
| Recurrent ASCVD event | 263(30.2) | 130(18.6) |
| Myocardial infarction | 48(5.5) | 39(5.6) |
| Ischaemic stroke | 105(12.1) | 53(7.5) |
| Cardiovascular death | 110(12.6) | 38(5.4) |

Table A.2: Only primary recurrent ASCVD events are shown. Values are n (%), mean ± standard deviation, or median (IQR) for skewed data (triglycerides, C-reactive protein, and follow-up time). SMART, Second Manifestations of ARTerial disease; BMI, body mass index; ASCVD, atherosclerotic cardiovascular disease.

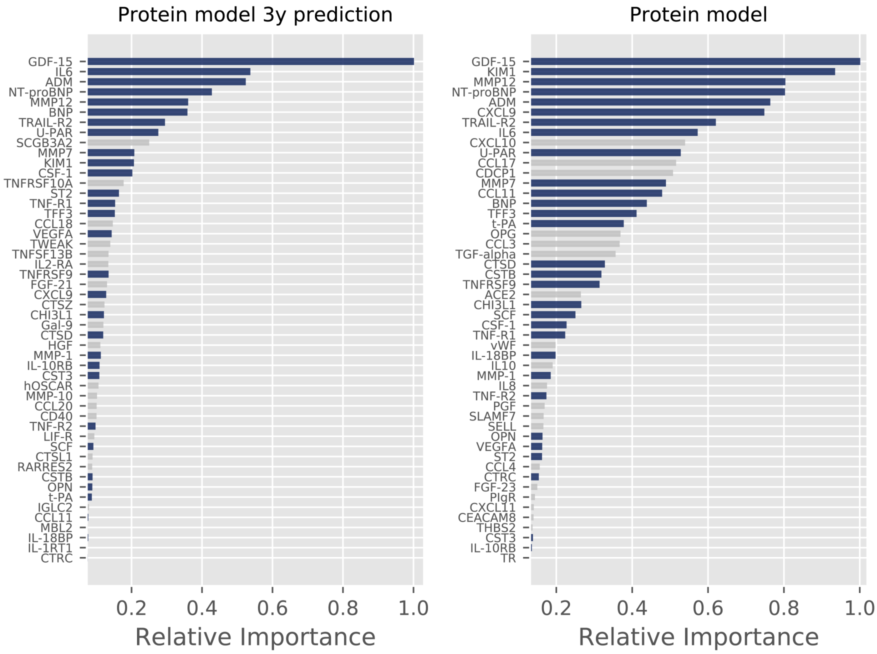


Figure A.1: **Overlap between predictive proteins**
Relative importance plots of 50 proteins predictive in the derivation cohort. Left: proteins predictive of events in the derivation cohort. Right: proteins predictive of events in derivation cohort <3 years. Proteins that overlap between two models are in blue.

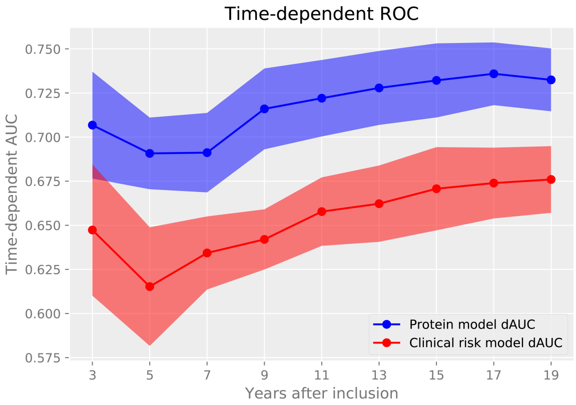


Figure A.2: **Time-dependent receiver operating characteristics** Receiver operating characteristics (dynamic AUC) and standard deviation of the survival models with a 2-year interval.

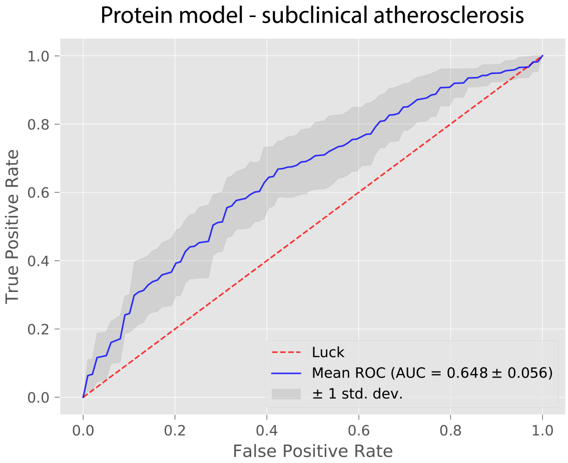


Figure A.3: **Protein model validation in asymptomatic atherosclerosis** Receiver operating characteristic of protein model in validation cohort on asymptomatic atherosclerosis.

# Bibliography

Abbas, A.; Aukrust, P.; Russell, D.; Krohg-Sørensen, K.; Almås, T.; Bundgaard, D.; Bjerkeli, V.; Sagen, E. L.; Michelsen, A. E.; Dahl, T. B.; Holm, S.; Ueland, T.; Skjelland, M.; and Halvorsen, B. 2014. Matrix metalloproteinase 7 is associated with symptomatic lesions and adverse events in patients with carotid atherosclerosis. *PloS one*, 9: e84935.

Acerbi, L.; and Ji, W. 2017. Practical Bayesian Optimization for Model Fitting with Bayesian Adaptive Direct Search. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 1836–1846.

Ahmad, T.; Desai, N.; Wilson, F.; Schulte, P.; Dunning, A.; Jacoby, D.; Allen, L.; Fiuzat, M.; Rogers, J.; Felker, G. M.; O'Connor, C.; and Patel, C. B. 2016. Clinical Implications of Cluster Analysis-Based Classification of Acute Decompensated Heart Failure and Correlation with Bedside Hemodynamic Profiles. *PloS one*, 11: e0145881.

Annemans, L.; Packard, C. J.; Briggs, A.; and Ray, K. K. 2018. 'Highest risk-highest benefit' strategy: a pragmatic, cost-effective approach to targeting use of PCSK9 inhibitor therapies. *European heart journal*, 39: 2546–2550.

Aratani, Y. 2018. Myeloperoxidase: Its role for host defense, inflammation, and neutrophil function. *Archives of biochemistry and biophysics*, 640: 47–52.

Argelaguet, R.; Velten, B.; Arnol, D.; Dietrich, S.; Zenz, T.; Marioni, J. C.; Buettner, F.; Huber, W.; and Stegle, O. 2018. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. 14.

Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.;

Richardson, J. E.; Ringwald, M.; Rubin, G. M.; and Sherlock, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25: 25–29.

Assarsson, E.; Holmquist, M. L. G.; Björkesten, J.; Thorsen, S. B.; Ekman, D.; Eriksson, A.; Dickens, E. R.; Ohlsson, S.; Edfeldt, G.; Andersson, A.-C.; Lindstedt, P.; Stenvang, J.; Gullberg, M.; and Fredriksson, S. 2014. Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. *PloS one*, 9: e95192.

Bakris, G. L.; Agarwal, R.; Anker, S. D.; Pitt, B.; Ruilope, L. M.; Rossing, P.; Kolkhof, P.; Nowack, C.; Schloemer, P.; Joseph, A.; Filippatos, G.; and Investigators, F. I. D. E. L. I. O.-D. K. D. 2020. Effect of Finerenone on Chronic Kidney Disease Outcomes in Type 2 Diabetes. *The New England journal of medicine*, 383: 2219–2229.

Bayes-Genis, A.; Liu, P. P.; Lanfear, D. E.; de Boer, R. A.; González, A.; Thum, T.; Emdin, M.; and Januzzi, J. L. 2020. Omics phenotyping in heart failure: the next frontier. *European heart journal*, 41: 3477–3484.

Bedani, R.; Rossi, E. A.; and Saad, S. M. I. 2013. Impact of inulin and okara on Lactobacillus acidophilus La-5 and Bifidobacterium animalis Bb-12 viability in a fermented soy product and probiotic survival under in vitro simulated gastrointestinal conditions. *Food Microbiol*, 34(2): 382-389.

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A theory of learning from different domains. *Mach. Learn.*, 79(1-2): 151–175.

Benavides, F.; Venables, A.; Poetschke Klug, H.; Glasscock, E.; Rudensky, A.; Gómez, M.; Martin Palenzuela, N.; Guénet, J. L.; Richie, E. R.; and Conti, C. J. 2001. The CD4 T cell-deficient mouse mutation nackt (nkt) involves a deletion in the cathepsin L (CtsI) gene. *Immunogenetics*, 53: 233–242.

Bereau, T.; Jr., R. A. D.; Tkatchenko, A.; and von Lilienfeld, O. A. 2018. Non-covalent interactions across organic and biological subsets of chemical space: Physics-based potentials parametrized from machine learning. *The Journal of Chemical Physics*, 148.

Bhatt, D. L.; Steg, P. G.; Miller, M.; Brinton, E. A.; Jacobson, T. A.; Ketchum, S. B.; Doyle, R. T.; Juliano, R. A.; Jiao, L.; Granowitz, C.; Tardif, J.-C.; Ballantyne, C. M.; and Investigators, R. E. D. U. C.

E.-I. T. 2019. Cardiovascular Risk Reduction with Icosapent Ethyl for Hypertriglyceridemia. *The New England journal of medicine*, 380: 11–22.

Bindea, G.; Mlecnik, B.; Hackl, H.; Charoentong, P.; Tosolini, M.; Kirilovsky, A.; Fridman, W.-H.; Pagès, F.; Trajanoski, Z.; and Galon, J. 2009. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics (Oxford, England)*, 25: 1091–1093.

Bleumink, G. S.; Schut, A. F. C.; Sturkenboom, M. C. J. M.; Deckers, J. W.; van Duijn, C. M.; and Stricker, B. H. C. 2004. Genetic polymorphisms and heart failure. *Genetics in medicine : official journal of the American College of Medical Genetics*, 6: 465–474.

Boekholdt, S. M.; Kuivenhoven, J.-A.; Wareham, N. J.; Peters, R. J. G.; Jukema, J. W.; Luben, R.; Bingham, S. A.; Day, N. E.; Kastelein, J. J. P.; and Khaw, K.-T. 2004. Plasma levels of cholesteryl ester transfer protein and the risk of future coronary artery disease in apparently healthy men and women: the prospective EPIC (European Prospective Investigation into Cancer and nutrition)-Norfolk population study. *Circulation*, 110: 1418–1423.

Bom, M. J.; Levin, E.; Driessen, R. S.; Danad, I.; Kuijk, C. C. V.; van Rossum, A. C.; Narula, J.; Min, J. K.; Leipsic, J. A.; Pereira, J. P. B.; Taylor, C. A.; Nieuwdorp, M.; Raijmakers, P. G.; Koenig, W.; Groen, A. K.; Stroes, E. S. G.; and Knaapen, P. 2019. Predictive value of targeted proteomics for coronary plaque morphology in patients with suspected coronary artery disease. *EBioMedicine*, 39: 109–117.

Borgwardt, K. M.; and Kriegel, H.-P. 2005. Shortest-path kernels on graphs. In *In Proceedings of the 5th International Conference on Data Mining*, 74–81.

Bradley, A. P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30: 1145–1159.

Brankovic, M.; Akkerhuis, K. M.; Mouthaan, H.; Constantinescu, A.; Caliskan, K.; van Ramshorst, J.; Germans, T.; Umans, V.; and Kardys, I. 2019. Utility of temporal profiles of new cardio-renal and pulmonary candidate biomarkers in chronic heart failure. *International journal of cardiology*, 276: 157–165.

Breiman, L. 2001. Random Forests. *Machine Learning*, 45(1): 5–32.

Brincks, E. L.; Kucaba, T. A.; James, B. R.; Murphy, K. A.; Schwertfeger, K. L.; Sangwan, V.; Banerjee, S.; Saluja, A. K.; and Griffith, T. S. 2015. Triptolide enhances the tumoricidal activity of TRAIL against renal cell carcinoma. *The FEBS journal*, 282: 4747–4765.

Brömme, D.; Panwar, P.; and Turan, S. 2016. Cathepsin K osteoporosis trials, pycnodysostosis and mouse deficiency models: Commonalities and differences. *Expert opinion on drug discovery*, 11: 457–472.

Canfora, E. E.; Jocken, J. W.; and Blaak, E. E. 2015. Short-chain fatty acids in control of body weight and insulin sensitivity. *Nat Rev Endocrinol.*, 11(10): 577–91.

Carlsson, A. C.; Ingelsson, E.; Sundström, J.; Carrero, J. J.; Gustafsson, S.; Feldreich, T.; Stenemo, M.; Larsson, A.; Lind, L.; and Ärnlöv, J. 2017. Use of Proteomics To Investigate Kidney Function Decline over 5 Years. *Clinical journal of the American Society of Nephrology : CJASN*, 12: 1226–1235.

Caruana, R.; Niculescu-Mizil, A.; Crew, G.; and Ksikes, A. 2004. Ensemble selection from libraries of models. In *Twenty-first international conference on Machine learning - ICML '04. ACM*, volume 18.

Cederlund, M.; Deronic, A.; Pallon, J.; Sørensen, O. E.; and Åkerström, B. 2015. A1M/α1-microglobulin is proteolytically activated by myeloperoxidase, binds its heme group and inhibits low density lipoprotein oxidation. *Frontiers in physiology*, 6: 11.

Chang, S. Y.; Kim, D.-H.; and Han, M. J. 2010. Physicochemical and Sensory Characteristics of Soy Yogurt Fermented with Bifidobacterium breve K-110, Streptococcus thermophilus 3781, or Lactobacillus acidophilus Q509011. *Food Science and Biotechnol*, 19(1): 107–113.

Chen, S.; Tang, Y.; and Zhou, X. 2019. Cystatin C for predicting all-cause mortality and rehospitalization in patients with heart failure: a meta-analysis. *Bioscience reports*, 39.

Chen, T.; and Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In *KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Cheng, X. W.; Shi, G.-P.; Kuzuya, M.; Sasaki, T.; Okumura, K.; and Murohara, T. 2012. Role for cysteine protease cathepsins in heart disease: focus on biology and mechanisms with clinical implication. *Circulation*, 125: 1551–1562.

Cobb, M. 2017. 60 years ago, Francis Crick changed the logic of biology. *PLoS biology*, 15: e2003243.

Cohen, M. R. 2015. *The New Chinese Medicine Handbook: An Innovative Guide to Integrating Eastern Wisdom with Western Practice for Modern Healing*. Fair Winds Press.

Consortium, G. 2013. The Genotype-Tissue Expression (GTEx) project. *Nature genetics*, 45: 580–585.

Consortium, G. 2015. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science (New York, N.Y.)*, 348: 648–660.

Consortium, G. O. 2021. The Gene Ontology resource: enriching a GOld mine. *Nucleic acids research*, 49: D325–D334.

Cook, G. M.; Wells, J. E.; and Russell, J. B. 1994. Ability of Acidaminococcus Fermentans to oxidize trans-aconitate and decrease the accumulation of tricarballylate, a toxic end product of ruminal fermentation. *Appl Environ Microbiol.*, 60(7): 2533–7.

Corder, G. W.; and Foreman, D. I. 2014. *Nonparametric statistics. A step-by-step approach*. Hoboken, NJ: John Wiley & Sons, 2nd ed. edition. ISBN 978-1-118-84031-3; 978-1-118-84042-9.

Cover, T. M.; and Thomas, J. A. 2012. *Elements of Information Theory*. John Wiley & Sons.

Covert, I.; Lundberg, S.; and Lee, S.-I. 2020. Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33.

Crawford, D. C.; Bailey, J. N. C.; Miskimen, K.; Miron, P.; McCauley, J. L.; Sedor, J. R.; O'Toole, J. F.; and Bush, W. S. 2018. Somatic T-cell Receptor Diversity in a Chronic Kidney Disease PatientPopulation Linked to Electronic Health Records. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2017: 63–71.

Cui, Z.; Chang, H.; Shan, S.; and Chen, X. 2014. Generalized Unsupervised Manifold Alignment. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 27: Annual Conference on Neural*

*Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 2429–2437.

Damman, K.; Gori, M.; Claggett, B.; Jhund, P. S.; Senni, M.; Lefkowitz, M. P.; Prescott, M. F.; Shi, V. C.; Rouleau, J. L.; Swedberg, K.; Zile, M. R.; Packer, M.; Desai, A. S.; Solomon, S. D.; and McMurray, J. J. V. 2018. Renal Effects and Associated Outcomes During Angiotensin-Neprilysin Inhibition in Heart Failure. *JACC. Heart failure*, 6: 489–498.

Damman, K.; and Testani, J. M. 2015. The kidney in heart failure: an update. *European heart journal*, 36: 1437–1444.

Damman, K.; Valente, M. A. E.; Voors, A. A.; O'Connor, C. M.; van Veldhuisen, D. J.; and Hillege, H. L. 2014. Renal impairment, worsening renal function, and outcome in patients with heart failure: an updated meta-analysis. *European heart journal*, 35: 455–469.

Damman, K.; van der Harst, P.; Smilde, T. D. J.; Voors, A. A.; Navis, G.; van Veldhuisen, D. J.; and Hillege, H. L. 2012. Use of cystatin C levels in estimating renal function and prognosis in patients with chronic systolic heart failure. *Heart (British Cardiac Society)*, 98: 319–324.

Danad, I.; Raijmakers, P. G.; Driessen, R. S.; Leipsic, J.; Raju, R.; Naoum, C.; Knuuti, J.; Mäki, M.; Underwood, R. S.; Min, J. K.; Elmore, K.; Stuijfzand, W. J.; van Royen, N.; Tulevski, I. I.; Somsen, A. G.; Huisman, M. C.; van Lingen, A. A.; Heymans, M. W.; van de Ven, P. M.; van Kuijk, C.; Lammertsma, A. A.; van Rossum, A. C.; and Knaapen, P. 2017. Comparison of Coronary CT Angiography, SPECT, PET, and Hybrid Imaging for Diagnosis of Ischemic Heart Disease Determined by Fractional Flow Reserve. *JAMA cardiology*, 2: 1100–1107.

Day, N.; Oakes, S.; Luben, R.; Khaw, K.; Bingham, S.; Welch, A.; and Wareham, N. 1999. EPIC-Norfolk: study design and characteristics of the cohort. European Prospective Investigation of Cancer. *British journal of cancer*, 80 Suppl 1: 95–103.

de Boer, R. A.; Cao, Q.; Postmus, D.; Damman, K.; Voors, A. A.; Jaarsma, T.; van Veldhuisen, D. J.; Arnold, W. D.; Hillege, H. L.; and Silljé, H. H. W. 2013. The WAP four-disulfide core domain protein HE4: a novel biomarker for heart failure. *JACC. Heart failure*, 1: 164–169.

de Goffau, M. C.; Luopajärvi, K.; Knip, M.; and et al. 2013. Fecal microbiota composition differs between children with beta-cell autoimmunity and those without. *Diabetes*, 62(4): 1238–44.

Defferrard, M.; Bresson, X.; and Vandergheynst, P. 2016. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. In *In Advances in Neural Information Processing Systems.*, 3844–3852.

Delaneau, O.; Zagury, J.-F.; and Marchini, J. 2013. Improved whole-chromosome phasing for disease and population genetic studies. *Nature methods*, 10: 5–6.

Delporte, C.; Boudjeltia, K. Z.; Noyon, C.; Furtmüller, P. G.; Nuyens, V.; Slomianny, M.-C.; Madhoun, P.; Desmet, J.-M.; Raynal, P.; Dufour, D.; Koyani, C. N.; Reyé, F.; Rousseau, A.; Vanhaeverbeek, M.; Ducobu, J.; Michalski, J.-C.; Nève, J.; Vanhamme, L.; Obinger, C.; Malle, E.; and Van Antwerpen, P. 2014. Impact of myeloperoxidase-LDL interactions on enzyme activity and subsequent posttranslational oxidative modifications of apoB-100. *Journal of lipid research*, 55: 747–757.

Deo, R. C. 2015. Machine Learning in Medicine. *Circulation*, 132: 1920–1930.

Dirice, E.; Ng, R. W. S.; Martinez, R.; and et al. 2017. Isoform-selective inhibitor of histone deacetylase 3 (HDAC3) limits pancreatic islet infiltration and protects female nonobese diabetic mice from diabetes. *J Biol Chem*, 292(43): 17598–17608.

Donohoe, D. R.; Garge, N.; Zhang, X.; and et al. 2012. The microbiome and butyrate regulate energy metabolism and autophagy in the mammalian colon. *Cell Metab.*, 13(5): 517–526.

Dorresteijn, J. A. N.; Visseren, F. L. J.; Wassink, A. M. J.; Gondrie, M. J. A.; Steyerberg, E. W.; Ridker, P. M.; Cook, N. R.; van der Graaf, Y.; and Group, S. S. 2013. Development and validation of a prediction rule for recurrent vascular events based on a cohort study of patients with arterial disease: the SMART risk score. *Heart (British Cardiac Society)*, 99: 866–872.

Dragovich, M. A.; Adam, K.; Strazza, M.; Tocheva, A. S.; Peled, M.; and Mor, A. 2019. SLAMF6 clustering is required to augment T cell activation. *PloS one*, 14: e0218109.

Durda, P.; Sabourin, J.; Lange, E. M.; Nalls, M. A.; Mychaleckyj, J. C.; Jenny, N. S.; Li, J.; Walston, J.; Harris, T. B.; Psaty, B. M.; Valdar, W.; Liu, Y.; Cushman, M.; Reiner, A. P.; Tracy, R. P.; and Lange, L. A. 2015. Plasma Levels of Soluble Interleukin-2 Receptor α: Associations

With Clinical Cardiovascular Events and Genome-Wide Association Scan. *Arteriosclerosis, thrombosis, and vascular biology*, 35: 2246–2253.

Eikelboom, J. W.; Connolly, S. J.; Bosch, J.; Dagenais, G. R.; Hart, R. G.; Shestakovska, O.; Diaz, R.; Alings, M.; Lonn, E. M.; Anand, S. S.; Widimsky, P.; Hori, M.; Avezum, A.; Piegas, L. S.; Branch, K. R. H.; Probstfield, J.; Bhatt, D. L.; Zhu, J.; Liang, Y.; Maggioni, A. P.; Lopez-Jaramillo, P.; O'Donnell, M.; Kakkar, A. K.; Fox, K. A. A.; Parkhomenko, A. N.; Ertl, G.; Störk, S.; Keltai, M.; Ryden, L.; Pogosova, N.; Dans, A. L.; Lanas, F.; Commerford, P. J.; Torp-Pedersen, C.; Guzik, T. J.; Verhamme, P. B.; Vinereanu, D.; Kim, J.-H.; Tonkin, A. M.; Lewis, B. S.; Felix, C.; Yusoff, K.; Steg, P. G.; Metsarinne, K. P.; Cook Bruns, N.; Misselwitz, F.; Chen, E.; Leong, D.; Yusuf, S.; and Investigators, C. O. M. P. A. S. S. 2017. Rivaroxaban with or without Aspirin in Stable Cardiovascular Disease. *The New England journal of medicine*, 377: 1319–1330.

Endesfelder, D.; Engel, M.; Davis-Richardson, A. G.; and et al. 2016. Towards a functional hypothesis relating anti-islet cell autoimmunity to the dietary impact on microbial communities and butyrate production. *Microbiome*, 4: 17.

Fernández-Friera, L.; Fuster, V.; López-Melgar, B.; Oliva, B.; García-Ruiz, J. M.; Mendiguren, J.; Bueno, H.; Pocock, S.; Ibáñez, B.; Fernández-Ortiz, A.; and Sanz, J. 2017. Normal LDL-Cholesterol Levels Are Associated With Subclinical Atherosclerosis in the Absence of Risk Factors. *Journal of the American College of Cardiology*, 70: 2979–2991.

Figarska, S. M.; Gustafsson, S.; Sundström, J.; Ärnlöv, J.; Mälarstig, A.; Elmståhl, S.; Fall, T.; Lind, L.; and Ingelsson, E. 2018. Associations of Circulating Protein Levels With Lipid Fractions in the General Population. *Arteriosclerosis, thrombosis, and vascular biology*, 38: 2505–2518.

Fisher, A.; Rudin, C.; and Dominici, F. 2018. Model Class Reliance: Variable Importance Measures for any Machine Learning Model Class, from the "Rashomon" Perspective. *Computer Science.*

Force, U. P. S. T.; Curry, S. J.; Krist, A. H.; Owens, D. K.; Barry, M. J.; Caughey, A. B.; Davidson, K. W.; Doubeni, C. A.; Epling, J. W.; Kemper, A. R.; Kubik, M.; Landefeld, C. S.; Mangione, C. M.; Silverstein, M.; Simon, M. A.; Tseng, C.-W.; and Wong, J. B. 2018. Risk Assessment for Cardiovascular Disease With Nontraditional Risk Factors: US Preventive Services Task Force Recommendation Statement. *JAMA*, 320: 272–280.

Fout, A.; Byrd, J.; Shariat, B.; and Ben-Hur, A. 2017. Protein Interface Prediction using Graph Convolutional Networks. In *In Advances in Neural Information Processing Systems*, 6533–6542.

Franz, W. M.; Müller, O. J.; and Katus, H. A. 2001. Cardiomyopathies: from genetics to the prospect of treatment. *Lancet (London, England)*, 358: 1627–1637.

Frazier, P. I. 2018. A Tutorial on Bayesian Optimization.

Friedewald, W. T.; Levy, R. I.; and Fredrickson, D. S. 1972. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clinical chemistry*, 18: 499–502.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3).

Ganz, P.; Heidecker, B.; Hveem, K.; Jonasson, C.; Kato, S.; Segal, M. R.; Sterling, D. G.; and Williams, S. A. 2016. Development and Validation of a Protein-Based Risk Score for Cardiovascular Outcomes Among Patients With Stable Coronary Heart Disease. *JAMA*, 315: 2532–2541.

Gärtner, T.; Flach, P.; and Wrobel, S. 2003. On Graph Kernels: Hardness Results and Efficient Alternatives. In *Computational Learning Theory and Kernel Machines, 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003*, volume 129-143(3), 129–143.

Goff, D. C.; Lloyd-Jones, D. M.; Bennett, G.; Coady, S.; D'Agostino, R. B.; Gibbons, R.; Greenland, P.; Lackland, D. T.; Levy, D.; O'Donnell, C. J.; Robinson, J. G.; Schwartz, J. S.; Shero, S. T.; Smith, S. C.; Sorlie, P.; Stone, N. J.; Wilson, P. W. F.; Jordan, H. S.; Nevo, L.; Wnek, J.; Anderson, J. L.; Halperin, J. L.; Albert, N. M.; Bozkurt, B.; Brindis, R. G.; Curtis, L. H.; DeMets, D.; Hochman, J. S.; Kovacs, R. J.; Ohman, E. M.; Pressler, S. J.; Sellke, F. W.; Shen, W.-K.; Smith, S. C.; Tomaselli, G. F.; and of Cardiology/American Heart Association Task Force on Practice Guidelines, A. C. 2014. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation*, 129: S49–S73.

Grainger, D. J. 2004. Transforming growth factor beta and atherosclerosis: so far, so good for the protective cytokine hypothesis. *Arteriosclerosis, thrombosis, and vascular biology*, 24: 399–404.

Gregorutti, B.; Michel, B.; and Saint-Pierre, P. 2017. Correlation and variable importance in random forests. *Stat Comput*, 27: 659–678.

Group, K. D. I. G. O. K. B. P. W. 2021. KDIGO 2021 Clinical Practice Guideline for the Management of Blood Pressure in Chronic Kidney Disease. *Kidney international*, 99: S1–S87.

Grömping, U. 2009. Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *The American Statistician*, 63(4): 308–319.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Pedreschi, D.; Turini, F.; and Giannotti, F. 2018. Local Rule-Based Explanations of Black Box Decision Systems. *http://arxiv.org/abs/1805.10820v1.*

Haas, R.; Zelezniak, A.; Iacovacci, J.; Kamrad, S.; Townsend, S.; and Ralser, M. 2017. Designing and interpreting 'multi-omic' experiments that may change our understanding of biology. *Current opinion in systems biology*, 6: 37–45.

Hajiramezanali, E.; Dadaneh, S. Z.; Karbalayghareh, A.; Zhou, M.; and Qian, X. 2018. Bayesian multi-domain learning for cancer subtype discovery from next-generation sequencing count data. In Bengio, S.; Wallach, H. M.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 9133–9142.

Hartstra, A. V.; Bouter, K. E. C.; Bäckhed, F.; and Nieuwdorp, M. 2015. Insights into the role of the microbiome in obesity and type 2 diabetes. *Diabetes Care*, 38(1): 159–165.

Hastie, T.; Tibshirani, R.; and Friedman, J. 2009. The Elements of Statistical Learning.

Hellwege, J. N.; Velez Edwards, D. R.; Giri, A.; Qiu, C.; Park, J.; Torstenson, E. S.; Keaton, J. M.; Wilson, O. D.; Robinson-Cohen, C.; Chung, C. P.; Roumie, C. L.; Klarin, D.; Damrauer, S. M.; DuVall, S. L.; Siew, E.; Akwo, E. A.; Wuttke, M.; Gorski, M.; Li, M.; Li, Y.; Gaziano, J. M.; Wilson, P. W. F.; Tsao, P. S.; O'Donnell, C. J.; Kovesdy, C. P.; Pattaro, C.; Köttgen, A.; Susztak, K.; Edwards, T. L.; and Hung, A. M. 2019. Mapping eGFR loci to the renal transcriptome and phenome in the VA Million Veteran Program. *Nature communications*, 10: 3842.

Heywood, J. T.; Fonarow, G. C.; Costanzo, M. R.; Mathur, V. S.; Wigneswaran, J. R.; Wynne, J.; Committee, A. S. A.; and Investigators. 2007. High prevalence of renal dysfunction and its impact on outcome in 118,465 patients hospitalized with acute decompensated heart failure: a report from the ADHERE database. *Journal of cardiac failure*, 13: 422–430.

Hillege, H. L.; Girbes, A. R.; de Kam, P. J.; Boomsma, F.; de Zeeuw, D.; Charlesworth, A.; Hampton, J. R.; and van Veldhuisen, D. J. 2000. Renal function, neurohormonal activation, and survival in patients with chronic heart failure. *Circulation*, 102: 203–210.

Hira, Z. M.; and Gillies, D. F. 2015. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. *Adv Bioinformatics*.

Ho, J. E.; Lyass, A.; Courchesne, P.; Chen, G.; Liu, C.; Yin, X.; Hwang, S.-J.; Massaro, J. M.; Larson, M. G.; and Levy, D. 2018. Protein Biomarkers of Cardiovascular Disease and Mortality in the Community. *Journal of the American Heart Association*, 7.

Hoes, M. F.; Tromp, J.; Ouwerkerk, W.; Bomer, N.; Oberdorf-Maass, S. U.; Samani, N. J.; Ng, L. L.; Lang, C. C.; van der Harst, P.; Hillege, H.; Anker, S. D.; Metra, M.; van Veldhuisen, D. J.; Voors, A. A.; and van der Meer, P. 2020. The role of cathepsin D in the pathophysiology of heart failure and its potentially beneficial properties: a translational approach. *European journal of heart failure*, 22: 2102–2111.

Hoogeveen, R. M.; Nahrendorf, M.; Riksen, N. P.; Netea, M. G.; de Winther, M. P. J.; Lutgens, E.; Nordestgaard, B. G.; Neidhart, M.; Stroes, E. S. G.; Catapano, A. L.; and Bekkering, S. 2018. Monocyte and haematopoietic progenitor reprogramming as common mechanism underlying chronic inflammatory and cardiovascular diseases. *European heart journal*, 39: 3521–3527.

Hoogeveen, R. M.; Pereira, J. P. B.; Nurmohamed, N. S.; Zampoleri, V.; Bom, M. J.; Baragetti, A.; Boekholdt, S. M.; Knaapen, P.; Khaw, K.-T.; Wareham, N. J.; Groen, A. K.; Catapano, A. L.; Koenig, W.; Levin, E.; and Stroes, E. S. G. 2020. Improved cardiovascular risk prediction using targeted plasma proteomics in primary prevention. *European Heart Journal*.

Hooker, G.; and Mentch, L. 2019. Please stop permuting features An explanation and alternatives. *arXiv*. Preprint.

Hotelling, H. 1936. Relations Between Two Sets of Variates. 28: 321.

Howie, B. N.; Donnelly, P.; and Marchini, J. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics*, 5: e1000529.

Hun, H. J.; D., L. D.; and K., S. L. 2003. Learning high dimensional correspondences from low dimensional manifolds.

Hur, K. Y.; and Lee, M.-S. 2015. Gut microbiota and metabolic disorders. *Diabetes Metab J.*, 39(3): 198–203.

Husain, M.; Birkenfeld, A. L.; Donsmark, M.; Dungan, K.; Eliaschewitz, F. G.; Franco, D. R.; Jeppesen, O. K.; Lingvay, I.; Mosenzon, O.; Pedersen, S. D.; Tack, C. J.; Thomsen, M.; Vilsbøll, T.; Warren, M. L.; Bain, S. C.; and Investigators, P. . 2019. Oral Semaglutide and Cardiovascular Outcomes in Patients with Type 2 Diabetes. *The New England journal of medicine*, 381: 841–851.

Igarashi, P.; and Somlo, S. 2002. Genetics and pathogenesis of polycystic kidney disease. *Journal of the American Society of Nephrology : JASN*, 13: 2384–2398.

Ikeda, Y.; Imai, Y.; Kumagai, H.; Nosaka, T.; Morikawa, Y.; Hisaoka, T.; Manabe, I.; Maemura, K.; Nakaoka, T.; Imamura, T.; Miyazono, K.; Komuro, I.; Nagai, R.; and Kitamura, T. 2004. Vasorin, a transforming growth factor beta-binding protein expressed in vascular smooth muscle cells, modulates the arterial response to injury in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, 101: 10732–10737.

Ince, R. A. 2017. The Partial Entropy Decomposition: Decomposing multivariate entropy and mutual information via pointwise common surprisal. *arXiv preprint arXiv:1702.01591*.

Jensen, J. K. 2016. Risk Prediction: Are We There Yet? *Circulation*, 134: 1441–1443.

Jensen, L. J.; Kuhn, M.; et al. 2009. STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, 37(Database issue):D412-6.

Jernberg, T.; Hasvold, P.; Henriksson, M.; Hjelm, H.; Thuresson, M.; and Janzon, M. 2015. Cardiovascular risk in post-myocardial infarction patients: nationwide real world data demonstrate the importance of a long-term perspective. *European heart journal*, 36: 1163–1170.

Jia, L.; Li, D.; Feng, N.; and et al. 2017. Anti-diabetic effects of clostridium butyricum CGMCC0313 through promoting the growth of gut butyrate-producing bacteria in Type 2 Diabetic Mice. *Sci Rep*, 7(1): 7046.

Johnson, K. W.; Torres Soto, J.; Glicksberg, B. S.; Shameer, K.; Miotto, R.; Ali, M.; Ashley, E.; and Dudley, J. T. 2018. Artificial Intelligence in Cardiology. *Journal of the American College of Cardiology*, 71: 2668–2679.

Jonschkowski, R. 2015. Learning state representations with robotic priors. *Autonomous Robots*, 39: 407–428.

Joseph, L. J.; Chang, L. C.; Stamenkovich, D.; and Sukhatme, V. P. 1988. Complete nucleotide and deduced amino acid sequences of human and murine preprocathepsin L. An abundant transcript induced by transformation of fibroblasts. *The Journal of clinical investigation*, 81: 1621–1629.

Joshi, A.; Rienks, M.; Theofilatos, K.; and Mayr, M. 2021. Systems biology in cardiovascular disease: a multiomics approach. *Nature reviews. Cardiology*, 18: 313–330.

Kaasenbrood, L.; Boekholdt, S. M.; van der Graaf, Y.; Ray, K. K.; Peters, R. J. G.; Kastelein, J. J. P.; Amarenco, P.; LaRosa, J. C.; Cramer, M. J. M.; Westerink, J.; Kappelle, L. J.; de Borst, G. J.; and Visseren, F. L. J. 2016. Distribution of Estimated 10-Year Risk of Recurrent Vascular Events and Residual Risk in a Secondary Prevention Population. *Circulation*, 134: 1419–1429.

Kanda, H.; Wang, H. L.; Hesseltine, C. W.; and Warner, K. 1976. Yoghurt production by Lactobacillus fermentation of soybean milk. *Process biochemistry*, 11(4): 23.

Kang, U.; Tong, H.; and Sun, J. 2012. Fast Random Walk Graph Kernel. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, 828–838.

Kavčič, N.; Pegan, K.; and Turk, B. 2017. Lysosomes in programmed cell death pathways: from initiators to amplifiers. *Biological chemistry*, 398: 289–301.

Kempf, T.; Zarbock, A.; Widera, C.; Butz, S.; Stadtmann, A.; Rossaint, J.; Bolomini-Vittori, M.; Korf-Klingebiel, M.; Napp, L. C.; Hansen, B.; Kanwischer, A.; Bavendiek, U.; Beutel, G.; Hapke, M.; Sauer, M. G.; Laudanna, C.; Hogg, N.; Vestweber, D.; and Wollert, K. C. 2011. GDF-15 is an inhibitor of leukocyte integrin activation required for survival after myocardial infarction in mice. *Nature medicine*, 17: 581–588.

Kendall, M. G. 1938. A New Measure of Rank Correlation. 30: 81.

Kessler, T.; and Schunkert, H. 2021. Coronary Artery Disease Genetics Enlightened by Genome-Wide Association Studies. *JACC. Basic to translational science*, 6: 610–623.

Khan, S.; and Jena, G. 2015. The role of butyrate a histone deacetylase inhibitor in Diabetes Mellitus: Experimental evidence for therapeutic intervention. *Epigenomics*, 7(4): 669–80.

Khan, S.; and Jena, G. 2016. Sodium butyrate reduces insulin-resistance, fat accumulation and dyslipidemia in Type-2 Diabetic rat: A comparative study with metformin. *Chem Biol Interact.*, 254: 124–34.

Kingma, D. P.; and Ba, J. L. 2015. ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION. In *Proceedings of the 3rd International Conference for Learning Representations.*

Kipf, T. N.; and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. In *in Proceedings of the 6th International Conference on Learning Representations.*

Kitano, H. 2002. Computational systems biology. *Nature*, 420: 206–210.

Kitano, H. 2004. Biological robustness. *Nature Reviews Genetics*, 5(11): 826–37.

Koller, D.; and Friedman, N. 2009. *Probabilistic Graphical Models: Principles and Techniques.* MIT press.

Kramer, L.; Turk, D.; and Turk, B. 2017. The Future of Cysteine Cathepsins in Disease Management. *Trends in pharmacological sciences*, 38: 873–898.

Kriege, N.; and Mutzel, P. 2012. Subgraph Matching Kernels for Attributed Graphs. In *In Proceedings of the 29th International Conference on Machine Learning*, 291–298.

Kumar, A.; Sattigeri, P.; Wadhawan, K.; Karlinsky, L.; Feris, R. S.; Freeman, B.; and Wornell, G. W. 2018. Co-regularized Alignment for Unsupervised Domain Adaptation. In Bengio, S.; Wallach, H. M.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 9367–9378.

Kumar, E.; Venkatasubramanian, S.; Scheidegger, C.; and Friedler, S. A. 2020. Problems with Shapley-value-based explanations as feature importance measures. *Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria.*

Kwon, D. Y.; III, J. W. D.; Kim, H. J.; and Park, S. 2010. Antidiabetic effects of fermented soybean products on type 2 diabetes. *Nutrition Research*, 30(1): 1–13.

LaVange, L. M.; and Koch, G. G. 2006. Rank score tests. *Circulation*, 114: 2528–2533.

Lefranc, M.-P. 2014. Immunoglobulin and T Cell Receptor Genes: IMGT(®) and the Birth and Rise of Immunoinformatics. *Frontiers in immunology*, 5: 22.

Leopold, J. A.; and Loscalzo, J. 2018. Emerging Role of Precision Medicine in Cardiovascular Disease. *Circulation research*, 122: 1302–1315.

Lind, L.; Siegbahn, A.; Lindahl, B.; Stenemo, M.; Sundström, J.; and Ärnlöv, J. 2015. Discovery of New Risk Markers for Ischemic Stroke Using a Novel Targeted Proteomics Chip. *Stroke*, 46: 3340–3347.

Linde, J.; Schulze, S.; Henkel, S. G.; and Guthke, R. 2015. Data- and knowledge-based modeling of gene regulatory networks: an update. *EXCLI J.*, 14: 346–378.

Lindsey, M. L.; Mayr, M.; Gomes, A. V.; Delles, C.; Arrell, D. K.; Murphy, A. M.; Lange, R. A.; Costello, C. E.; Jin, Y.-F.; Laskowitz, D. T.; Sam, F.; Terzic, A.; Van Eyk, J.; Srinivas, P. R.; on Functional Genomics, A. H. A. C.; Translational Biology, C. o. C. C. C. o. C., Council on Cardiovascular Disease in the Young; Stroke Nursing, C. o. H.; and Council, S. 2015. Transformative Impact of Proteomics on Cardiovascular Health and Disease: A Scientific Statement From the American Heart Association. *Circulation*, 132: 852–872.

Lipovetsky, S.; and Conklin, M. 2001. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4): 319–330.

Liu, C.-L.; Guo, J.; Zhang, X.; Sukhova, G. K.; Libby, P.; and Shi, G.-P. 2018. Cysteine protease cathepsins in cardiovascular disease: from basic research to clinical trials. *Nature reviews. Cardiology*, 15: 351–370.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, 4765–4774.

Markousis-Mavrogenis, G.; Tromp, J.; Ouwerkerk, W.; Devalaraja, M.; Anker, S. D.; Cleland, J. G.; Dickstein, K.; Filippatos, G. S.; van der Harst, P.; Lang, C. C.; Metra, M.; Ng, L. L.; Ponikowski, P.; Samani, N. J.; Zannad, F.; Zwinderman, A. H.; Hillege, H. L.; van Veldhuisen, D. J.; Kakkar, R.; Voors, A. A.; and van der Meer, P. 2019. The clinical significance of interleukin-6 in heart failure: results from the BIOSTAT-CHF study. *European journal of heart failure*, 21: 965–973.

Marques, T. A.; Borchers, S. T. B.; Borchers, D. L.; Rexstad, E. G.; and Thomas, L. 2011. Distance Sampling.

Marques, T. A.; Buckland, S.; Borchers, D.; Rextad, E. A.; and Thomas, L. 2010. *International Encyclopedia of Statistical Science*. Springer.

Marso, S. P.; Daniels, G. H.; Brown-Frandsen, K.; Kristensen, P.; Mann, J. F. E.; Nauck, M. A.; Nissen, S. E.; Pocock, S.; Poulter, N. R.; Ravn, L. S.; Steinberg, W. M.; Stockner, M.; Zinman, B.; Bergenstal, R. M.; Buse, J. B.; Committee, L. S.; and Investigators, L. T. 2016. Liraglutide and Cardiovascular Outcomes in Type 2 Diabetes. *The New England journal of medicine*, 375: 311–322.

Meinshausen, N.; and Bühlmann, P. 2010. Stability selection. *Royal Statistical Society*, 72: 417–473.

Members:, A. F.; McDonagh, T. A.; Metra, M.; Adamo, M.; Gardner, R. S.; Baumbach, A.; Böhm, M.; Burri, H.; Butler, J.; Čelutkienė, J.; Chioncel, O.; Cleland, J. G. F.; Coats, A. J. S.; Crespo-Leiro, M. G.; Farmakis, D.; Gilard, M.; Heymans, S.; Hoes, A. W.; Jaarsma, T.; Jankowska, E. A.; Lainscak, M.; Lam, C. S. P.; Lyon, A. R.; McMurray, J. J. V.; Mebazaa, A.; Mindham, R.; Muneretto, C.; Francesco Piepoli, M.; Price, S.; Rosano, G. M. C.; Ruschitzka, F.; Kathrine Skibelund, A.; and Group, E. S. D. 2022.

2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: Developed by the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). With the special contribution of the Heart Failure Association (HFA) of the ESC. *European journal of heart failure*, 24: 4–131.

Moens, F.; Verce, M.; and Vuyst, L. D. 2017. Lactate- and acetate-based cross-feeding interactions betweeen selected strains of Lactobacilli Bifidobacteria and colon bacteria in the presence of inulin-type fructans. *Int J Food Microbiol.*, 241: 225–236.

Montagnana, M.; Danese, E.; Giudici, S.; Franchi, M.; Guidi, G. C.; Plebani, M.; and Lippi, G. 2011. HE4 in ovarian cancer: from discovery to clinical application. *Advances in clinical chemistry*, 55: 1–20.

Morita, H.; Seidman, J.; and Seidman, C. E. 2005. Genetic causes of human heart failure. *The Journal of clinical investigation*, 115: 518–526.

Mortensen, M. B.; Falk, E.; Li, D.; Nasir, K.; Blaha, M. J.; Sandfort, V.; Rodriguez, C. J.; Ouyang, P.; and Budoff, M. 2018. Statin Trials, Cardiovascular Events, and Coronary Artery Calcification: Implications for a Trial-Based Approach to Statin Therapy in MESA. *JACC. Cardiovascular imaging*, 11: 221–230.

Mueller, N. T.; Odegaard, A. O.; Gross, M. D.; Koh, W.-P.; Yu, M. C.; Yuan, J.-M.; and Pereira, M. A. 2012. Soy intake and risk of type 2 diabetes mellitus in Chinese Singaporeans. *European Journal of Nutrition*, 51: 1033–1040.

Murri, M.; Leiva, I.; Gomez-Zumaquero, J. M.; and et al. 2013. Gut microbiota in children with type 1 diabetes differs from that in healthy children: A case-control study. *BMC Medicine*, 11: 46.

Münger, L. H.; Trimigno, A.; Picone, G.; and et al. 2017. Identification of urinary food intake biomarkers for milk, cheese, and soy-based drink by untargeted GC-MS and NMR in healthy humans. *J Proetome Res*, 16(9): 3321–3335.

Nakagawa, S.; Nishihara, K.; Miyata, H.; Shinke, H.; Tomita, E.; Kajiwara, M.; Matsubara, T.; Iehara, N.; Igarashi, Y.; Yamada, H.; Fukatsu, A.; Yanagita, M.; Matsubara, K.; and Masuda, S. 2015. Molecular Markers of Tubulointerstitial Fibrosis and Tubular Cell Damage in Patients with Chronic Kidney Disease. *PloS one*, 10: e0136994.

Nakagawa, T.; Roth, W.; Wong, P.; Nelson, A.; Farr, A.; Deussing, J.; Villadangos, J. A.; Ploegh, H.; Peters, C.; and Rudensky, A. Y. 1998. Cathepsin L: critical role in Ii degradation and CD4 T cell selection in the thymus. *Science (New York, N.Y.)*, 280: 450–453.

Nerlekar, N.; Ha, F. J.; Cheshire, C.; Rashid, H.; Cameron, J. D.; Wong, D. T.; Seneviratne, S.; and Brown, A. J. 2018. Computed Tomographic Coronary Angiography-Derived Plaque Characteristics Predict Major Adverse Cardiovascular Events: A Systematic Review and Meta-Analysis. *Circulation. Cardiovascular imaging*, 11: e006973.

Niculescu-Mizil, A.; and Caruana, R. 2005. Predicting good probabilities with supervised learning.

Nidorf, S. M.; Fiolet, A. T. L.; Mosterd, A.; Eikelboom, J. W.; Schut, A.; Opstal, T. S. J.; The, S. H. K.; Xu, X.-F.; Ireland, M. A.; Lenderink, T.; Latchem, D.; Hoogslag, P.; Jerzewski, A.; Nierop, P.; Whelan, A.; Hendriks, R.; Swart, H.; Schaap, J.; Kuijper, A. F. M.; van Hessen, M. W. J.; Saklani, P.; Tan, I.; Thompson, A. G.; Morton, A.; Judkins, C.; Bax, W. A.; Dirksen, M.; Alings, M.; Hankey, G. J.; Budgeon, C. A.; Tijssen, J. G. P.; Cornel, J. H.; Thompson, P. L.; and Investigators, L. T. 2020. Colchicine in Patients with Chronic Coronary Disease. *The New England journal of medicine*, 383: 1838–1847.

Noureldein, M. H.; Bitar, S.; Youssef, N.; Azar, S.; and Eid, A. A. 2020. Butyrate modulates Diabetes-linked gut dysbiosis: Epigenetic and mechanistic modifications. *J Mol Endocrinol*, 64(1): 29–42.

Nowak, C.; Carlsson, A. C.; Östgren, C. J.; Nyström, F. H.; Alam, M.; Feldreich, T.; Sundström, J.; Carrero, J.-J.; Leppert, J.; Hedberg, P.; Henriksen, E.; Cordeiro, A. C.; Giedraitis, V.; Lind, L.; Ingelsson, E.; Fall, T.; and Ärnlöv, J. 2018. Multiplex proteomics for prediction of major cardiovascular events in type 2 diabetes. *Diabetologia*, 61: 1748–1757.

Ohashi, K.; Naruto, M.; Nakaki, T.; and Sano, E. 2003. Identification of interleukin-8 converting enzyme as cathepsin L. *Biochimica et biophysica acta*, 1649: 30–39.

Ojala, M.; and Garriga, G. C. 2010. Permutation Tests for Studying Classifier Performance. *Journal of Machine Learning Research*, 11: 1833–1863.

Olmastroni, E.; Baragetti, A.; Casula, M.; Grigore, L.; Pellegatta, F.; Pirillo, A.; Tragni, E.; and Catapano, A. L. 2019. Multilevel Models to Estimate

Carotid Intima-Media Thickness Curves for Individual Cardiovascular Risk Evaluation. *Stroke*, 50: 1758–1765.

Olmastroni, E.; Shlyakhto, E. V.; Konradi, A. O.; Rotar, O. P.; Alieva, A. S.; Boyarinova, M. A.; Baragetti, A.; Grigore, L.; Pellegatta, F.; Tragni, E.; Catapano, A. L.; and Casula, M. 2018. Epidemiology of cardiovascular risk factors in two population-based studies. *Atherosclerosis. Supplements*, 35: e14–e20.

Olson, O. C.; and Joyce, J. A. 2015. Cysteine cathepsin proteases: regulators of cancer progression and therapeutic response. *Nature reviews. Cancer*, 15: 712–729.

Opstal, T. S. J.; Hoogeveen, R. M.; Fiolet, A. T. L.; Silvis, M. J. M.; The, S. H. K.; Bax, W. A.; de Kleijn, D. P. V.; Mosterd, A.; Stroes, E. S. G.; and Cornel, J. H. 2020. Colchicine Attenuates Inflammation Beyond the Inflammasome in Chronic Coronary Artery Disease: A LoDoCo2 Proteomic Substudy. *Circulation*, 142: 1996–1998.

Ouwerkerk, W.; Zwinderman, A. H.; Ng, L. L.; Demissei, B.; Hillege, H. L.; Zannad, F.; van Veldhuisen, D. J.; Samani, N. J.; Ponikowski, P.; Metra, M.; Ter Maaten, J. M.; Lang, C. C.; van der Harst, P.; Filippatos, G.; Dickstein, K.; Cleland, J. G.; Anker, S. D.; and Voors, A. A. 2018. Biomarker-Guided Versus Guideline-Based Treatment of Patients With Heart Failure: Results From BIOSTAT-CHF. *Journal of the American College of Cardiology*, 71: 386–398.

Parker, B. L.; Calkin, A. C.; Seldin, M. M.; Keating, M. F.; Tarling, E. J.; Yang, P.; Moody, S. C.; Liu, Y.; Zerenturk, E. J.; Needham, E. J.; Miller, M. L.; Clifford, B. L.; Morand, P.; Watt, M. J.; Meex, R. C. R.; Peng, K.-Y.; Lee, R.; Jayawardana, K.; Pan, C.; Mellett, N. A.; Weir, J. M.; Lazarus, R.; Lusis, A. J.; Meikle, P. J.; James, D. E.; de Aguiar Vallim, T. Q.; and Drew, B. G. 2019. An integrative systems genetic analysis of mammalian lipid metabolism. *Nature*, 567: 187–193.

Pedregosa, F.; and et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Pencina, M. J.; D'Agostino, R. B.; Pencina, K. M.; Janssens, A. C. J. W.; and Greenland, P. 2012. Interpreting incremental value of markers added to risk prediction models. *American journal of epidemiology*, 176: 473–481.

Pencina, M. J.; D'Agostino, R. B.; and Steyerberg, E. W. 2011. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in medicine*, 30: 11–21.

Pereira, J.; Groen, A. K.; Stroes, E. S. G.; and Levin, E. 2019. Graph Space Embedding. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI 2019)*, 3253–3259.

Pereira, J.; Stroes, E. S. G.; Groen, A. K.; Zwinderman, A. H.; and Levin, E. 2020. Manifold Mixing for Stacked Regularization. In Cellier, D. K., P., ed., *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019. Communications in Computer and Information Science*, volume 1167, 444–452. Springer.

Perkovic, V.; Jardine, M. J.; Neal, B.; Bompoint, S.; Heerspink, H. J. L.; Charytan, D. M.; Edwards, R.; Agarwal, R.; Bakris, G.; Bull, S.; Cannon, C. P.; Capuano, G.; Chu, P.-L.; de Zeeuw, D.; Greene, T.; Levin, A.; Pollock, C.; Wheeler, D. C.; Yavin, Y.; Zhang, H.; Zinman, B.; Meininger, G.; Brenner, B. M.; Mahaffey, K. W.; and Investigators, C. T. 2019. Canagliflozin and Renal Outcomes in Type 2 Diabetes and Nephropathy. *The New England journal of medicine*, 380: 2295–2306.

Piek, A.; Meijers, W. C.; Schroten, N. F.; Gansevoort, R. T.; de Boer, R. A.; and Silljé, H. H. W. 2017. HE4 Serum Levels Are Associated with Heart Failure Severity in Patients With Chronic Heart Failure. *Journal of cardiac failure*, 23: 12–19.

Piepoli, M. F.; Hoes, A. W.; Agewall, S.; Albus, C.; andAlberico L Catapano, C. B.; Cooney, M.-T.; Corrà, U.; Cosyns, B.; Deaton, C.; Graham, I.; Hall, M. S.; Hobbs, F. D. R.; Løchen, M.-L.; Löllgen, H.; Marques-Vidal, P.; Perk, J.; Prescott, E.; Redon, J.; Richter, D. J.; Sattar, N.; Smulders, Y.; Tiberi, M.; van der Worp, H. B.; van Dis, I.; Verschuren, W. M. M.; Binno, S.; and Group, E. S. D. 2016. 2016 European Guidelines on cardiovascular disease prevention in clinical practice. *European Heart Journal*, 37: 2315–2381.

Pitt, B.; Filippatos, G.; Agarwal, R.; Anker, S. D.; Bakris, G. L.; Rossing, P.; Joseph, A.; Kolkhof, P.; Nowack, C.; Schloemer, P.; Ruilope, L. M.; and Investigators, F. I. G. A. R. O.-D. K. D. 2021. Cardiovascular Events with Finerenone in Kidney Disease and Type 2 Diabetes. *The New England journal of medicine*, 385: 2252–2263.

Pölsterl, S. 2020. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. *Journal of Machine Learning Research*, 21(212): 1–6.

Qin, J.; Li, Y.; ...; and Wang, J. 2012. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490: 55–60.

Qin, Y.; and Shi, G.-P. 2011. Cysteinyl cathepsins and mast cell proteases in the pathogenesis and therapeutics of cardiovascular diseases. *Pharmacology & therapeutics*, 131: 338–350.

Rajkomar, A.; Dean, J.; and Kohane, I. 2019. Machine Learning in Medicine. *N Engl J Med*, 380: 1347–1358.

Rangaswami, J.; Bhalla, V.; Blair, J. E. A.; Chang, T. I.; Costa, S.; Lentine, K. L.; Lerma, E. V.; Mezue, K.; Molitch, M.; Mullens, W.; Ronco, C.; Tang, W. H. W.; McCullough, P. A.; on the Kidney in Cardiovascular Disease, A. H. A. C.; and on Clinical Cardiology, C. 2019. Cardiorenal Syndrome: Classification, Pathophysiology, Diagnosis, and Treatment Strategies: A Scientific Statement From the American Heart Association. *Circulation*, 139: e840–e878.

Ray, K. K.; Molemans, B.; Schoonen, W. M.; Giovas, P.; Bray, S.; Kiru, G.; Murphy, J.; Banach, M.; De Servi, S.; Gaita, D.; Gouni-Berthold, I.; Hovingh, G. K.; Jozwiak, J. J.; Jukema, J. W.; Kiss, R. G.; Kownator, S.; Iversen, H. K.; Maher, V.; Masana, L.; Parkhomenko, A.; Peeters, A.; Clifford, P.; Raslova, K.; Siostrzonek, P.; Romeo, S.; Tousoulis, D.; Vlachopoulos, C.; Vrablik, M.; Catapano, A. L.; Poulter, N. R.; and study, D. V. I. N. C. I. 2021. EU-Wide Cross-Sectional Observational Study of Lipid-Modifying Therapy Use in Secondary and Primary Care: the DA VINCI study. *European journal of preventive cardiology*, 28: 1279–1289.

Reel, P. S.; Reel, S.; Pearson, E.; Trucco, E.; and Jefferson, E. 2021. Using machine learning approaches for multi-omics data analysis: A review. *Biotechnology advances*, 49: 107739.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. “Why Should I Trust You?” Explaining the Predictions of Any Classifier. In *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’16).*

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Anchors: High-precision model-agnostic explanations. *AAAI.*

Ridker, P. M. 2016. Residual inflammatory risk: addressing the obverse side of the atherosclerosis prevention coin. *European heart journal*, 37: 1720–1722.

Ridker, P. M. 2018. Clinician's Guide to Reducing Inflammation to Reduce Atherothrombotic Risk: JACC Review Topic of the Week. *Journal of the American College of Cardiology*, 72: 3320–3331.

Ridker, P. M.; Buring, J. E.; Rifai, N.; and Cook, N. R. 2007. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score. *JAMA*, 297: 611–619.

Ridker, P. M.; Danielson, E.; Fonseca, F. A. H.; Genest, J.; Gotto, A. M.; Kastelein, J. J. P.; Koenig, W.; Libby, P.; Lorenzatti, A. J.; MacFadyen, J. G.; Nordestgaard, B. G.; Shepherd, J.; Willerson, J. T.; Glynn, R. J.; and Group, J. S. 2008. Rosuvastatin to prevent vascular events in men and women with elevated C-reactive protein. *The New England journal of medicine*, 359: 2195–2207.

Ridker, P. M.; Everett, B. M.; Thuren, T.; MacFadyen, J. G.; Chang, W. H.; Ballantyne, C.; Fonseca, F.; Nicolau, J.; Koenig, W.; Anker, S. D.; Kastelein, J. J. P.; Cornel, J. H.; Pais, P.; Pella, D.; Genest, J.; Cifkova, R.; Lorenzatti, A.; Forster, T.; Kobalava, Z.; Vida-Simiti, L.; Flather, M.; Shimokawa, H.; Ogawa, H.; Dellborg, M.; Rossi, P. R. F.; Troquay, R. P. T.; Libby, P.; Glynn, R. J.; and Group, C. T. 2017a. Antiinflammatory Therapy with Canakinumab for Atherosclerotic Disease. *The New England journal of medicine*, 377: 1119–1131.

Ridker, P. M.; MacFadyen, J. G.; Thuren, T.; Everett, B. M.; Libby, P.; Glynn, R. J.; and Group, C. T. 2017b. Effect of interleukin-1β inhibition with canakinumab on incident lung cancer in patients with atherosclerosis: exploratory results from a randomised, double-blind, placebo-controlled trial. *Lancet (London, England)*, 390: 1833–1842.

Riley, R. D.; Ensor, J.; Snell, K. I. E.; Harrell, F. E.; Martin, G. P.; Reitsma, J. B.; Moons, K. G. M.; Collins, G.; and van Smeden, M. 2020. Calculating the sample size required for developing a clinical prediction model. *BMJ (Clinical research ed.)*, 368: m441.

Roweis, S. T.; and Saul, L. K. 2000. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290: 2323–2326.

Rudnicki, M.; Perco, P.; D Haene, B.; Leierer, J.; Heinzel, A.; Mühlberger, I.; Schweibert, N.; Sunzenauer, J.; Regele, H.; Kronbichler, A.; Mestdagh, P.; Vandesompele, J.; Mayer, B.; and Mayer, G. 2016. Renal microRNA- and RNA-profiles in progressive chronic kidney disease. *European journal of clinical investigation*, 46: 213–226.

Sabatine, M. S.; Giugliano, R. P.; Keech, A. C.; Honarpour, N.; Wiviott, S. D.; Murphy, S. A.; Kuder, J. F.; Wang, H.; Liu, T.; Wasserman, S. M.; Sever, P. S.; Pedersen, T. R.; Committee, F. S.; and Investigators. 2017. Evolocumab and Clinical Outcomes in Patients with Cardiovascular Disease. *The New England journal of medicine*, 376: 1713–1722.

Saleheen, D.; Scott, R.; Javad, S.; Zhao, W.; Rodrigues, A.; Picataggi, A.; Lukmanova, D.; Mucksavage, M. L.; Luben, R.; Billheimer, J.; Kastelein, J. J. P.; Boekholdt, S. M.; Khaw, K.-T.; Wareham, N.; and Rader, D. J. 2015. Association of HDL cholesterol efflux capacity with incident coronary heart disease events: a prospective case-control study. *The lancet. Diabetes & endocrinology*, 3: 507–513.

Samet, H. 2006. *Foundations of multidimensional and metric data structures.* Morgan Kaufm. Ser. Data Manag. Syst. Amsterdam: Elsevier; San Francisco, CA: Morgan Kaufmann. ISBN 0-12-369446-9.

Santamaría, I.; Velasco, G.; Cazorla, M.; Fueyo, A.; Campo, E.; and López-Otín, C. 1998. Cathepsin L2, a novel human cysteine proteinase produced by breast and colorectal carcinomas. *Cancer research*, 58: 1624–1630.

Santema, B. T.; Kloosterman, M.; Van Gelder, I. C.; Mordi, I.; Lang, C. C.; Lam, C. S. P.; Anker, S. D.; Cleland, J. G.; Dickstein, K.; Filippatos, G.; Van der Harst, P.; Hillege, H. L.; Ter Maaten, J. M.; Metra, M.; Ng, L. L.; Ponikowski, P.; Samani, N. J.; Van Veldhuisen, D. J.; Zwinderman, A. H.; Zannad, F.; Damman, K.; Van der Meer, P.; Rienstra, M.; and Voors, A. A. 2018. Comparing biomarker profiles of patients with heart failure: atrial fibrillation vs. sinus rhythm and reduced vs. preserved ejection fraction. *European heart journal*, 39: 3867–3875.

Saulnier, D. M. A.; Spinler, J. K.; Gibson, G. R.; and Versalovic, J. 2009. Mechanisms of Probiosis and Prebiosis: Considerations for enhanced functional foods. *Curr Opin Biotechnol.*, 20(2): 135–41.

Schermann, J. P. 2008. Amino Acids, Peptides and Proteins. *Spectroscopy and Modeling of Biomolecular Building Blocks.*

Schwartz, G. G.; Steg, P. G.; Szarek, M.; Bhatt, D. L.; Bittner, V. A.; Diaz, R.; Edelberg, J. M.; Goodman, S. G.; Hanotin, C.; Harrington, R. A.; Jukema, J. W.; Lecorps, G.; Mahaffey, K. W.; Moryusef, A.; Pordy, R.; Quintero, K.; Roe, M. T.; Sasiela, W. J.; Tamby, J.-F.; Tricoci, P.; White, H. D.; Zeiher, A. M.; Committees, O. O. U. T. C. O. M. E. S.; and Investigators. 2018.

Alirocumab and Cardiovascular Outcomes after Acute Coronary Syndrome. *The New England journal of medicine*, 379: 2097–2107.

Seferovic, J. P.; Claggett, B.; Seidelmann, S. B.; Seely, E. W.; Packer, M.; Zile, M. R.; Rouleau, J. L.; Swedberg, K.; Lefkowitz, M.; Shi, V. C.; Desai, A. S.; McMurray, J. J. V.; and Solomon, S. D. 2017. Effect of sacubitril/valsartan versus enalapril on glycaemic control in patients with heart failure and diabetes: a post-hoc analysis from the PARADIGM-HF trial. *The lancet. Diabetes & endocrinology*, 5: 333–340.

Senior, R. M.; Gresham, H. D.; Griffin, G. L.; Brown, E. J.; and Chung, A. E. 1992. Entactin stimulates neutrophil adhesion and chemotaxis through interactions between its Arg-Gly-Asp (RGD) domain and the leukocyte response integrin. *The Journal of clinical investigation*, 90: 2251–2257.

Sevenich, L.; Hagemann, S.; Stoeckle, C.; Tolosa, E.; Peters, C.; and Reinheckel, T. 2010. Expression of human cathepsin L or human cathepsin V in mouse thymus mediates positive selection of T helper cells in cathepsin L knock-out mice. *Biochimie*, 92: 1674–1680.

Shah, S. J.; Katz, D. H.; and Deo, R. C. 2014. Phenotypic spectrum of heart failure with preserved ejection fraction. *Heart failure clinics*, 10: 407–418.

Shah, S. J.; Katz, D. H.; Selvaraj, S.; Burke, M. A.; Yancy, C. W.; Gheorghiade, M.; Bonow, R. O.; Huang, C.-C.; and Deo, R. C. 2015. Phenomapping for novel classification of heart failure with preserved ejection fraction. *Circulation*, 131: 269–279.

Sharma, S.; and Taliyan, R. 2016. Histone deacetylase inhibitors: Future therapeutics for insulin resistance and Type 2 Diabetes. *Pharmacol Res*, 113(Pt A): 320-326.

Shervashidze, N.; Vishwanathan, S.; Petri, T. H.; Mehlhorn, K.; and Borgwardt, K. M. 2009. Efficient Graphlet Kernels for Large Graph Comparison. In *In Proceedings of the International Conference on Artificial Intelligence and Statistics*, 488–495.

Simons, P. C.; Algra, A.; van de Laak, M. F.; Grobbee, D. E.; and van der Graaf, Y. 1999. Second manifestations of ARTerial disease (SMART) study: rationale and design. *European journal of epidemiology*, 15: 773–781.

Singh, S.; Ribeiro, M. T.; and Guestrin, C. 2016. Programs as Black-Box Explanations. *http://arxiv.org/abs/1611.07579v1.*

Skau, E.; Henriksen, E.; Wagner, P.; Hedberg, P.; Siegbahn, A.; and Leppert, J. 2017. GDF-15 and TRAIL-R2 are powerful predictors of long-term mortality in patients with acute myocardial infarction. *European journal of preventive cardiology*, 24: 1576–1583.

Soehnlein, O.; and Libby, P. 2021. Targeting inflammation in atherosclerosis - from experimental insights to the clinic. *Nature reviews. Drug discovery*, 20: 589–610.

Stakos, D. A.; Kambas, K.; Konstantinidis, T.; Mitroulis, I.; Apostolidou, E.; Arelaki, S.; Tsironidou, V.; Giatromanolaki, A.; Skendros, P.; Konstantinides, S.; and Ritis, K. 2015. Expression of functional tissue factor by neutrophil extracellular traps in culprit artery of acute myocardial infarction. *European heart journal*, 36: 1405–1414.

Stelling, J.; Sauer, U.; Szallasi, Z.; III, F. J. D.; and Doyle, J. 2004. Robustness of Cellular Functions. *Cell*, 118(6): 675–85.

Stenemo, M.; Nowak, C.; Byberg, L.; Sundström, J.; Giedraitis, V.; Lind, L.; Ingelsson, E.; Fall, T.; and Ärnlöv, J. 2018. Circulating proteins as predictors of incident heart failure in the elderly. *European journal of heart failure*, 20: 55–62.

Strobl, C.; Boulesteix, A.; and et al, T. K. 2008. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(307).

Strobl, C.; Boulesteix, A.-L.; Zeileis, A.; and Hothorn, T. 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC bioinformatics*, 8: 25.

Štrumbelj, E.; and Kononenko, I. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3): 647–665.

Sudmant, P. H.; Rausch, T.; Gardner, E. J.; Handsaker, R. E.; Abyzov, A.; Huddleston, J.; Zhang, Y.; Ye, K.; Jun, G.; Fritz, M. H.-Y.; Konkel, M. K.; Malhotra, A.; Stütz, A. M.; Shi, X.; Casale, F. P.; Chen, J.; Hormozdiari, F.; Dayama, G.; Chen, K.; Malig, M.; Chaisson, M. J. P.; Walter, K.; Meiers, S.; Kashin, S.; Garrison, E.; Auton, A.; Lam, H. Y. K.; Mu, X. J.; Alkan, C.; Antaki, D.; Bae, T.; Cerveira, E.; Chines, P.; Chong, Z.; Clarke, L.; Dal, E.; Ding, L.; Emery, S.; Fan, X.; Gujral, M.; Kahveci, F.; Kidd, J. M.; Kong, Y.; Lameijer, E.-W.; McCarthy, S.; Flicek, P.; Gibbs, R. A.; Marth, G.; Mason, C. E.; Menelaou, A.; Muzny, D. M.; Nelson, B. J.; Noor, A.; Parrish, N. F.;

Pendleton, M.; Quitadamo, A.; Raeder, B.; Schadt, E. E.; Romanovitch, M.; Schlattl, A.; Sebra, R.; Shabalin, A. A.; Untergasser, A.; Walker, J. A.; Wang, M.; Yu, F.; Zhang, C.; Zhang, J.; Zheng-Bradley, X.; Zhou, W.; Zichner, T.; Sebat, J.; Batzer, M. A.; McCarroll, S. A.; Consortium, . G. P.; Mills, R. E.; Gerstein, M. B.; Bashir, A.; Stegle, O.; Devine, S. E.; Lee, C.; Eichler, E. E.; and Korbel, J. O. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526: 75–81.

Szklarczyk, D.; Gable, A. L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N. T.; Morris, J. H.; Bork, P.; Jensen, L. J.; and Mering, C. v. 2019. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, 47: D607–D613.

Ting, H. K. 2008. On the Amount of Information. *Theory of Probability & Its Applications*, 7(4): 439–447.

Toloşi, L.; and Lengauer, T. 2011. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, 27: 1986–1994.

Toma, I.; and McCaffrey, T. A. 2012. Transforming growth factor-β and atherosclerosis: interwoven atherogenic and atheroprotective aspects. *Cell and tissue research*, 347: 155–175.

Towbin, J. A.; and Bowles, N. E. 2002. The failing heart. *Nature*, 415: 227–233.

Trachana, K.; Bargaje, R.; Glusman, G.; Price, N. D.; Huang, S.; and Hood, L. E. 2018. Taking Systems Medicine to Heart. *Circulation research*, 122: 1276–1289.

Tromp, J.; Ouwerkerk, W.; Demissei, B. G.; Anker, S. D.; Cleland, J. G.; Dickstein, K.; Filippatos, G.; van der Harst, P.; Hillege, H. L.; Lang, C. C.; Metra, M.; Ng, L. L.; Ponikowski, P.; Samani, N. J.; van Veldhuisen, D. J.; Zannad, F.; Zwinderman, A. H.; Voors, A. A.; and van der Meer, P. 2018. Novel endotypes in heart failure: effects on guideline-directed medical therapy. *European heart journal*, 39: 4269–4276.

Tsivtsivadze, E.; Urban, J.; Geuvers, H.; and Heskes, T. 2011. Semantic Graph Kernels for Automated Reasoning. In *Proceedings of the Eleventh SIAM International Conference on Data Mining, SDM.*

Turk, V.; Stoka, V.; Vasiljeva, O.; Renko, M.; Sun, T.; Turk, B.; and Turk, D. 2012. Cysteine cathepsins: from structure, function and regulation to new frontiers. *Biochimica et biophysica acta*, 1824: 68–88.

Turk, V.; Turk, B.; and Turk, D. 2001. Lysosomal cysteine proteases: facts and opportunities. *The EMBO journal*, 20: 4629–4633.

van Lammeren, G. W.; den Ruijter, H. M.; Vrijenhoek, J. E. P.; van der Laan, S. W.; Velema, E.; de Vries, J.-P. P. M.; de Kleijn, D. P. V.; Vink, A.; de Borst, G. J.; Moll, F. L.; Bots, M. L.; and Pasterkamp, G. 2014. Time-dependent changes in atherosclerotic plaque composition in patients undergoing carotid surgery. *Circulation*, 129: 2269–2276.

van Nunen, L. X.; Zimmermann, F. M.; Tonino, P. A. L.; Barbato, E.; Baumbach, A.; Engstrøm, T.; et al. 2015. Fractional flow reserve versus angiography for guidance of PCI in patients with multivessel coronary artery disease (FAME): 5-year follow-up of a randomised controlled trial. *Lancet*, 386(10006): 1853–1860.

Vangay, P.; Hillmann, B. M.; and Knights, D. 2019. Microbiome Learning Repo (ML Repo): A public repository of microbiome regression and classification tasks. 8.

Vasiljeva, O.; Reinheckel, T.; Peters, C.; Turk, D.; Turk, V.; and Turk, B. 2007. Emerging roles of cysteine cathepsins in disease and their potential as drug targets. *Current pharmaceutical design*, 13: 387–403.

Verbree-Willemsen, L.; Zhang, Y.-N.; Ibrahim, I.; Ooi, S. B. S.; Wang, J.-W.; Mazlan, M. I.; Kuan, W. S.; Chan, S.-P.; Peelen, L. M.; Grobbee, D. E.; Richards, A. M.; Lam, C. S. P.; and de Kleijn, D. P. V. 2020. Extracellular vesicle Cystatin C and CD14 are associated with both renal dysfunction and heart failure. *ESC heart failure*, 7: 2240–2249.

Verhoeven, B. A. N.; Velema, E.; Schoneveld, A. H.; de Vries, J. P. P. M.; de Bruin, P.; Seldenrijk, C. A.; de Kleijn, D. P. V.; Busser, E.; van der Graaf, Y.; Moll, F.; and Pasterkamp, G. 2004. Athero-express: differential atherosclerotic plaque expression of mRNA and protein in relation to cardiovascular events and patient characteristics. Rationale and design. *European journal of epidemiology*, 19: 1127–1133.

Vishwanathan, S. N.; Schraudolph, N. N.; Kondor, R.; and Borgwardt, K. M. 2010. Graph Kernels. *Journal of Machine Learning Research*, 1201–1242.

Vock, D. M.; Wolfson, J.; Bandyopadhyay, S.; Adomavicius, G.; Johnson, P. E.; Vazquez-Benitez, G.; and O'Connor, P. J. 2016. Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting. *Journal of biomedical informatics*, 61: 119–131.

Voors, A. A.; Anker, S. D.; Cleland, J. G.; Dickstein, K.; Filippatos, G.; van der Harst, P.; Hillege, H. L.; Lang, C. C.; Ter Maaten, J. M.; Ng, L.; Ponikowski, P.; Samani, N. J.; van Veldhuisen, D. J.; Zannad, F.; Zwinderman, A. H.; and Metra, M. 2016. A systems BIOlogy Study to TAilored Treatment in Chronic Heart Failure: rationale, design, and baseline characteristics of BIOSTAT-CHF. *European journal of heart failure*, 18: 716–726.

Wang, C.; and Mahadevan, S. 2011. Heterogeneous Domain Adaptation Using Manifold Alignment. In Walsh, T., ed., *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, 1541–1546. IJCAI/AAAI.

Wang, T. J.; Gona, P.; Larson, M. G.; Tofler, G. H.; Levy, D.; Newton-Cheh, C.; Jacques, P. F.; Rifai, N.; Selhub, J.; Robins, S. J.; Benjamin, E. J.; D'Agostino, R. B.; and Vasan, R. S. 2006. Multiple biomarkers for the prediction of first major cardiovascular events and death. *The New England journal of medicine*, 355: 2631–2639.

Weng, S. F.; Reps, J.; Kai, J.; Garibaldi, J. M.; and Qureshi, N. 2017. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS one*, 12: e0174944.

(WHO), W. H. O. 2021. Fact sheet cardiovascular diseases. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Accessed: 2021-01-05.

Williams, P. L.; and Beer, R. D. 2010. Nonnegative Decomposition of Multivariate Information. *arXiv:1004.2515*.

Williams, S. A.; Kivimaki, M.; Langenberg, C.; Hingorani, A. D.; Casas, J. P.; Bouchard, C.; Jonasson, C.; Sarzynski, M. A.; Shipley, M. J.; Alexander, L.; Ash, J.; Bauer, T.; Chadwick, J.; Datta, G.; DeLisle, R. K.; Hagar, Y.; Hinterberg, M.; Ostroff, R.; Weiss, S.; Ganz, P.; and Wareham, N. J. 2019. Plasma protein patterns as comprehensive indicators of health. *Nature medicine*, 25: 1851–1857.

Witten, D. M.; Tibshirani, R.; and Hastie, T. 2009. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. 10: 515–534.

Wiviott, S. D.; Raz, I.; Bonaca, M. P.; Mosenzon, O.; Kato, E. T.; Cahn, A.; Silverman, M. G.; Zelniker, T. A.; Kuder, J. F.; Murphy, S. A.; Bhatt, D. L.; Leiter, L. A.; McGuire, D. K.; Wilding, J. P. H.; Ruff, C. T.; Gause-Nilsson, I. A. M.; Fredriksson, M.; Johansson, P. A.; Langkilde, A.-M.; Sabatine, M. S.; and Investigators, D. . 2019. Dapagliflozin and Cardiovascular Outcomes in Type 2 Diabetes. *The New England journal of medicine*, 380: 347–357.

WOLD, H. E. R. M. A. N. 1975. Path Models with Latent Variables: The NIPALS Approach.

Wolpert, D. H. 1992. Stacked generalization. *Neural Networks*, 5: 241–259.

Wong, H. S.-C.; Chang, C.-M.; Kao, C.-C.; Hsu, Y.-W.; Liu, X.; Chang, W.-C.; Wu, M.-S.; and Chang, W.-C. 2017. V-J combinations of T-cell receptor predict responses to erythropoietin in end-stage renal disease patients. *Journal of biomedical science*, 24: 43.

Wuttke, M.; Li, Y.; Li, M.; Sieber, K. B.; Feitosa, M. F.; Gorski, M.; Tin, A.; Wang, L.; Chu, A. Y.; Hoppmann, A.; Kirsten, H.; Giri, A.; Chai, J.-F.; Sveinbjornsson, G.; Tayo, B. O.; Nutile, T.; Fuchsberger, C.; Marten, J.; Cocca, M.; Ghasemi, S.; Xu, Y.; Horn, K.; Noce, D.; van der Most, P. J.; Sedaghat, S.; Yu, Z.; Akiyama, M.; Afaq, S.; Ahluwalia, T. S.; Almgren, P.; Amin, N.; Ärnlöv, J.; Bakker, S. J. L.; Bansal, N.; Baptista, D.; Bergmann, S.; Biggs, M. L.; Biino, G.; Boehnke, M.; Boerwinkle, E.; Boissel, M.; Bottinger, E. P.; Boutin, T. S.; Brenner, H.; Brumat, M.; Burkhardt, R.; Butterworth, A. S.; Campana, E.; Campbell, A.; Campbell, H.; Canouil, M.; Carroll, R. J.; Catamo, E.; Chambers, J. C.; Chee, M.-L.; Chee, M.-L.; Chen, X.; Cheng, C.-Y.; Cheng, Y.; Christensen, K.; Cifkova, R.; Ciullo, M.; Concas, M. P.; Cook, J. P.; Coresh, J.; Corre, T.; Sala, C. F.; Cusi, D.; Danesh, J.; Daw, E. W.; de Borst, M. H.; De Grandi, A.; de Mutsert, R.; de Vries, A. P. J.; Degenhardt, F.; Delgado, G.; Demirkan, A.; Di Angelantonio, E.; Dittrich, K.; Divers, J.; Dorajoo, R.; Eckardt, K.-U.; Ehret, G.; Elliott, P.; Endlich, K.; Evans, M. K.; Felix, J. F.; Foo, V. H. X.; Franco, O. H.; Franke, A.; Freedman, B. I.; Freitag-Wolf, S.; Friedlander, Y.; Froguel, P.; Gansevoort, R. T.; Gao, H.; Gasparini, P.; Gaziano, J. M.; Giedraitis, V.; Gieger, C.; Girotto, G.; Giulianini, F.; Gögele, M.; Gordon, S. D.; Gudbjartsson, D. F.; Gudnason, V.; Haller, T.; Hamet, P.; Harris, T. B.; Hartman, C. A.; Hayward, C.; Hellwege, J. N.; Heng, C.-K.; Hicks, A. A.; Hofer, E.; Huang, W.; Hutri-Kähönen, N.; Hwang, S.-J.; Ikram, M. A.; Indridason, O. S.; Ingelsson, E.; Ising, M.; Jaddoe, V. W. V.; Jakobsdottir, J.; Jonas, J. B.; Joshi, P. K.; Josyula, N. S.; Jung, B.; Kähönen, M.; Kamatani, Y.; Kammerer, C. M.; Kanai, M.; Kastarinen, M.; Kerr, S. M.; Khor, C.-C.; Kiess, W.; Kleber, M. E.; Koenig, W.; Kooner, J. S.; Körner, A.; Kovacs, P.; Kraja, A. T.; Krajcoviechova, A.; Kramer, H.; Krämer, B. K.; Kronenberg, F.; Kubo, M.; Kühnel, B.; Kuokkanen, M.; Kuusisto, J.; La Bianca, M.; Laakso, M.; Lange, L. A.; Langefeld, C. D.; Lee, J. J.-M.; Lehne, B.; Lehtimäki, T.; Lieb, W.; Study, L. C.; Lim, S.-C.; Lind, L.; Lindgren, C. M.; Liu, J.; Liu, J.; Loeffler, M.; Loos, R. J. F.; Lucae, S.; Lukas, M. A.; Lyytikäinen, L.-P.; Mägi, R.; Magnusson, P. K. E.; Mahajan, A.; Martin, N. G.; Martins, J.; März, W.; Mascalzoni, D.; Matsuda, K.; Meisinger, C.; Meitinger, T.; Melander, O.; Metspalu, A.; Mikaelsdottir, E. K.; Milaneschi, Y.; Miliku, K.; Mishra, P. P.; Program, V. A. M. V.; Mohlke, K. L.; Mononen, N.; Montgomery, G. W.; Mook-Kanamori, D. O.; Mychaleckyj, J. C.; Nadkarni, G. N.; Nalls, M. A.; Nauck, M.; Nikus, K.; Ning, B.; Nolte, I. M.; Noordam, R.; O'Connell, J.; O'Donoghue, M. L.; Olafsson, I.; Oldehinkel, A. J.; Orho-Melander, M.; Ouwehand, W. H.; Padmanabhan, S.; Palmer, N. D.; Palsson, R.; Penninx, B. W. J. H.; Perls, T.; Perola, M.; Pirastu, M.; Pirastu, N.; Pistis, G.; Podgornaia, A. I.; Polasek, O.; Ponte, B.; Porteous, D. J.; Poulain, T.; Pramstaller, P. P.; Preuss, M. H.; Prins, B. P.; Province, M. A.; Rabelink, T. J.; Raffield, L. M.; Raitakari, O. T.; Reilly, D. F.; Rettig, R.; Rheinberger, M.; Rice, K. M.; Ridker, P. M.; Rivadeneira, F.; Rizzi, F.; Roberts, D. J.; Robino, A.; Rossing, P.; Rudan, I.; Rueedi, R.; Ruggiero, D.; Ryan, K. A.; Saba, Y.; Sabanayagam, C.; Salomaa, V.; Salvi, E.; Saum, K.-U.; Schmidt, H.; Schmidt, R.; Schöttker, B.; Schulz, C.-A.; Schupf, N.; Shaffer, C. M.; Shi, Y.; Smith, A. V.; Smith, B. H.; Soranzo, N.; Spracklen, C. N.; Strauch, K.; Stringham, H. M.; Stumvoll, M.; Svensson, P. O.; Szymczak, S.; Tai, E.-S.; Tajuddin, S. M.; Tan, N. Y. Q.; Taylor, K. D.; Teren, A.; Tham, Y.-C.; Thiery, J.; Thio, C. H. L.; Thomsen, H.; Thorleifsson, G.; Toniolo, D.; Tönjes, A.; Tremblay, J.; Tzoulaki, I.; Uitterlinden, A. G.; Vaccargiu, S.; van Dam, R. M.; van der Harst, P.; van Duijn, C. M.; Velez Edward, D. R.; Verweij, N.; Vogelezang, S.; Völker, U.; Vollenweider, P.; Waeber, G.; Waldenberger, M.; Wallentin, L.; Wang, Y. X.; Wang, C.; Waterworth, D. M.; Bin Wei, W.; White, H.; Whitfield, J. B.; Wild, S. H.; Wilson, J. F.; Wojczynski, M. K.; Wong, C.; Wong, T.-Y.; Xu, L.; Yang, Q.; Yasuda, M.; Yerges-Armstrong, L. M.; Zhang, W.; Zonderman, A. B.; Rotter, J. I.; Bochud, M.; Psaty, B. M.; Vitart, V.; Wilson, J. G.; Dehghan, A.; Parsa, A.;

Chasman, D. I.; Ho, K.; Morris, A. P.; Devuyst, O.; Akilesh, S.; Pendergrass, S. A.; Sim, X.; Böger, C. A.; Okada, Y.; Edwards, T. L.; Snieder, H.; Stefansson, K.; Hung, A. M.; Heid, I. M.; Scholz, M.; Teumer, A.; Köttgen, A.; and Pattaro, C. 2019. A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nature genetics*, 51: 957–972.

Yamada, A.; Ishimaru, N.; Arakaki, R.; Katunuma, N.; and Hayashi, Y. 2010. Cathepsin L inhibition prevents murine autoimmune diabetes via suppression of CD8(+) T cell activity. *PloS one*, 5: e12894.

Yarden, Y.; and Pines, G. 2012. The ERBB network: at last, cancer therapy meets systems biology. *Nature reviews. Cancer*, 12: 553–563.

Yin, X.; Subramanian, S.; Hwang, S.-J.; O'Donnell, C. J.; Fox, C. S.; Courchesne, P.; Muntendam, P.; Gordon, N.; Adourian, A.; Juhasz, P.; Larson, M. G.; and Levy, D. 2014. Protein biomarkers of new-onset cardiovascular disease: prospective study from the systems approach to biomarker research in cardiovascular disease initiative. *Arteriosclerosis, thrombosis, and vascular biology*, 34: 939–945.

Yu, T.; Simoff, S.; and Jan, T. 2010. VQSVM: A case study for incorporating prior domain knowledge into inductive machine learning. *Neurocomputing*, 13-15: 2614–2623.

Yuan, T.; and Li, Y. 2017. Human Epididymis Protein 4 as a Potential Biomarker of Chronic Kidney Disease in Female Patients With Normal Ovarian Function. *Laboratory medicine*, 48: 238–243.

Yusuf, S.; Rangarajan, S.; Teo, K.; Islam, S.; Li, W.; Liu, L.; Bo, J.; Lou, Q.; Lu, F.; Liu, T.; Yu, L.; Zhang, S.; Mony, P.; Swaminathan, S.; Mohan, V.; Gupta, R.; Kumar, R.; Vijayakumar, K.; Lear, S.; Anand, S.; Wielgosz, A.; Diaz, R.; Avezum, A.; Lopez-Jaramillo, P.; Lanas, F.; Yusoff, K.; Ismail, N.; Iqbal, R.; Rahman, O.; Rosengren, A.; Yusufali, A.; Kelishadi, R.; Kruger, A.; Puoane, T.; Szuba, A.; Chifamba, J.; Oguz, A.; McQueen, M.; McKee, M.; Dagenais, G.; and Investigators, P. U. R. E. 2014. Cardiovascular risk and events in 17 low-, middle-, and high-income countries. *The New England journal of medicine*, 371: 818–827.

Zannad, F.; Ferreira, J. P.; Pocock, S. J.; Anker, S. D.; Butler, J.; Filippatos, G.; Brueckmann, M.; Ofstad, A. P.; Pfarr, E.; Jamal, W.; and Packer, M. 2020. SGLT2 inhibitors in patients with heart failure with reduced ejection fraction: a meta-analysis of the EMPEROR-Reduced and DAPA-HF trials. *Lancet (London, England)*, 396: 819–829.

Zhou, H.; Liu, Z.; Ning, S.; Yang, Y.; Lang, C.; Lin, Y.; and Ma, K. 2018. Leveraging prior knowledge for protein–protein interaction extraction with memory network. *Database*.

Zimmermann, F. M.; Ferrara, A.; Johnson, N. P.; van Nunen, L. X.; Escaned, J.; Albertsson, P.; et al. 2015. Deferral vs. performance of percutaneous coronary intervention of functionally non-significant coronary stenosis: 15-year follow-up of the DEFER trial. *Eur Heart J*, 36: 3182–3188.

Zinman, B.; Wanner, C.; Lachin, J. M.; Fitchett, D.; Bluhmki, E.; Hantel, S.; Mattheus, M.; Devins, T.; Johansen, O. E.; Woerle, H. J.; Broedl, U. C.; Inzucchi, S. E.; and Investigators, E. M. P. A.-R. O. U. T. C. O. M. E. 2015. Empagliflozin, Cardiovascular Outcomes, and Mortality in Type 2 Diabetes. *The New England journal of medicine*, 373: 2117–2128.

# Samenvatting

Medisch onderzoek ervaart een sterke toename in de hoeveelheid beschikbare data. Het enorme volume en de complexiteit van meetbare variabelen vormen een uitdaging voor het gebruik van traditionele statistische methoden en zijn voor geen enkel mens te bevatten. Het oplossen van dit probleem vereist krachtige modellen die in staat zijn om de interacties tussen variabelen vast te leggen en te onderzoeken hoe deze variabelen niet-lineair gerelateerd zijn aan de aandoening die wordt bestudeerd. In dit werk gebruiken we eerst Machine Learning (ML) -methoden om cardiovasculaire risico's beter te voorspellen, nieuwe biomarkers te identificeren en beschrijven we vervolgens nieuwe biologisch geïnspireerde algoritmen om enkele van de ondervonden uitdagingen op te lossen.

Aan de **klinische kant** demonstreren we hoe het combineren van gerichte plasma proteomics met ML-modellen het voorspellen van het risico op een eerste acuut myocardinfarct overtreft ten opzichte van traditionele klinische risicofactoren overtreft. We breiden dit resultaat vervolgens uit door aan te tonen dat deze combinatie ook superieur is bij het voorspellen van terugkerende atherosclerotische cardiovasculaire aandoeningen (acuut myocardinfarct, ischemische beroerte en cardiovasculaire sterfte). Ten slotte verdiepen we ons op de pathofysiologische paden die betrokken zijn bij de ontwikkeling van hartfalen met behulp van een multi-domein ML-model.

Aan de **technische kant**, aangezien proteomics een belangrijke rol heeft gespeeld in onze klinische onderzoeken en er informatie beschikbaar is over eiwit-eiwit interacties, willen we deze aanvullende kennis incorporeren om de prestaties van onze analyses te verbeteren zonder meer data toe te voegen. We hebben dit bereikt door een nieuwe "graph kernel" te ontwikkelen die alle instanties in de data in kaart brengt in een "infinite-dimensional space" waar de basisvectoren gewogen producten zijn van combinaties van eiwitinteracties van grootte 1 tot oneindig. Vervolgens laten we zien hoe dit de voorspelling van ischemie verbetert in vergelijking met het baseline algoritme en andere "graph kernels". Omdat deze methode de data projecteert in een "infinite-dimensional space" en transparantie een harde vereiste is in onze analyses, breiden we de "local model interpretability method" (LIME) uit. LIME bouwt een interpretable model dat is getraind op de voorspellingen

van de originele modellen voor random input perturbations. Onze methode verstoort de input in de richting van maximale outputverandering met in plaats daarvan equidistante outputveranderingen.

Er zijn verschillende onderling verbonden informatielagen in biologische systemen. Biologisch bewuste “multi-domain” modellen mogen daarom niet uitgaan van onafhankelijkheid tussen de domeinen. We hebben dit probleem aangepakt door eerst aan te nemen dat de gegevens zich in een lower-dimensional manifold bevinden (een ruimte die zich lokaal gedraagt als een Euclidean space) vanwege de vele value constraints die voortvloeien uit de interactiesvan alle componenten. Vervolgens vervormen we elk manifold lokaal met behulp van de topologie van de resterende domeinen. Door de getransformeerde datasets in een stacked model in te voeren, vergroten we effectief de informatiestroom tussen domeinen en laten we zien hoe dit de prestaties verbeterde.

Ten slotte bespreken we de kwestie van global model interpretability in black-boxmodellen om de belangrijkste variabelen te bepalen die van belang zijn voor de voorspellingen van het model. Het centrale onderzoeksobject is permutation importance dat de prestaties van het model vergelijkt wanneer de waarden van een specifiek kenmerk worden gerandomiseerd naar de baseline. Deze techniek is aantrekkelijk vanwege zijn intuïtiviteit en lineaire complexiteit. Deze benadering is echter biased voor features die nauw aan elkaar verwant zijn, aangezien het model nog steeds nauwkeurige voorspellingen kan doen met behulp van de non-permuted variabelen. Dit probleem is met name relevant in biologische systemen vanwege hun robuustheidseigenschap: de weerstand tegen verstoringen die wordt bereikt door elementen met vergelijkbare functies, alternatieve routes en interacties. We hebben dit probleem eerst benaderd door de correlatie tussen elk feature pair in overweging te nemen en vervolgens permuteren we feature pairs die een vooraf gedefinieerde correlatiedrempel overschrijden. Dit is echter een onvolledige oplossing omdat het geen betrekking heeft op higher-order interactions.

Onze laatste algoritmische ontwikkeling is misschien wel één van de belangrijkste bijdrage van dit proefschrift vanwege de voordelen ten opzichte van de state-of-the-art interpretability methods. Het doel is om het probleem van permutation importance bias te behandelen vanuit een set-theory perspective. Dit doen we door gebruik te maken van elementen uit de Information Theory waarvan de eigenschappen analoog zijn aan die in de set theory. Vervolgens beargumenteren we dat de waarden van permutation importance een functie zijn van de redundante/synergetische entropie tussen het kenmerk/de uitvoer en hoeveel van deze entropie wordt ”bedekt” door de andere kenmerken. Nadat we een kaart hebben geleerd tussen de entropietermen en de permutatiebelangen, kunnen we het werkelijke belang van een kenmerk, gemeten in de oorspronkelijke schaal voor verlies van het model, herstellen door de afbeelding van de functie te berekenen voor een bedekte entropie met nulwaarde. Ten slotte gebruiken we Markov Random Fields om de computational complexity van de methode te verminderen. Het belang van deze methode is om een global, model-agnostic, unbiased feature importance methode te produceren die alle feature interacties in overweging neemt met een korte computing time. Het uitbreiden van de transparantie in complexe modellen zou het medisch onderzoek aanzienlijk kunnen versnellen door het identificeren van belangrijke factoren die betrokken zijn bij complexe biologische processen.

# Abstract

Medical research has seen a stark increase in the amount of available data. The sheer volume and complexity of measured variables challenge the use of traditional statistical methods and are beyond the ability of any human to comprehend. Solving this problem demands powerful models capable of capturing the variable interactions and how those are non-linearly related to the condition under study. In this work, we first use Machine Learning (ML) methods to achieve better cardiovascular risk prediction/novel disease biomarker identification and then describe novel bio-inspired algorithms to solve some of the encountered challenges.

On the **clinical side**, we start by demonstrating how combining targeted plasma proteomics with ML models outperforms traditional clinical risk factors in predicting the risk of first-time acute myocardial infarction. We then extend this result by showing this combination is also superior when predicting recurrent atherosclerotic cardiovascular disease (acute myocardial infarction, ischaemic stroke, and cardiovascular death). Finally, we shed some light on the pathophysiological pathways involved in heart failure development using a multi-domain ML model.

On the **technical side**, since proteomics played a significant role in our clinical investigations and there is information on protein-protein interactions available, we would like to incorporate this additional knowledge to boost the performance of our analyses without adding more data. We achieved this by developing a novel graph kernel that maps all instances in the data into an infinite-dimensional space where the basis vectors are weighted products of protein interactions' combinations of size 1 to infinity. We then show how this improves the prediction of ischaemia compared to the baseline algorithm and other graph kernels. Because this method projects the data into an infinite-dimensional space and transparency is a hard requirement in our analyses, we extend the local model interpretability method LIME. LIME builds an interpretable model trained on the original models' predictions for random input perturbations. Our method perturbs the input in the direction of maximum output change with equidistant output changes instead.

There are several inter-connected layers of information in biological systems. Biologically aware multi-domain models should therefore not assume inter-domain independence. We tackled this problem by first assuming the data lies in a lower-dimensional manifold (a space that locally behaves like Euclidean space) due to the many value constraints stemming from all the components' interactions. We then locally deform each manifold using the topology of the remaining domains. By plugging the transformed datasets in a stacked model, we effectively increase inter-domain information flow and show how this improved performance.

Finally, we address the issue of global model interpretability in black-box models to uncover the most important variables governing the model prediction. The central study object is permutation importance which compares the model's performance when a specific feature's values are randomized to baseline. This technique is attractive because of its intuitiveness and linear complexity. However, this approach is biased toward closely related features since the model can still achieve accurate predictions using the non-permuted variables. This issue is particularly relevant in biological systems due to their robustness property: the resistance against perturbations achieved through elements with similar functions, alternative pathways, and interactions. We first mitigated this problem by considering the correlation between each feature pair and then permuting feature pairs exceeding a pre-defined correlation threshold. However, this is an incomplete solution since it does not address higher-order interactions.

Our final algorithmic development is perhaps one of the most significant contributions to this thesis for its advantages over the state-of-the-art interpretability methods. The goal is to treat the problem of permutation importance bias from a set-theory perspective. We do this by using elements of Information Theory whose properties are analogous to those in set theory. We then argue that the values of permutation importance are a function of the redundant/synergistic entropy between the feature/output and how much of this entropy is ”covered” by the other features. After learning a map between the entropy terms and the permutation importance values, we can recover the true feature importance measured in the original model loss scale by computing the function's image for a zero-valued covered entropy. Finally, we use Markov Random Fields to mitigate the method's computational complexity. The significance of this method is that it offers a global, model-agnostic, truly unbiased feature importance method that considers all feature interactions with a fast computing time. Increasing the transparency of complex models could significantly speed up Medical research by discovering which key players are involved in intricate biological processes.

# List of symbols

| | |
|---|---|
| **General math notation** | |
| $\mathbb{I}(x)$ | Indicator function, $\mathbb{I}(x) = 1$ if $x$ is true, else 0 |
| $\infty$ | Infinity |
| $\rightarrow$ | Tends towards, e.g. $n \rightarrow \infty$ |
| $\propto$ | Proportional to |
| $\lvert x \rvert$ | Absolute value |
| $\lvert \mathcal{S} \rvert$ | Size (cardinality) of a set |
| $n!$ | Factorial function |
| $\nabla$ | Vector of first derivatives |
| $\equiv$ | Defined as |
| $\mathcal{O}(\cdot)$ | Big-O: roughly means order of magnitude |
| $\mathbb{R}$ | The real numbers |
| $\approx$ | Approximately equal to |
| $\text{argmax}_x f(x)$ | Argmax: the value of $x$ that maximize $f$ |
| $\Gamma(x)$ | Gamma function, $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$ |
| $exp(x)$ | Exponential function $e^x$ |
| $\mathcal{X}$ | A set from which values are drawn (e.g. $\mathcal{X} = \mathbb{R}^D$) |
| $\{\cdot\}$ | Set notation |
| **Linear Algebra notation** | |
| $< \cdot, \cdot >$ | Inner-product |
| $< \cdot, \cdot >_F$ | Frobenius inner product |
| $\mathbf{A}_{:,j}$ | the $j$th column of the matrix |
| $\mathbf{A}_{i,:}$ | the $i$th row of the matrix |
| **Probability/ML notation** | |
| $X \perp Y$ | $X$ is independent of $Y$ |
| $X \perp Y \mid Z$ | $X$ is independent of $Y$ given $Z$ |
| $X \not\perp Y$ | $X$ is not independent of $Y$ |
| $X \sim p$ | $X$ is distributed according to $p$ |
| $\mathbf{E}[X]$ | Expected value of $X$ |
| $\mathbf{E}_q[X]$ | Expected value of $X$ wrt distribution $q$ |
| $\boldsymbol{\theta}$ | Parameter vector |
| $\boldsymbol{\Sigma}$ | Covariance matrix |
| $J(\boldsymbol{\theta})$ | Cost function |
| $\mathcal{N}(\mu, \sigma^2)$ | Normal distribution with mean $\mu$ and standard deviation $\sigma$ |
| $\mathcal{N}(X)$ | neighbors of r.v. $X$, i.e. $X \not\perp \mathcal{N}(X)$ |

# Portfolio

| | Year | ECTS |
|---|---|---|
| **Courses** | | |
| -Data visualization with ggplot2 | 2020 | 0.25 |
| -Data processing in Shell | 2020 | 0.25 |
| -Experimental design in Python | 2020 | 0.25 |
| -Python Data science toolbox | 2020 | 0.5 |
| **Professional work** | | |
| -Campina research project: modelling infant formula fat content's impact on the child | 2019-2022 | 10 |
| -TNO research project: predicting glucose | 2019 | 5 |
| -Cargill project: Data processing pipeline | 2021 | 5 |
| **Presentations** | | |
| -"Protein Space Embedding Kernel for Plaque Volume Prediction" poster presentation at Workshop on Computational Biology at joint ICML/IJCAI/ECAI/AAMAS conference | 2018 | 0.5 |
| -"Graph Space Embedding" - Invited oral presentation at BioSB 2019 | 2019 | 0.5 |
| -"Graph Space Embedding" - oral and poster presentation at IJCAI 2019 | 2019 | 1 |
| -"Bio-inspired algorithms for cardiovascular risk prediction" invited oral presentation at 87th EAS | 2020 | 0.5 |
| -"Biology guided ML for multi-omics analysis" presentation at ACS symposium | 2020 | 1 |
| -"Covered Information Disentanglement: Correcting Permutation Feature Importance in the Presence of Covariates" poster presentation at Machine Learning in Computational Biology (MLCB20) | 2020 | 0.5 |
| -"Covered Information Disentanglement: Model Transparency via Unbiased Permutation Importance" oral presentation and poster presentation at AAAI22 | 2022 | 1.5 |
| **Conferences and Workshops** | | |
| -ICML 2018 | 2018 | 0.5 |
| -ICML 2018 "Variational Bayes and Beyond: Bayesian Inference for Big Data" tutorial | 2018 | 0.5 |
| -Workshop on Computational Biology at joint ICML/IJCAI/ECAI/AAMAS conference | 2018 | 0.5 |
| -IJCAI 2019 | 2019 | 0.5 |
| -IJCAI 2019 "Hands-On Deep Learning with TensorFlow 2.0" tutorial | 2019 | 0.5 |
| -MLCB20 | 2020 | 0.5 |
| -AAAI22 | 2022 | 0.5 |

# Acknowledgments

Throughout the successes and failures of this journey, several people made it possible for me to keep moving forward and reach my destination. No matter where our paths will lead or whether they cross again, I will carry you in my heart with great kindness, for you were an indispensable light in the middle of the storm.

Critical to my success was the momentum I gained in the first year, as this fueled the excitement and curiosity to keep innovating and pushing the boundaries. The motor behind this accelerating train were my supervisors Evgeni Levin and Erik Stroes. Evgeni gave me the freedom and motivation to pursue my own ideas and heightened my ambition to polish them into high-quality publications. The life cycle of original work is generally an unforgiving one. However, the calmness and simplicity with which he handled the failures, made all the frustration more bearable and prevented the fear of further exploration. Erik's contagious enthusiasm, outstanding medical expertise, and capacity to see the big picture made for a clear project roadmap, preventing missteps on the path toward high-quality medical research.

Similarly, I would like to thank Renate and Nick for our fruitful discussions and your relentless quest for rigor and quality. Despite our different expertise and opinions, the synergy between our backgrounds resulted in carefully crafted work which I can confidently say I am proud of. I would also like to thank Troy, Wouter, and Adriaan for our collaboration which, after several refinements, yielded fruitful results.

Thanks Rens and Charlotte for our brief but effective collaborations!

Dear members of my defense committee, prof. dr. A.H. Zwinderman, prof. dr. W.J. de Jonge, dr. C.J. Veenman, dr. H.J. Herrema and prof. dr. P.A.N. Bosman, I highly appreciate your willingess to read/evaluate my thesis and to serve as opponents during my defense.

Delicious as coffee may be, what made ”Koffietijd” an oasis to bask in, was the company of my colleagues. The delightful conversations I had with Kim, Ulrika, Torsten, Koen, Veera, Sultan, Xiang, Manon, Eduard, Anne Linde, and others while the aroma of roasted coffee beans lingered in the air restored my energy to keep on, and I am very grateful for it. A special thanks to my dear paranymphs Kim and Ulrika for helping with the defense's preparations and our endless conversations. I hope we keep in touch in the future! Thank you Sultan for your career advice and emotional support during the hardest phases in my Ph.D. trajectory.

It is remarkable the profound impact someone thousands of kilometers away can have on you. My frequent calls with my parents, grandmother, sister and brother in law grounded me and allowed me to feel at home even with the distance separating us. Thank you for always being there. I am blessed to have friends that are just as caring and loving as family. Thank you Ana and Mário, João, Alexandre, Diogo and Faca. Your unconditional support and cheering lifted my spirit and prevented me from going under during the darkest times.

Living in a foreign country generally means separation from your family and friends. Over the years, I was lucky to find friends in some of the most wonderful people I have ever met. Oliver, Andrea, Judith, Dhruv and Nino, Diogo and Marcel, I will forever cherish our time together. Our dinners, parties, and hangouts made life as an expat not only much easier but a delightful and rewarding experience by itself. Fortunately, I had the luck to have a Portuguese family present with me when Diogo and Catarina moved to the Netherlands. Thank you for your constant presence. Your support was invaluable.

Finally, none of this would be possible without you my dear Cláudia. You have always been my source of inspiration and guidance. You picked me up countless times, you paved the way for numerous insights, and you dissipated my constant self-doubt, but most of all, your love is an infinite well of comfort, energy, and bliss.

# About the Author

João Belo Pereira was born in Entroncamento, Portugal on the 24th of November 1992. After completing his high school studies in 2009 at Escola Secundária do Entroncamento, he enrolled in an integrated Master in Biomedical Engineering at Instituto Superior Técnico (Lisbon, Portugal) in 2010.

During his Master's final year, he did an exchange program at the Eindhoven University of Technology (Eindhoven, the Netherlands). While in Eindhoven, he started working on this Master Thesis: "Tryptophan Metabolism Profiling for Psychiatric Diagnosis Support", for which he analyzed psychiatric patients' data, developed tryptophan metabolism models and created two novel metabolism-inspired classification algorithms. In 2016, he went back to Portugal to finish the project and obtained his Master's degree.

In 2018, he started working as a Ph.D. candidate in the Department of Vascular Medicine at the Amsterdam University Medical Center under the supervision of prof. dr. Erik Stroes and co-supervision dr. Evgeni Levin. The focus of his research was two-fold: to build predictive models/identify biomarkers for cardiovascular risk, and to develop novel algorithms that incorporate biological domain knowledge, as well as increase black-box models' transparency for improved feature discovery/selection. From 2019 to 2022, he worked part-time for the healthcare AI company Horaizon with projects ranging from real-time non-invasive glucose prediction to modeling the impact of different milk formulas on infants' phenotype.

Medical research has seen a stark increase in the amount of available data.

The sheer volume and complexity of measured variables challenge the use of traditional statistical methods and are beyond the ability of any human to comprehend. Solving this problem demands powerful models capable of capturing the variable interactions and how those are non-linearly related to the condition under study.

Machine learning offers flexible and powerful models making them promising candidates to solve this problem. However, there may be a prohibitively large space of possible solutions which motivates the main theme of this thesis: Can we use our current biology knowledge to constrain this space?