Verena Niederkofler, BSc

A Dynamic Network-Based Approach for Proteomic Biomarker Discovery and Pathway Analysis in Myocardial Injury

MASTER'S THESIS

to achieve the university degree of

Diplom-Ingenieurin

Master's degree programme: Biomedical Engineering

submitted to

Graz University of Technology

Supervisor

Univ.-Prof. Dipl.-Ing. Dr.techn. Christian Baumgartner

Institute of Health Care Engineering with
European Testing Center of Medical Devices

Robert E. Gerszten, MD, Chief of Cardiology
Beth Israel Deaconess Medical Center
Harvard Medical School - Teaching Hospital

Beth Israel Deaconess
Medical Center

HARVARD MEDICAL SCHOOL
TEACHING HOSPITAL

Graz, July 2018

EIDESSTATTLICHE ERKLÄRUNG

AFFIDAVIT

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Masterarbeit/Diplomarbeit/Dissertation identisch.

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis/diploma thesis/doctoral dissertation.

Datum / Date	Unterschrift / Signature

Die Technische Universität Graz übernimmt mit der Betreuung und Bewertung einer Masterarbeit keine Haftung für die erarbeiteten Ergebnisse: Eine positive Bewertung und Anerkennung (Approbation) einer Arbeit bescheinigt nicht notwendigerweise die vollständige Richtigkeit der Ergebnisse.

Acknowledgment

To begin with, I would like to thank the whole team at the Beth Israel Deaconess Medical Center for all the support and the expertise they provided during my stay in Boston. A special thanks goes to Robert E. Gerszten, MD and Univ.-Prof. Dipl.-Ing. Dr.techn. Christian Baumgartner, who made this opportunity possible to work at such a highly esteemed institution and to gain so much knowledge and experience for my future working life.

Also, I would like to thank my friends and especially my family for all the support through this time. They always support me in every situation and taught me that I can achieve all my goals if I believe in myself.

Zusammenfassung

Ein netzwerk-basierter Ansatz zur Identifikation von proteomischen Biomarkern und Pathway-Analyse bei Myokardinfarkt

Biomarker spielen eine große Rolle bei der Identifizierung von Krankheiten, und haben somit eine große Bedeutung in der Forschung. Dadurch, dass Herzkrankheiten zu den häufigsten Todesursachen weltweit zählen, wird der Entdeckung von Biomarkern ein großer Teil der Forschung gewidmet. Diese Masterarbeit beschäftigt sich mit einer aus zwei Schritten bestehenden Methode, für die Identifizierung von proteomischen Biomarkern bei Myokardinfarkt und deren Wechselwirkungen. Durch die Anwendung eines Subset-Selektions-Algorithmus werden die vielversprechendsten Proteine vorselektiert und für die nachfolgende Netzwerkanalyse verwendet. Diese netzwerk-basierte Methode identifiziert wichtige Interaktionen zwischen diesen Proteinen, wobei viel neue Information aus den Daten gewonnen wird. Die Schlüsselproteine und deren Interaktionen wurden anhand von Literatur und Datenbanken verifiziert. Bekannte Biomarker bei Myokardinfarkt konnten mit dieser Methode bestätigt werden, aber auch neue Proteine konnten als mögliche Biomarker-Kandidaten identifiziert werden, wobei weitere Studien zur biologischen Interpretation notwendig sind.

Schlüsselwörter: Proteomik, Biomarker, Pathway-Analyse, dynamische Netzwerke, Myokardinfarkt

Abstract

A dynamic network-based approach for proteomic biomarker discovery and pathway analysis in myocardial injury

Due to the significance of biomarkers for the identification of diseases, they play an important role in the research area. A big part of the research is dedicated to the identification of biomarkers for cardiovascular diseases being the most common cause of death worldwide. This thesis introduces a two-step method for the identification of proteomic biomarkers in myocardial infarction and pathway analysis. By using a subset selection method the most significant proteins, using data from a longitudinal cohort study and a case-control study, were preselected and included for the pathway analysis. Subsequently, a network-based approach to identify crucial relations between proteins was applied. The selected proteins and relations were analyzed by reviewing literature and different databases. Known biomarkers for myocardial infarction could be identified, confirming this method, but also some new biomarker candidates were identified, though, they need further investigation. Substantial amount of biological information could be added by the networks, showing many unexpected interactions between the selected proteins, but more research on the biological interpretation and verification is needed.

Key words: proteomics, biomarker, pathway analysis, dynamic networks, myocardial infarction

Contents

List of Abbreviations

AMI	acute myocardial infarction
BI	Biomarker Identifier
BL	baseline
BNP	B-type natriuretic peptide
CT	computer tomography
CV	coefficient of variation
DA	discriminatory ability
ECG	electrocardiogram
EMA	European Medicines Agency
FABP	fatty acid-binding protein
FDA	Food and Drug Administration
GCP	Good Clinical Practice
GWAS	genome-wide association study
HOCM	hypertrophic obstructive cardiomyopathy
HPLC	high performance liquid chromatography
IL	interleukin
IPA®	Ingenuity® Pathway Analysis
KEGG	Kyoto Encyclopedia of Genes and Genomes
MeSH	Medical Subject Heading
MI	myocardial infarction
MS	mass spectrometry
NPV	negative predictive value
pBI	paired Biomarker Identifier
PMI	planned myocardial infarction
PPI	protein-protein interactions

PPV	positive predictive value
RFU	relative fluorescent unit
SMI	spontaneous myocardial infarction
SNP	single nucleotide polymorphism
SOMAmer®	Slow-Off-rate Modified Aptamer
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
TP^2	true-positive rate
uBI	unpaired Biomarker Identifier
UniProt	Universal Protein database
WHO	World Health Organization

List of Figures

List of Tables

1. Assignment

The research dedicated to the biomarker discovery has become more important over time due to the significance of biomarkers to identify diseases. Biomarkers go back in time long before high-throughput technologies were invented. However, these technologies have made it possible to analyze hundreds of molecules to identify specific disease-related signatures with the objective to recognize a disease as early as possible and to even predict the outcome of the disease. Knowing all this, the most effective treatment can be selected for each patient individually.

Cardiovascular diseases belong to the most common causes for death worldwide. For this reason, a big part of the biomarker research is devoted to cardiovascular diseases along with cancer research. This thesis focuses on myocardial infarctions and a network-based pathway analysis of proteomic data. The data originates from a study about planned myocardial infarction from the Beth Israel Deaconess Medical Center, a Medical Center of the Harvard University. The model of a planned myocardial infarction represents an accepted approach for investigating a spontaneous myocardial infarction. The data consists of a derivation and validation cohort with proteomic data of peripheral blood from different measurement times, which allows the analysis through time. Additionally, blood was drawn from the coronary sinus of a small group of patients, allowing another comparison. Furthermore, proteomic data of peripheral blood from patients with a spontaneous myocardial infarction and a control group is included.

The goal was to use a two-step network-based method to analyze this data. First a subset selection method called *Biomarker Identifier* was applied to prioritize the proteins and to include the most significant for the network analysis in a second step. These networks, as a way to visualize the protein interactions through time, were used to analyze the relations between proteins in addition to verify the proteins selected by the *Biomarker Identifier* method.

Thus, the results are eventually discussed and interpreted with the support of the relevant databases and literature reviews.

2. Biomarker Introduction

The importance of biomarkers has increased drastically over the last few decades as they are used in many aspects of our lives due to their significance in the medical domain and the ongoing discovery process has yet more to reveal. This chapter describes the relevance of biomarkers and outlines the considerable steps of the discovery process in the research area.

2.1. Preamble

In the ancient Egyptian medicine physicians tried to identify the cause of a disease with a fraction of the knowledge and understanding of modern medicine. The Edwin Smith Papyrus (1600 B.C.) is a clear reflection of many medical problems, which illustrates their most effective treatments and, therefore, a remarkable testament. For instance, an abnormal pulse rhythm was referred to heart diseases. [1]

Accordingly, biomarkers are a phenomenon that has been around for a very long time. Also, Sushruta, the "father of surgery", described in the seventh century B.C., that ants were attracted by the urine of patients with diabetes. Urine was brought into focus for many centuries because of its simple availability. Hippocrates (460 - 355 B.C.) is often mentioned as the founder of uroscopy and he stated the hypothesis of urine as body fluid, which was filtered through the kidneys. With the observation of bubbles on the surface of urine he drew the conclusion that it might be a sign of a long-term kidney disease. This appearance is now known to happen due to proteinuria. The Greek writer Theophilus Protospatharius observed cloudiness while heating the urine from patients, which showed signs of a kidney disease. Likewise, Paracelsus discovered the same cloudiness using vinegar in the sixteenth century. The heat causes the precipitation of proteins, similar to acid. Bringing it full circle, Richard Bright (1789-1859) discovered the correlation between kidney disease through the autopsy findings of abnormal kidney and proteinuria. Many other parameters were discovered and used as biomarkers with the passing of time, like Roentgen's discovery of the x-ray in 1895, producing an image which can be used as a diagnostic tool. [2]

Thus, the biomarker research has steadily been increasing. Nowadays, a PubMed search of the term *biomarker* lists more than 745 000 publications and over 19 000 publications including the term *biomarker* in the title. [i] Over the course of time, the same search was performed and listed more than 765 000 publications and over 20 000 with the term *biomarker* in the title.[ii]

2.2. Definition

The term "biomarker" is composed of the two words "biological marker". Although biomarkers have been known for millennia, this term was first used in 1977 by Karpetsky, Humphrey and Levy in the *Journal of the National Cancer Institute* [2]. In 1989, the term biomarker was first introduced as a Medical Subject Heading (MeSH) [3]. In 1998, the National Institutes of Health Biomarkers Definitions Working Group defined a biomarker as

> "*a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention.*" [4]

The World Health Organization (WHO) similarly defined a biomarker as

> "*any substance, structure, or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease.*" [4]

There are not only these but also several other precise definitions, which fortunately overlap considerably. In general, a biomarker refers to a broad range of subcategories of medical signs. These are objective indications of health or disease states observed from outside, measured on a sample, an obtained recording from a patient or an imaging test. Hence, a biomarker can be pulse, blood pressure, urine, an electrocardiogram (ECG), a computed tomography (CT) scan and much more. A biomarker measured on a biosample indicates an abnormal condition correlating to a specific disease because of its characteristics. This biosample can be a tissue test or molecule substances like a gene, metabolite, protein or more complex chemistries. [3, 4, 5]

Considering all these definitions and examples, the term biomarker compromises a large field of measurements, which have the common goal to reveal and understand the interaction between the body and a disease.

[i] *www.ncbi.nlm.nih.gov/pubmed*, as of 31.01.2018
[ii] *www.ncbi.nlm.nih.gov/pubmed*, as of 31.05.2018

2.3. Characteristics and Categorization

Diseases cause specific changes in tissue and biological fluids indicating to those specific diseases. The analysis of the genetic, proteomic and metabolomic composition allows the detection of the diseases as they occur or identification of individuals with predispositions. The patterns of an overexpression of a substance correlate with important characteristics. There are many steps from the pathogenesis to the clinical manifestation of a disease and biomarkers can be used at any level, from the molecular to the organ level. Therefore, different categories of biomarkers are available. [5, 6]

To identify those patterns and to predict the clinical outcome a biomarker should have some significant properties to be useful. Some of the most important characteristics are shortly stated in the following paragraph, nonetheless, it should be mentioned that the desired properties and their significance are dependent on the intended use of the biomarker and the disease.

One of the most crucial characteristics are the sensitivity and the specificity. These two parameters should be as near to 100% as possible for a biomarker to get clinically accepted and to differentiate health or disease states. The sensitivity is the percentage of positive test results, given that the disease is truly present. The specificity is defined as the ability of a marker or test to exclude the disease, given that the disease is truly absent. Consequently, the clinical value of a new biomarker is only given if it is accurate, reproducibly obtainable and objectively quantifiable, easy to interpret and certainly acceptable for the patient. Furthermore, the biomarker should be specific to the disease and not due to biological perturbations or environmental conditions. [3, 4, 5]

Additionally, the positive predictive value (PPV) and the negative predictive value (NPV) can be a relevant feature for a biomarker. The PPV is defined as the percentage of individuals with a positive test result and a present disease. The NPV, on the other hand, is defined as the percentage of individuals with a negative test result and the absence of the disease. Moreover, false negatives and false positives should be prevented, where the false negatives indicate the individuals having the disease but a negative test result, and the false positives, by contrast, indicate the individuals without the disease but with a positive test result. False positive and false negative conclusions can, therefore, be a risk for the clinical efficacy. A small summary of the ideal methods and important parameters is shown in figure 2.1. [5, 7]

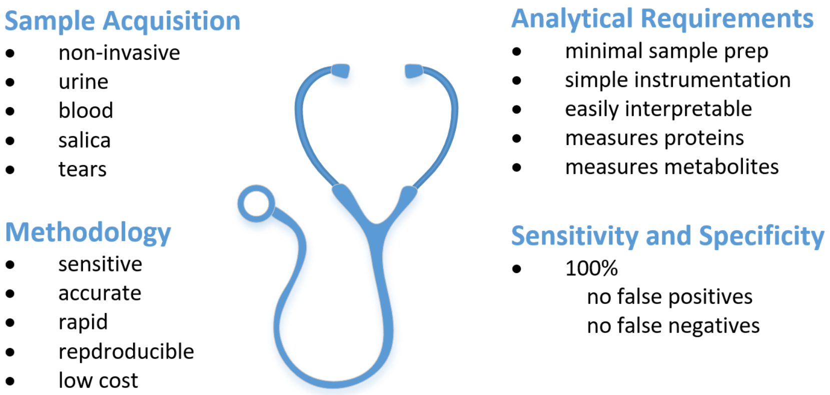


Figure 2.1.: Summary of ideal methods and significant parameters for disease diagnosis. [5]

As previously mentioned, biomarkers can be used at any level to identify characteristics of a disease. Accordingly, different types of biomarkers can be distinguished. There are three major types: antecedent or risk biomarkers, diagnostic and prognostic biomarker. The antecedent or risk biomarkers are used to identify the risk of a patient to develop a disease. Diagnostic biomarkers recognize an overt or early disease state and classify the severity of a disease, whereas prognostic biomarkers are used to predict a future disease course, the recurrence and response to therapy, but also to monitor the efficacy of a therapy. A small overview of these three types is shown in table 2.1. More types and subcategories can be found depending on differentiations in literature. [3, 8]

Types of Biomarkers	
Antecedent/Risk biomarkers	Identification of the risk of developing a disease
Diagnostic biomarkers	Detection of overt disease Classification of severity of disease
Prognostic biomarkers	Prediction of future disease course Prediction of recurrence and therapy response Monitoring efficacy of therapy

Table 2.1.: Short overview of the major categorizations of biomarkers. [3, 8]

It is also possible that one biomarker may be used for both diagnostic and prognostic purposes. However, a biomarker belongs generally to only one category. [9]

2.4. Biomarker Discovery

Biomarkers have always played a major role in the medical practice since its inception. Most of the historical biomarkers were externally visible indicators of physiological processes, like fever, rash or swelling. However, the rise of high-throughput omic technologies over the last decades has made it possible to search unique molecular biomarkers which are correlated to diseases. The objective is to improve the early detection, determine the prognosis, monitor the therapy response or select the most efficient treatments. Even though the biomarker discovery and characterization process provides opportunities to improve all of the above-mentioned situations, it also results in large datasets demanding a variety of sophisticated methods of analysis to identify a biomarker associated with the disease of interest and to be clinically approved. This discovery-driven biomarker process and identification are comparable to looking for a needle in a haystack. [8, 10]

2.4.1. Current State

Biomarker research is rising, however, a great number of literature indicates that a lot of presently used molecular biomarkers are inadequate to replace a clinical test. Even though many have been suggested, they lack the required specificity or sensitivity for the early detection of cancer or other diseases. For an early detection biomarkers must possess such a high specificity to differentiate between cancer, which is clinically significant or just related and benign. At the same time the sensitivity has to be high enough to ideally detect small tumor masses before any clinical symptoms. Consequently, many proteomic and metabolomic biomarkers are, up to the present, only used to detect an advanced stage cancer with a low survival rate. [8, 5]

The current state of biomarker research may be illustrated with an example of bladder cancer biomarkers. Despite the huge effort which was made to find a biomarker, there have not been any results of an acceptable test or rather biomarker which was specific and sensitive enough to substitute cystoscopy. Bladder cancer has a recurring nature and requires monitoring for three to six months, which makes the disease expensive to treat. There are only some biomarkers which have been approved by the *European Medicines Agency* (EMA) or the *U.S. Food and Drug Administration* (FDA) to monitor the recurrence of bladder cancer. Some of these biomarkers are the nuclear matrix protein (NMP22) with a sensitivity of 56% and hyaluronic acid/hyaluronidase measurements with a sensitivity of 92%. However, none of them are sensitive enough to be suited for population screening. [5]

Nevertheless, the present state also shows that the biomarker research constantly proposes many novel promising biomarker candidates for a variety of diseases. Since the focus of this thesis lies on biomarkers in cardiovascular diseases, a more detailed explanation and some examples are stated in section 2.5. First, the general approach of the biomarker discovery process will be explained in the following section.

2.4.2. Discovery Process

Biomarkers for the early detection and identification of a recurrent disease, but also biomarkers to predict the outcome of a disease and the response to therapy require many cooperating fields of expertise. This collaboration of interdisciplinary knowledge usually includes biologists, clinicians, analytical- and biochemists, but also bioinformaticians. All of these fields and sometimes more, cover the biomarker discovery process from study design to biomarker validation. A schematic representation of these significant steps in the biomarker discovery process is depicted in figure 2.2. [11, 12]

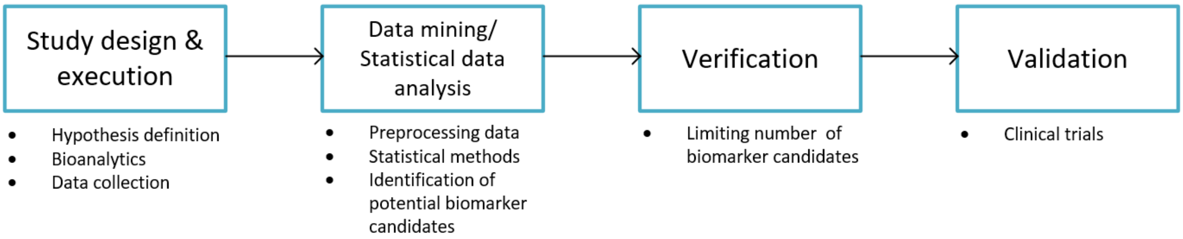


Figure 2.2.: Demonstrative representation of the major stages in the biomarker discovery process. [11]

The advent of high-throughput technologies has made it possible to easily gather information of thousands of individual molecules and generate large amounts of data, which require many steps in the biomarker discovery process. Although there has been an “explosion” in the biomarker research and the number of novel biomarker candidates is rising, there is a low estimated rate of a successful clinical translation and application of biomarkers due to issues along the numerous stages in the biomarker discovery process. However, the ultimate goal is to establish these clinically accessible tests with clinical utility. [13, 14, 15]

An illustration of a general possibility of the procedure of the biomarker discovery using high performance liquid chromatography (HPLC) and mass spectrometry (MS) is shown in figure 2.3. A speciem is taken from diseased and healthy subjects and after analyzing and comparing the results the discriminating factors can be determined. [5]

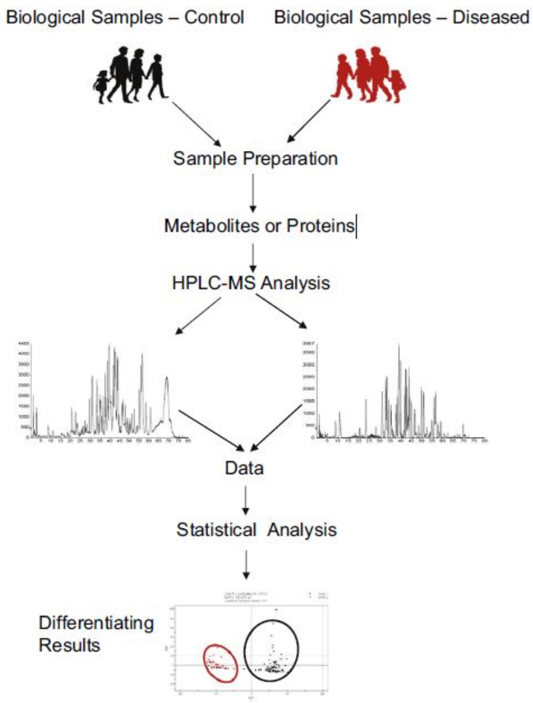


Figure 2.3.: Generic procedure of a biomarker discovery process using HPLC/MS and statistical data analysis. [5]

Study Design and Execution

The whole biomarker discovery process requires a careful consideration of various aspects, hence, the study design is fundamental. Some important steps to consider are the disease of interest, sample type, number of patients, class of molecule (metabolites, proteins, nucleotides) and defining a hypothesis. These are some crucial steps in the discovery process because they affect the integrity of the results. For instance, the number of patients should be high enough to be statistically significant, and the proper sample collection, preparation and storage can have a great impact on the specificity and sensitivity, two parameters, which ideally should be 100%. [5]

Clearly defining the research question and the fundamental use of a biomarker can be quite challenging in this early stage of the discovery process. Advantageously, in the era of high-throughput omic technologies it is possible to interrogate thousands of molecules without *a priori* hypotheses, but can be defined *post hoc*. In general there are two types of development for biomarker identification, the discovery-based or data-driven approach and the hypothesis-based or knowledge-driven approach. [8, 15]

Data-driven approaches make use of big datasets, which can let researchers face terabytes of raw data. These datasets are used to facilitate the clarification of the underlying structure and to get first insights of possible new biomarkers. These large datasets have to be reduced to a manageable size to obtain and understand the complex biological information. To contrast with the data-driven approach the knowledge-driven approach integrates data into available knowledge bases. The understanding of biological processes and diseases is facilitated this way. A variety of different study designs can be used from cohort and case-control studies to more complex designs like serial sampling or cross-over designs. Serial sampling studies investigate kinetic characteristics and patterns of analytes by tracking changes in levels over time. The most frequently used study design is the retrospective case-control study, where patients with a particular medical condition (cases) are compared with individuals without this condition but with other phenotypic and particular characteristics (controls). Longitudinal cohort studies observe patients over a period of time and allow them to serve as their own biological control. This reduces the interindividual variability examined in other cohort studies. The study design is then followed by the study execution. Also, the sample collection, preparation and storage as well as the data collection is included here. [8, 16, 11, 17]

Data Mining

Significant effort is being made towards the discovery of novel biomarkers. With the large toolbox of profiling technologies researchers need to handle immense data volume. To reduce its dimension and complexity specialized software systems and algorithms are necessary. In addition, the preprocessing and data mining should take less effort and time than the data acquisition. [17]

Data preprocessing is the first essential step to ensure completeness and consistency, because the throughput of various technologies precludes the analysis of the huge amout of data at once. Some of the various methods of preprocessing is data transformation (e.g. logarithmic scaling), data normalization (e.g. z-transformation) or data sampling and outlier detection. Outlier detection is a significant task in order to avoid distortion of the statistical analysis. [11, 18]

Data reduction is the next important step to reduce the complexity to a manageable size, make the dataset easier to understand but also to eliminate noise. These include for example, trend analysis, clustering and feature selection. Trend analysis is used to reveal patterns which are statistically relevant and can therefore also be applied in population data. Clustering refers to any approach to group the data according to certain characteristics or criteria, such as specific cancer subtypes, similar expression patterns or to identify similar pathways. [8]

The goal of feature selection can be either to improve model predictive accuracy or, on the other hand, to determine relevance for hypothesis generation. Features, or numerous independent variables, characterize a data instance. For example a patient, who potentially has cancer, and tumor markers as characterizing features. There might also be a response variable, for example if the tumor is benign or malignant. Feature selection can be categorized into supervised, semi-supervised or unsupervised. Supervised feature selection means that all the response variables are known for the data instances of the dataset. If only some data instances have known response variables, the feature selection is semi-supervised and unsupervised if none of the response variables are known. [19]

Feature selection methods have a broad classification into filters, wrappers and more advanced methods like an embedded algorithm, which is commonly accepted in literature. [11]

A summery of some data mining methods for independent and dependent samples is shown in table 2.2. The method used in this thesis is the Biomarker Identifier (BI) which will be explained in detail in section 3.2. A detailed review of the other methods is beyond the scope of this thesis and can be reviewed in other literature.

Up to this point, these methods belonged to the data-driven approaches, however, the knowledge-driven approaches make use of existing knowledge to facilitate the understanding of underlying structures and processes. Methods which belong here are, for example, protein-protein interactions (PPI) or pathway analysis. For the PPI method existing knowledge of protein interactions with experimental datasets are used to understand the underlying biological mechanisms since interactions between proteins play a key role in most cellular processes. PPI studies focus on experimental identification of protein interactions and their characterization, but also on the application of computational techniques to predict interactions between proteins and domains based on experimental data. Otherwise, pathway analysis represents another form of knowledge, which can be included into biomarker discovery studies. The objective is again to develop an understanding of mechanisms, which are disease-related at a molecular level. Pathway analysis happens in three steps. First, for example, a set of genes is chosen, usually by a data-driven approach, then a hypothesis with a biologically relevant question is formulated and eventually, a statistical test is chosen to answer the question. [8]

Independent samples	
Unpaired Biomarker Identifier (uBI)	- univariate filter method - statistical evaluation score by combining a discriminance measure with a biological effect term - quick and effective ranking of features - uBI scores permit prioritization of features - uBI scores closely related to pBI scores - appropriate for two class problems only
Support vector machine-recursive feature elimination (SVM-REF)	- embedded selection method - SVM-REF uses optimized weights of SVM classifier to rank features - appropriate for two class problems only
Random forest models (RFM)	- embedded selection method - RFM uses bagging and random subspace methods to construct a collection of decision trees aiming at identifying a complete set of significant features - appropriate for multiple class problems
Dependent samples	
Paired null hypothesis testing (Paired t-test*, Wilcoxon signed-rank test△)	- univariate filter method - P value serves as evaluation measure for the discriminatory ability of variables - is an accepted statistical measure - appropriate for two class problems only - P value is sample size dependent - two dependent samples
Paired Biomarker Identifier (pBI)	- univariate filter method - pBI uses a statistical evaluation score by combining a discriminance measure with a biological effect term - pBI scores permit prioritization of features - pBI scores closely related to uBI scores - appropriate for two class problems only

Table 2.2.: Summary of selected commonly used data mining methods for independent and dependent samples for the identification of biomarker candidates. [11] * normal distributed data, △ non-normal distributed data

Verification and Validation

Upon completion of the data mining and statistical analysis dozens to hundreds of potential biomarker candidates are selected. Hence, it is necessary to limit the number of the biomarker candidates and to confirm the potential of each putative biomarker candidate. Even though the number of publications declaring the identification of new biomarker candidates is rising, the number of approved biomarkers is actually decreasing. A reason for this decrease might be the high number of candidates identified as false-positive in the identification phase or difficulties in the verification and validation phase. The lack of a method to verify hundreds of potential biomarkers is described as the " bottleneck" of biomarker development. [20, 21]

To reduce the number of candidates appropriate statistical analysis and the comparison with available databases are required. This knowledge and data can be obtained from public repositories such as Kyoto Encyclopedia of Genes and Genomes (KEGG), Universal Protein database (UniProt), Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) and literature. After the verification phase only a manageable list of biomarker candidates proceed to the validation phase. This phase is a multifaceted process requiring the collaboration of various clinical centers. Independent sample cohorts and validation studies performed in a retrospective and prospective manner would be ideal to ensure statistical relevance. The biomarker candidates have to be validated in clinical trials with continuous monitoring of their performance and significance for a period of time before the final approval. [11, 21]

2.5. Biomarkers in Cardiovascular Disease

Up to this point biomarkers were discussed in general, yet, from now on the focus will lie on proteins as biomarkers in cardiovascular diseases.

The world's leading cause of death and hospitalization are cardiovascular diseases including stroke and ischemic heart diseases. Hence, a lot of time is spent on biomarker research to improve early detection and select the most efficient treatments to achieve the best clinical outcome possible. Proteomic technologies provide opportunities to measure in a large-scale manner to discover new biomarkers. The human body is expected to have over 20 000 proteins, based on protein-coding genes. Plasma is often used for proteomic biomarker studies, since a broad range of proteins from all tissues can be found in this complex mixture. The smallest abundant proteins in plasma are troponins and cytokines, which are in the range of a few $\frac{pg}{mL}$ and less, in contrast to the most abundant protein albumin which lies in the range of $35 - 45$ $\frac{mg}{mL}$. The following example should illustrate this broad range and the challenge on finding low-abundant proteins. Looking for a troponin molecule with an abundance of 5 $\frac{pg}{mL}$ among albumin molecules at 45 $\frac{mg}{mL}$, can be compared to looking for one human in the entire human population. The first used technologies to detect proteins were electrophoresis and mass spectrometry, on which many studies on cardiovascular diseases are based on. Due to the lack and challenges of detecting low-abundant proteins the number of studies based on more advanced technologies is increasing. One aptamer-based technology, which is used to generate the data analyzed in this thesis, will be explained shortly in section 3.1.1. [22, 23]

2.5.1. Biomarkers in Myocardial Infarction

Myocardial infarction (MI) is caused by prolonged ischemia resulting in myocardial cell death, which occurs within hours after the onset of myocardial ischemia, with the highest risk of fatality. Thus, these first few hours are crucial for the early diagnosis and quick identification of an acute myocardial infarction (AMI) to initiate a timely treatment. [24]

Two established and used biomarkers for AMI are creatine kinase (CK) and cardiac troponin. CK was already indicated as a biomarker for cardiac diseases in 1979. It is primarily found in the cardiac and skeletal muscle, whereas the isoenzym CK-MB is in the cardiac muscle. Cardiac troponin is highly specific to cardiac tissue and an accepted diagnostic marker for a MI. It is released when irreversible myocardial damage has occured and reaches peak values at 12 hours and can have high levels for days. [25, 24]

Another example is the heart-type fatty acid-binding protein (H-FABP). H-FABP has been identified as early biomarker for ischemia since it is released only 30 minutes after a MI. B-type natriuretic peptide (BNP) is a well-known biomarker of biochemical stress and is released if cardiomyocytes are under tension. Copeptin levels increase within minutes after an AMI and is, therefore, another early biomarker for myocardial necrosis. Interleukin (IL)-6 is a marker of early atherosclerosis and activates inflammatory cells as a reaction to ischemia. [25, 24, 26]

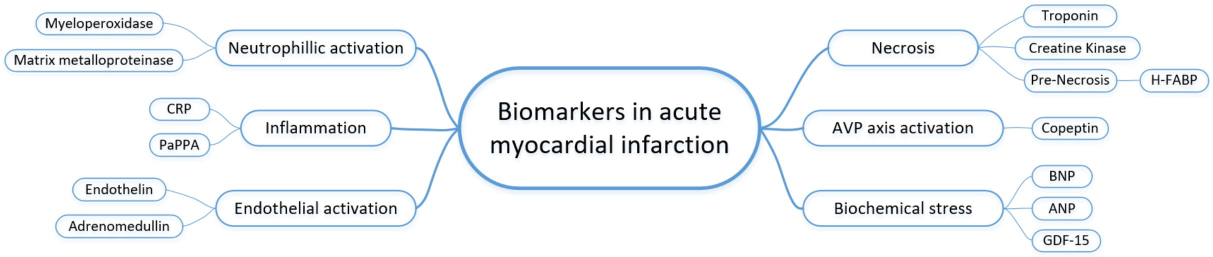


Figure 2.4.: Illustration of some biomarkers with their associated pathophysiological process associated with acute myocardial infarction (AMI). [25] CRP - C-reactive protein, PaPPA - pregnancy associated plasma protein, H-FABP - heart-type fatty acid-binding protein, BNP - B-type natriuretic peptide, ANP - atrial natriuretic peptide, GDF-15 - growth differentiation factor-15

A graphic representation of various associated biomarkers with AMI, including some of the before mentioned, is shown in figure 2.4. To effectively use the initial hours of an AMI to reduce the mortality, research of cardiac biomarkers is significant for a rapid diagnosis.

3. Methods

The method used to analyze the data to ensure a meaningful interpretation is based on a two-step preprocessing procedure. First, a subset selection method was applied to the preprocessed data for prioritization purpose. In the next step, the networks were generated with the before prioritized proteins to show significant interactions. The method for the subset selection was implemented in the language `R`, based on the `R`-package `BiomarkeR` from [27] and is publicly available. The implementation of the networks is based on the approach from [28] and [29]. All of this `R`-code was adapted and extended to fit for the relative RFU proteomic data.

3.1. Data Background

The data used in this thesis originates from [22], a study about planned myocardial infarction (PMI) from the Beth Israel Deaconess Medical Center, a Medical Center of the Harvard University. The patients of this study suffer from hypertrophic obstructive cardiomyopathy (HOCM). This hypertrophy of the myocardium obstructs the outflow of blood from the left ventricle of the heart. During a PMI, patients undergo an alcohol septum ablation as a treatment of HOCM. The clinical model of PMI represents a spontaneous myocardial infarction (SMI), where blood samples can be drawn at different times to study the changes in molecule levels.

The data includes 1129 different proteins and consists of a derivation and a validation cohort, each containing data from 15 patients. Peripheral blood was drawn before the alcohol septum ablation, which represents the baseline (BL) value, 10 minutes (10min), 1 hour (1h) and 24 hours (24h) after the ablation. As a longitudinal cohort study each patient could serve as their own biological control to reduce the interindividual variability. With the placement of a catheter into the coronary sinus blood was drawn at baseline, 10 minutes and 1 hour from an additional 6 patients, allowing the comparison of two different measurement locations and serving as a second validation cohort. Furthermore, the data includes patients with a SMI and a control group, which underwent a cardiac catheterization without having a myocardial ischemia, with each 23 individuals. To ensure the principles of the Good Clinical Practice (GCP), informed consent was obtained from all participants and [22] offers more details about the inclusion criteria. [22]

3.1.1. Data Generation

The data was generated by a high-throughput proteomic aptamer-based technology called SOMAscan® assay by SomaLogic. It is a powerful tool for discovering biomarkers measuring over 1.300 proteins in an exceptional dynamic range across eight orders of magnitude in abundance, from femtomolar (fM) to micromolar (μM). The here used SOMAmer® (Slow-Off-rate Modified Aptamer) reagents benefit from aptamer technology. Aptamers are oligonucleotides of short single-stranded DNA or RNA molecules able to bind protein targets with high specificity because of a specific sequence. However, identifying aptamers is difficult because the more limited chemical diversity of nucleic acid than that of proteins. To overcome this, chemically modified nucleotides are added expanding the chemical diversity and, therefore, enhancing the specificity. [30, 31]

A SOMAmer assay quantitatively measures the proteins in a biological sample by transforming them into a specific SOMAmer-based DNA signal. This signal is quantified by fluorescence, leading to relative fluorescent units (RFU), which is directly proportional to the amount of protein in the initial biological sample. [30, 31]

3.1.2. Data Preprocessing

Before the data could be used, it had to be prepared for further analysis. This first step of the data preprocessing usually includes handling of missing values and outlier detection. Because of the high-quality data, there were no missing values which had to be filled in. To ensure statistical significant results outlier detection was performed by detecting and replacing outliers with the upper or lower limit.

The whole dataset included more information, which was not all needed for these analyses. Therefore, only the required information was extracted and used. This included the sample IDs for each patient, the time points at which the measurement was done and all the raw RFU values of the proteins. It should be noted, that the RFU values are relative values and, therefore, also the results and, especially, the interactions between the proteins should be interpreted as interactions in terms of expression and concentration changes, changes in function and changes of conformity. All these parameters can play an important role in the interpretation of the results.

The preprocessing of the data also included the separation in a *non-targeted* and *targeted* approach. For the *non-tageted* approach all 1129 proteins were considered for the subset selection, which will be explained in section 3.2. This subset of proteins from the derivation cohort was then validated in the validation cohorts. For reasons of simplification, the validation cohort with the measurement of the peripheral blood will from now on be referred to as validation cohort 1 and the cohort with the measurement in the coronary sinus will be referred to as validation cohort 2. For the *targeted* approach, on the other hand, a subset of proteins was selected from the SMI data, and these proteins were validated based on the PMI data.

3.2. Subset Selection Method

The subset selection method, called Biomarker Identifier, used in this thesis originates from Baumgartner et al. [16]. The objective is to use this score to prioritize the protein data and select a subset of proteins for further analysis. There are two different variations of this model for paired and unpaired data, which will be explained in the following sections.

3.2.1. Paired Biomarker Identifier

The paired Biomarker Identifier (pBI) is defined for paired data and incorporates two main features of the diagnostic test, the discriminatory ability (DA) and the coefficient of variation (CV). These parameters in combination with the scaling factor λ and the effect term $\sqrt{|\Delta_{change}|/|CV|}$ result in equation 3.1. [16]

$$\begin{aligned} pBI &= \lambda \cdot DA^{*} \cdot \sqrt{\frac{|\Delta_{change}|}{|CV|}} \cdot sign(\Delta_{change}) \\ \Delta_{change} &= \begin{cases} \Delta & \text{if } \Delta \geqslant 1 \\ -\frac{1}{\Delta} & \text{else} \end{cases} \end{aligned} \tag{3.1}$$

λ represents a scaling factor and is 100 by default. The DA is defined as percentage change in one direction versus baseline. For instance, if 50% of the protein levels are increasing, the DA is calculated as 0.5, likewise, if 75% of protein levels are increasing the DA is 0.75. This parameter is usually defined in the range of $[0.5, 1]$ and is rescaled to be in the range of $[0, 1]$ which is represented by DA^{*}, weighted by the biological effect term $\sqrt{|\Delta_{change}|/|CV|}$. If $CV > 1$, CV is set to be 1 by default, to only consider data distributions with a smaller variance. The mean percentage change of one point in time versus BL is represented by Δ_{change} and the direction of change by the *sign* function. [16]

This pBI score was used for the PMI time-series data and each patient served as their own biological control. In order to calculate the pBI score the data was split up to each one measurement time t_i ($i = 1, 2, 3$; representing 10min, 1h, 24h) and t_0 for the BL value as reference. A positive value indicates an increasing level, while, a negative value indicates a decreasing protein level compared to BL.

3.2.2. Unpaired Biomarker Identifier

The unpaired Biomarker Identifier (uBI) is defined for unpaired data and incorporates in addition to the CV another two significant features from the diagnostic test, the sensitivity and the specificity. These two parameters are combined in the true-positive rate (TP^2) and the resulting equation is stated in 3.2. [16]

$$\begin{gathered} uBI = \lambda \cdot TP^{2*} \cdot \sqrt{|\Delta_{change}| \; \frac{CV_{ref}}{CV}} \cdot sign(\Delta_{change}) \\ \Delta_{change} = \begin{cases} \Delta & \text{if } \Delta \geqslant 1 \\ -\frac{1}{\Delta} & \text{else} \end{cases} \quad \text{with } \Delta = \frac{\overline{x}}{\overline{x}_{ref}} \end{gathered} \tag{3.2}$$

λ is again a scaling factor and 100 by default. The TP^2 is defined as the product of sensitivity and specificity and represents an objective measure of discrimination between two independent groups. The interpretation of this parameter is the following: TP^2 is 0.25 if sensitivity and specificity is 0.5, which does not indicate any valuable discrimination, comparable with DA of 0.5. Note that TP^2 is set 0, indicating no discriminatory value, if either sensitivity or specificity is < 0.5. TP^2 is defined in the range of $[0.25, 1]$ and is rescaled to $[0, 1]$, represented by the parameter TP^{2*}. Likewise to the pBI equation 3.1, the biological effect term is represented by $\sqrt{|\Delta_{change}| \cdot (CV_{ref}/CV)}$. Δ_{change} represents the relative percentage change from one point in time to the reference and the division of the CV of the two independent groups represents changes in the variance of the data. The *sign* function displays as well the direction of change and $\overline{x}$ the mean value in either of the groups. [16]

The uBI score was used for the SMI data, consisting of two independent groups. The control group served as the reference for the group of patients, who suffered from a SMI. A positive value, again, indicates an increasing and a negative value a decreasing level.

3.2.3. Interpretation and Visualization

The rescaling of the parameters DA and TP^2 makes it possible to compare the absolute pBI and the uBI and results in generalizing this subset selection method for both paired and unpaired data. [16]

The use of this subset selection method was the first step to prioritize and to only include the most significant proteins. Therefore, a threshold τ had to be defined to exclude pBI scores, which are not high enough and, thus, proteins which are not suitable for further analyses. The threshold could be chosen as the 0.5, 0.75 or 0.9 quantile $\tau = \{q50,\ q75,\ q90\}$, whereas, a lower threshold includes a higher number of proteins for further analyses, but a higher threshold results in a more significant selection of proteins.

First, the pBI scores were calculated for all proteins for every time point t_i versus BL, then the threshold τ was computed of the absolute values and only proteins with $|pBI| > \tau$ were included. Likewise, the uBI score was calculated for the two groups, the threshold τ was computed and only proteins with $|uBI| > \tau$ were considered. τ was selected as $\tau = q75$, to include a moderate number but also to ensure a reasonably significant set of proteins.

To show the changes of the pBI scores for the PMI data throughout time, barplot diagrams were generated for each time point t_i. In order to select the highest scores, the absolute values were ranked and a selected subset of these scores were considered in the diagrams. For reasons of visualization the subset of selected proteins included 30 proteins, which were then also considered for the further network analysis. In that way, the diagrams illustrate the most significant proteins based on the pBI score at each time point t_i. Also, a barplot diagram was constructed for the highest uBI scores, representing the most significant proteins in the SMI data. These scores indicate the ability of a protein to be a good biomarker, in other words, the higher the score in either direction, the higher the chance of a protein to be a good candidate as a novel biomarker.

3.3. Dynamic Network Generation

The subset selection was the first step to prioritize the proteins and to include a subset for the further network-based analysis. Networks are a useful tool to analyze the interactions between proteins and discover relationships of different pathways. Different genes interacting in a network to regulate biological processes and, therefore, also regulate specific diseases represent the basis for a network-based approach [32].

Since one of the objectives of this thesis is to analyze the proteomic data in a network-based approach through time, only the PMI time-series data was considered for the second step in the analysis. For the *non-targeted* approach the top 30 proteins from the derivation cohort were used for the network analysis in all cohorts. For the *targeted* approach the uBI score was first computed for the SMI data and only this subset of those 30 proteins with the highest uBI score was considered in the analysis for the PMI data. Consequently, the pBI scores were only computed for this small subset of proteins and, also, the networks include solely those selected proteins.

The following sections focus on how the networks were constructed and visualized, which can then be seen in chapter 4.

3.3.1. Dynamic Network Construction

After selecting the most significant proteins for the network construction, the next step was to determine and visualize the interactions between them.

In order to do so, a network G with a set of vertices V and edges E, connecting the vertices, was constructed. A network G is, therefore, defined as $G = (V, E)$, where the vertices V or nodes represent the proteins and the edges E the interactions between the proteins. At first, the logarithmic ratios R of all proteins P, using the RFU values, were calculated, as $r_{ij} = |log_2(\frac{p_i}{p_j})|$ with $i > j$, $r \in R, p \in P$ and p must be positive $(p \geqslant 0)$. The logarithm was used to induce the symmetry of the protein ratios and their reciprocals. [27]

Next, the pBI score was calculated for all ratios r according to equation 3.1 and described in section 3.2.1, which will be referred to as pBI* score. To determine if an interaction between two proteins took place, a threshold had to be defined. Once again, the threshold τ could be defined as $\tau = \{q50,\ q75,\ q90\}$. If $|pBI^*| > \tau$ an interaction between these two protein was assumed and an edge was constructed to connect the two nodes corresponding to these two proteins. As a threshold τ the 0.75 and 0.9 quantile was used to guarantee significant results. This process was performed for every time point t_i, resulting in i networks for each cohort allowing the analysis of the changes of the interactions around each protein between the cohorts and at each time point t_i individually.

3.3.2. Dynamic Network Visualization

So far, networks were created representing the interactions around each protein for each time point t_i. These networks were combined in a final step of the analysis, which allows the investigation of the interactions through time. The edges can be weighted by two different methods, which will be introduced in this section.

The first method to combine the networks is by *discretely* weighting the edges. Hereto, a binary adjacency matrix A is created for each time step t_i with a 0 if there has been no connection and a 1 if there has been a connection between two proteins p_i. An example of such an adjacency matrix A is illustrated below. [28]

$$A = \begin{array}{c} \\ p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{array} \begin{array}{c} \begin{array}{ccccc} p_1 & p_2 & p_3 & p_4 & p_5 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix} \end{array} \tag{3.3}$$

All the adjacency matrices for each time point t_i are summed up over time, therefore, an entry can have values from zero to i, the number of time points and, hence, the number of networks combined. This final matrix is normalized to make the weight distribution independent from the number of networks combined. To visualize this weight distribution in the combined network, the edges are displayed in different thicknesses. The thicker the line in the network between two proteins the more often an interaction has occurred between them. If the edge is displayed as a thin line, fewer interactions have taken place. In other words, a network with discrete weighted edges represents the frequency of interactions between this set of proteins. [28]

The second method is by continuously weighting the edges by combining the networks over time. This method is based on the pBI* scores, which are computed for each time point t_i. In order to combine the networks over time, the scores are normalized to ensure comparability. Equivalent to the discrete weighted edges, an adjacency matrix is created for every time point. The only difference is, that the entries can now take values in the range from [0,1]. All these adjacency matrices are, again, summed over time and normalized to make the weight distribution independent from the number of networks combined. The higher the pBI* scores have been for two proteins for every time point t_i, the thicker the edge will be displayed in the network. Consequently, a thin line illustrates a weaker interaction between two proteins through time. In contrast to the discrete weighting, the continuous weighting represents the strength of the interaction between two proteins throughout all the time points. For reasons of visualization, the edges of all the combined network graphs were weighted by a polynomial function of fourth degree. Consequently, significant relations are illustrated in a more evident way. [28]

3.4. Protein Verification and Pathway Analysis

To achieve the best possible and significant results different cohorts were used to analyze the most relevant proteins. As a first verification step only the most significant proteins from the derivation cohort were selected and validated in a validation cohort. Due to the extensive dataset, this selected subset of proteins could be validated in two different validation cohorts, leading to more significant results.

After reducing the data to the most promising proteins and validating them in different cohorts, these findings had to be verified in one more step. The verification of these results is a meaningful task which should be done carefully to confirm the findings with already known biomarkers and newly published biomarker candidates. Not only literature can be used for this crucial task, but there are many publicly available databases like UniProt, STRING and KEGG, just to mention a few.

To ensure various sources of information, not only literature was reviewed to confirm already known proteins as biomarkers, but also the different databases were used to gather additional information and to investigate pathways. Additionally, to guarantee a wide-reaching and meaningful verification the software `Ingenuity® Pathway Analysis` (IPA®), a powerful analysis software, was used for the verification of proteins and proteomic pathways.

4. Results

This chapter provides an overview of the results which are based on the method described in chapter 3. These results include different barplot diagrams showing the distribution of the highest pBI scores, the networks at different time points and the combined networks for the *non-targeted* and the *targeted* approach.

4.1. SMI vs. Controls

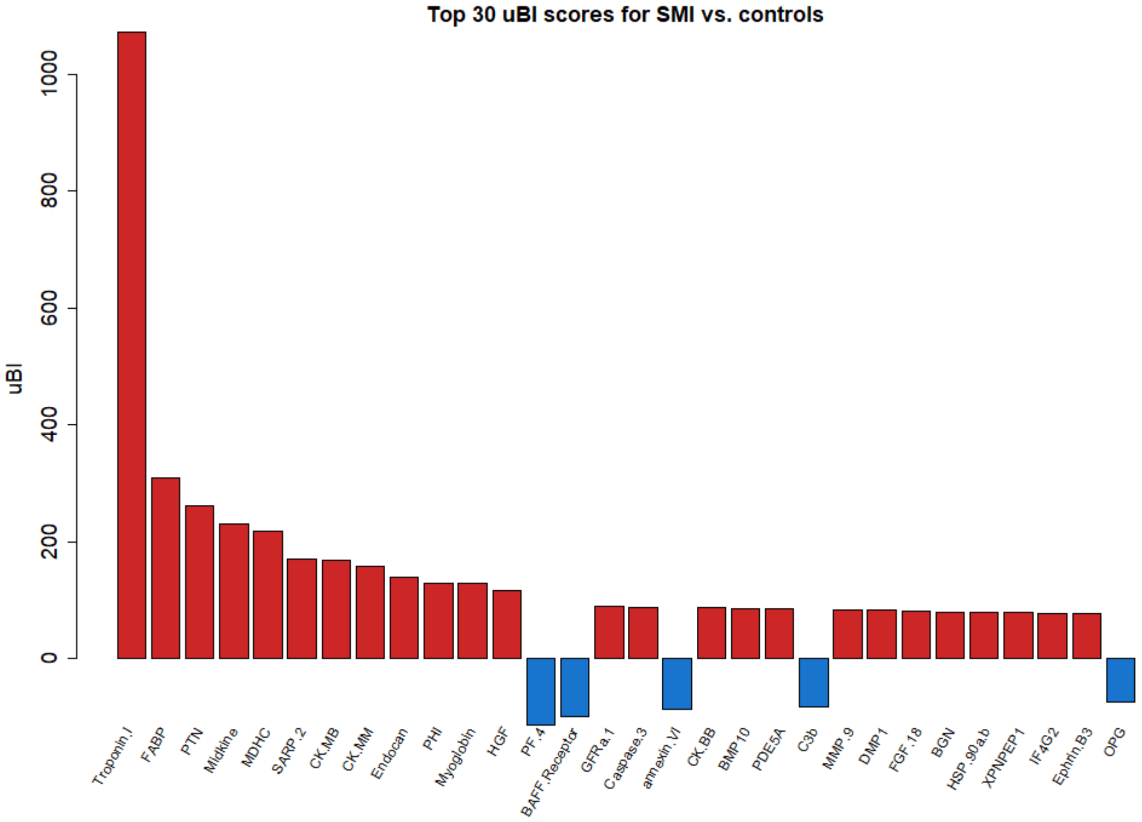


Figure 4.1.: Barplot diagram of the top 30 ranked proteins (uBI scores) of the SMI data vs. controls at baseline.

Figure 4.1 shows a barplot diagram of the uBI scores of the unpaired SMI data. The uBI scores were calculated with equation 3.2 of the group of patients who suffered a SMI and a control group. Shown are the top 30 proteins with the highest uBI score in either direction, where a positive score represents an increasing level and a negative score represents a decreasing level, which is illustrated by the red and blue bars.

The diagram shows the uBI scores at baseline, where it should be noted that all patients with a SMI came into the hospital at different times within the first few hours. The proteins shown in figure 4.1 were used for the *targeted* approach. These results can be seen in section 4.3.

4.2. Non-targeted Approach

For the *non-targeted* approach all 1129 proteins were used for the first verification step. These proteins were prioritized and the most significant ones selected for further analysis. The following subsections will provide an overview of the pBI scores and the networks.

Also, as a general summary a table was created, which show the top 20 percentage changes, calculated from the RFU values in each cohort at each measurement time t_i versus baseline. This table outlines the most significant changes for each cohort individually including all proteins and can be seen in the supplement A.3.

4.2.1. Visualization of pBI Scores

The following barplot diagrams show the 30 highest pBI scores of the derivation and the validation cohorts 1 and 2 at a measurement time of 10 minutes. Like mentioned in section 3.2.3 the barplot diagrams are a way to visualize the distribution of the highest pBI scores of the selected subset of proteins, where the absolute values of the scores were ranked beforehand. The red bars represent an increasing level and the blue bars a decreasing level of the protein values. The barplot diagrams for all cohorts for the additional measurement times can be seen in A.1, whereat there are barplot diagrams at 1 hour and 24 hours for the derivation and validation cohort 1 and one more barplot diagram at 1 hour for the validation cohort 2.

These barplot diagrams allow the comparison of the most significant proteins in different cohorts as a first verification step.

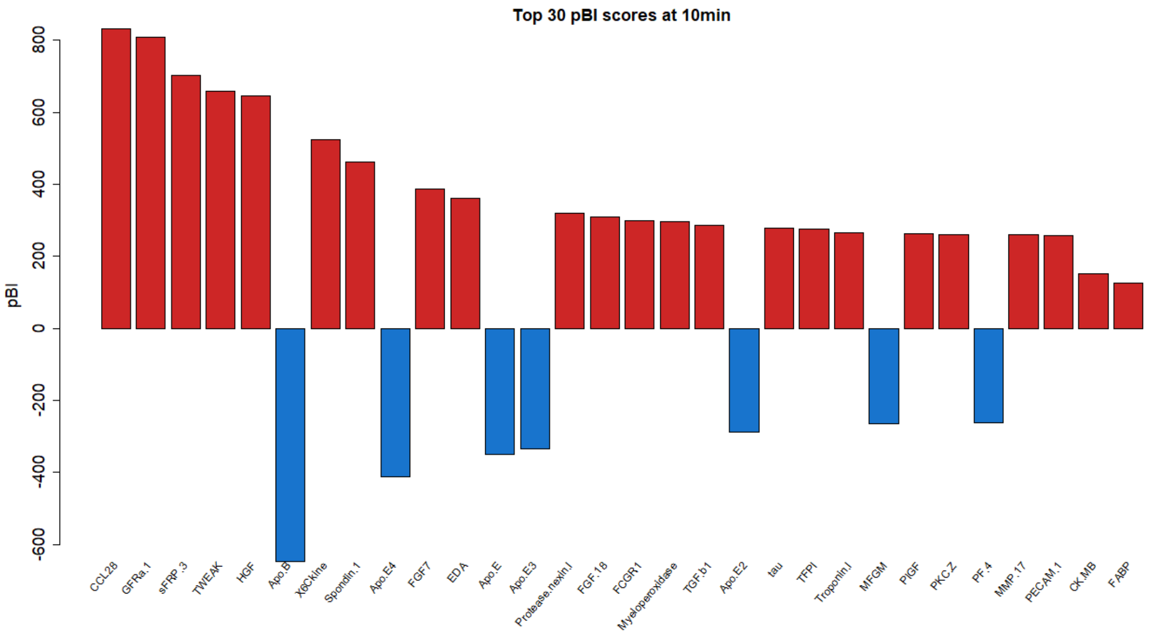


Figure 4.2.: Barplot diagram of the top 30 ranked proteins (pBI scores) of the derivation cohort at 10 minutes using the *non-targeted* approach.

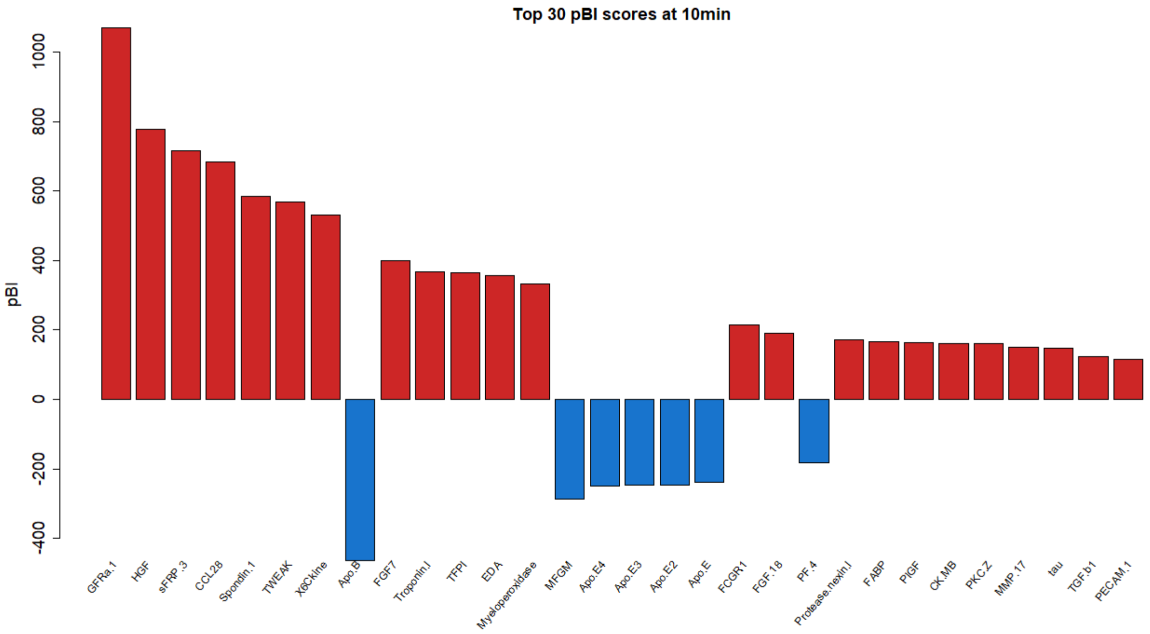


Figure 4.3.: Barplot diagram of the top 30 ranked proteins (pBI scores) of the validation cohort 1 at 10 minutes using the *non-targeted* approach.

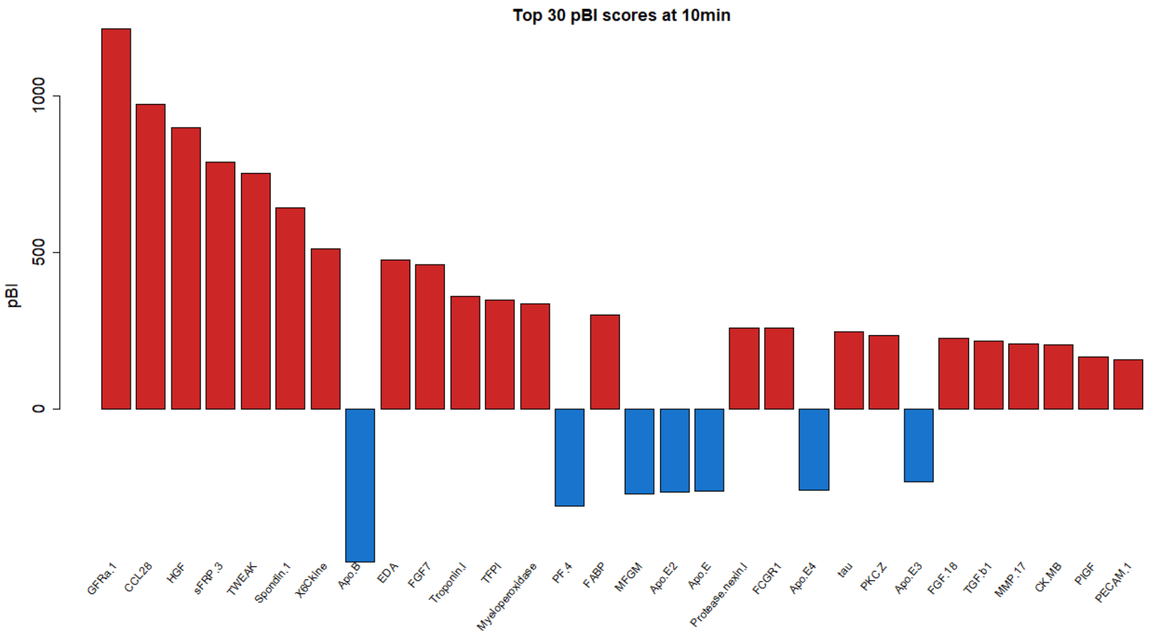
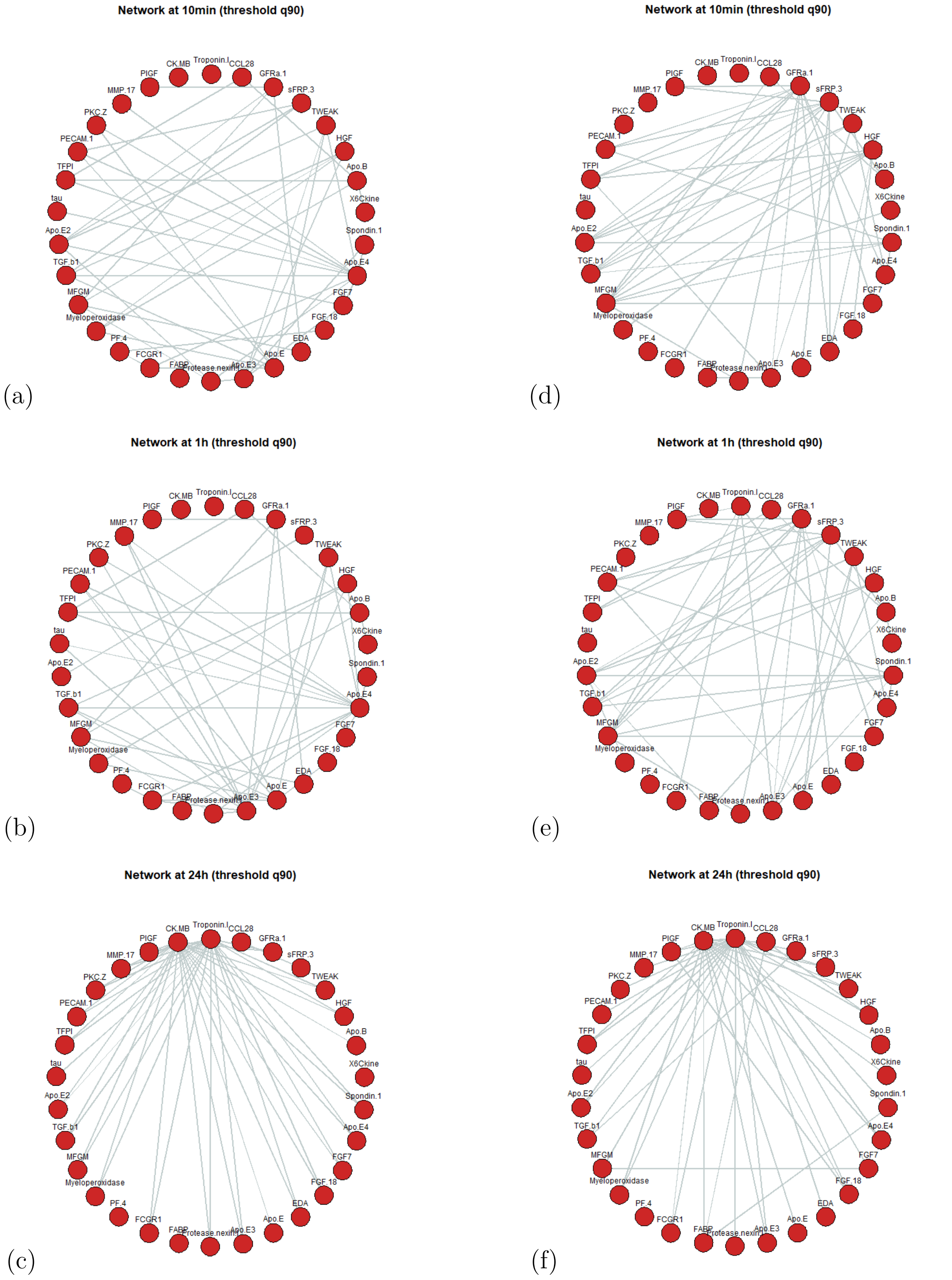


Figure 4.4.: Barplot diagram of the top 30 ranked proteins (pBI scores) of the validation cohort 2 at 10 minutes using the *non-targeted* approach.

4.2.2. Dynamic Network Graphs

The dynamic network graphs illustrate the interactions around each protein of the selected subset at each measurement time t_i. The subfigures in figure 4.5 show all the single networks at all time points t_i for the derivation cohort and the validation cohort 1 allowing a direct comparison of the interactions. Additionally, the interactions can be analyzed through time for one cohort or in comparison with another cohort. The two networks at 10 minutes and 1 hour for the validation cohort 2 are shown in figure 4.6, which gives the opportunity for a second verification with another cohort at two different measurement times.

All these networks with a different threshold of $\tau = q75$ are shown in A.6. A lower threshold includes more interactions for a broader analysis.



Network at 10min (threshold q90)

Network at 1h (threshold q90)

Network at 24h (threshold q90)

Network at 10min (threshold q90)

Network at 1h (threshold q90)

Network at 24h (threshold q90)

Figure 4.5.: Network representation of the selected subset of proteins of the derivation cohort (a-c) and validation cohort 1 (d-f) using the *non-targeted* approach at different measurement times with a threshold of $\tau = q90$. (a) derivation at 10min, (b) derivation at 1h, (c) derivation at 24h, (d) validation 1 at 10min, (e) validation 1 at 1h, (f) validation 1 at 24h

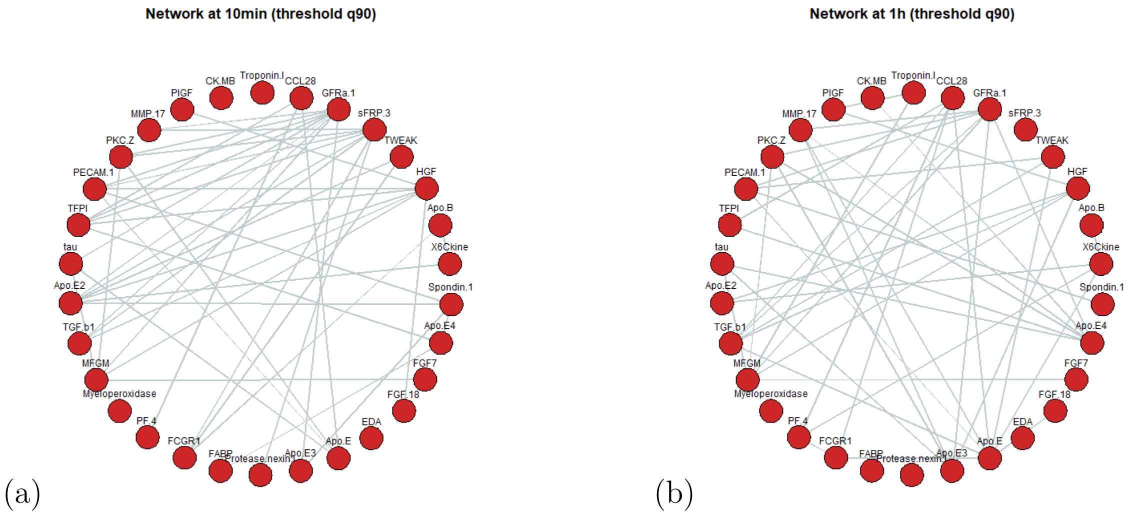


Figure 4.6.: Network representation of the selected subset of proteins of the validation cohort 2 using the *non-targeted* approach at different measurement times with a threshold of $\tau = q90$. (a) validation 2 at 10min, (b) validation 2 at 1h

4.2.3. Combined Dynamic Network Graphs

To visualize the interactions around each protein through time in one single network, the combined networks were created, explained in section 3.3.2. The two different approaches, discrete and continuous weighted edges, are displayed in figure 4.7.

The continuous weighted edges are a method to show the strength of the interaction through time, and the discrete weighted edges, on the other hand, are a representation of the frequency of the interactions between the proteins through time. It should be noted, that for the derivation cohort and validation cohort 1 three networks were combined (10min, 1h, 24h) and for the validation cohort 2 two networks were combined (10min, 1h).

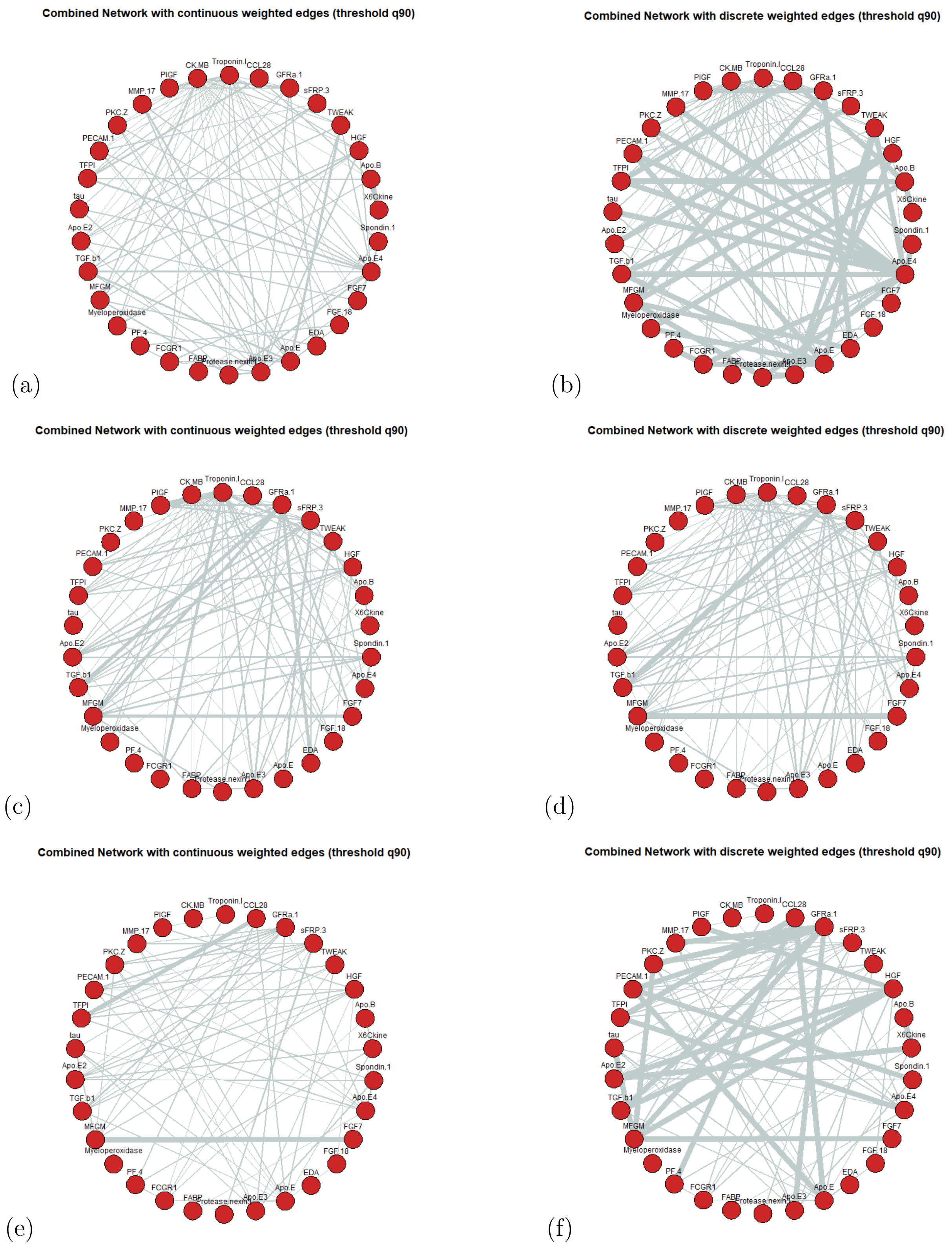

Combined Network with continuous weighted edges (threshold q90)
Combined Network with discrete weighted edges (threshold q90)
CK.MB
Troponin.I
CCL28
GFRa.1
sFRP.3
TWEAK
HGF
Apo.B
X6Ckine
Spondin.1
Apo.E4
FGF7
FGF.18
EDA
Apo.E
Apo.E3
Protease.nexin
FABP
FCGR1
PF.4
Myeloperoxidase
MFGM
TGF.b1
Apo.E2
tau
TFPI
PECAM.1
PKC.Z
MMP.17
PIGF
(a)
(b)
Combined Network with continuous weighted edges (threshold q90)
Combined Network with discrete weighted edges (threshold q90)
(c)
(d)
Combined Network with continuous weighted edges (threshold q90)
Combined Network with discrete weighted edges (threshold q90)
(e)
(f)

Figure 4.7.: Combined network representation of the selected subset of proteins with continuous (a,c,e) and discrete (b,d,f) weighted edges using the *non-targeted* approach with a threshold of $\tau = q90$. derivation continuous (a) and discrete (b) weighted edges, validation 1 continuous (c) and discrete (d) weighted edges, validation 2 continuous (e) and discrete (f) weighted edges

4.3. Targeted Approach

In the *targeted* approach a selected subset of proteins was used for the analysis of the PMI data. This subset was extracted from the SMI data by calculating the uBI scores and using the top 30 proteins with the highest score for the further analysis. Therefore, the pBI score was only calculated for this subset and also the networks were created in the same way like in section 4.2 for the *non-targeted* approach.

As an overview another table with the top 20 percentage changes of these 30 proteins was created and can be seen in table A.2 in the supplement.

4.3.1. Visualization of pBI Scores

As in section 4.2.1 the following barplot diagrams show the distribution of the pBI scores of the selected proteins. Again, the red bars represents an increasing and the blue bars a decreasing level. The barplot diagrams show the pBI scores at 10 minutes for all cohorts. The additional diagrams are illustrated in the supplement A.2, whereat there are barplot diagrams at 1 hour and 24 hours for the derivation and validation cohort 1 and one diagram for the validation cohort 2 at 1 hour.

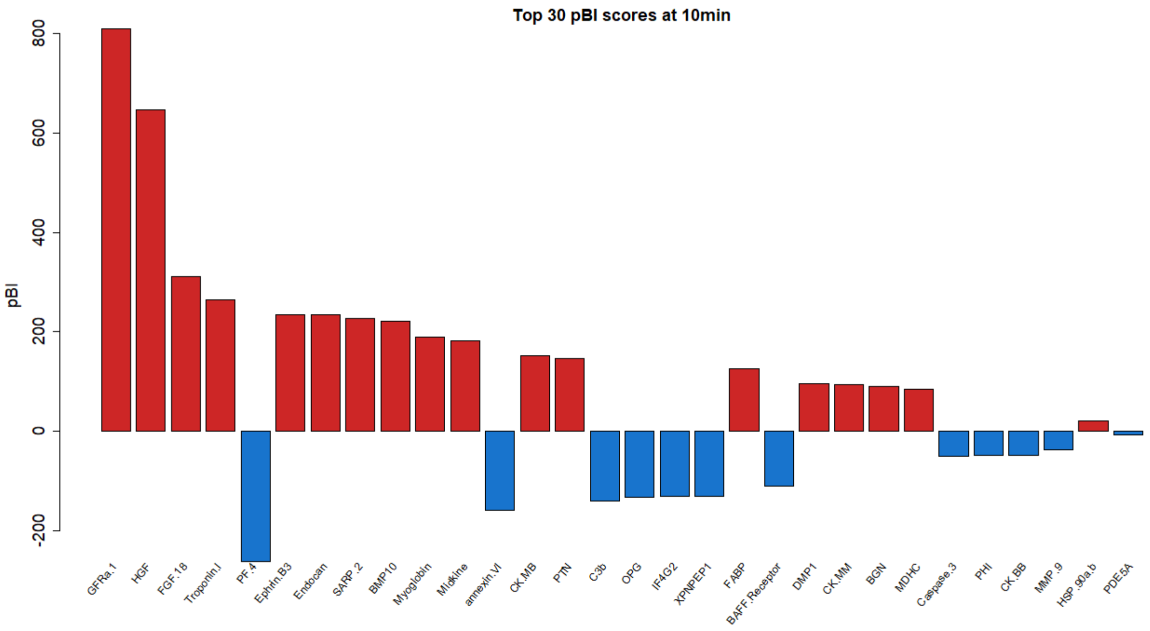


Figure 4.8.: Barplot diagram of the top 30 ranked proteins (pBI scores) of the derivation cohort at 10 minutes using the *targeted* approach.

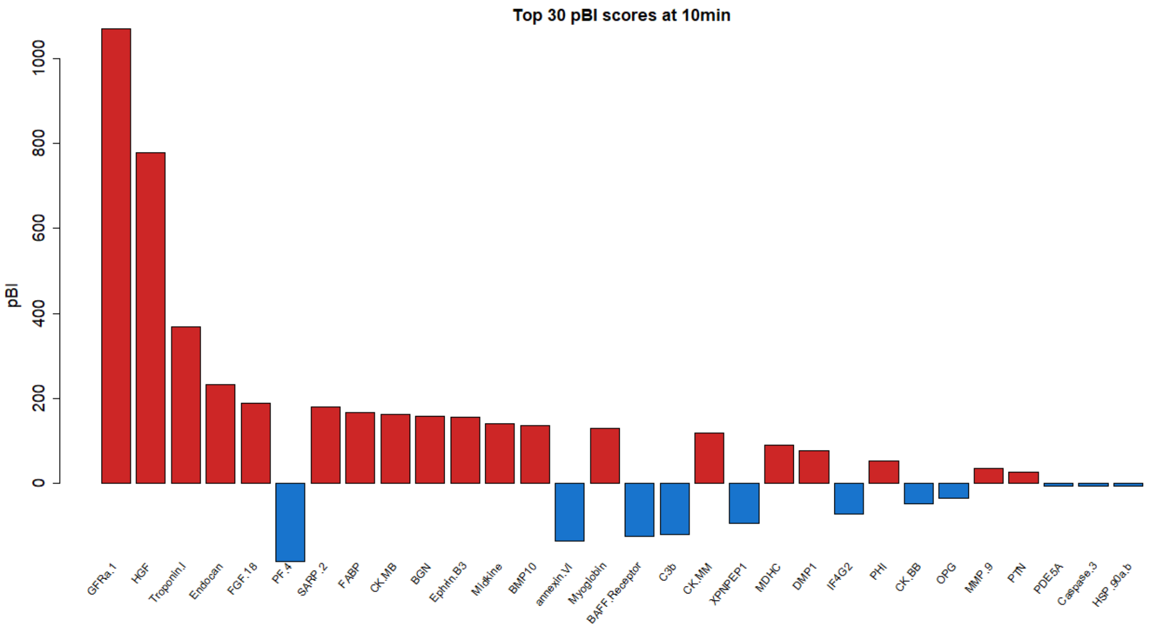


Figure 4.9.: Barplot diagram of the top 30 ranked proteins (pBI scores) of the validation cohort 1 at 10 minutes using the *targeted* approach.

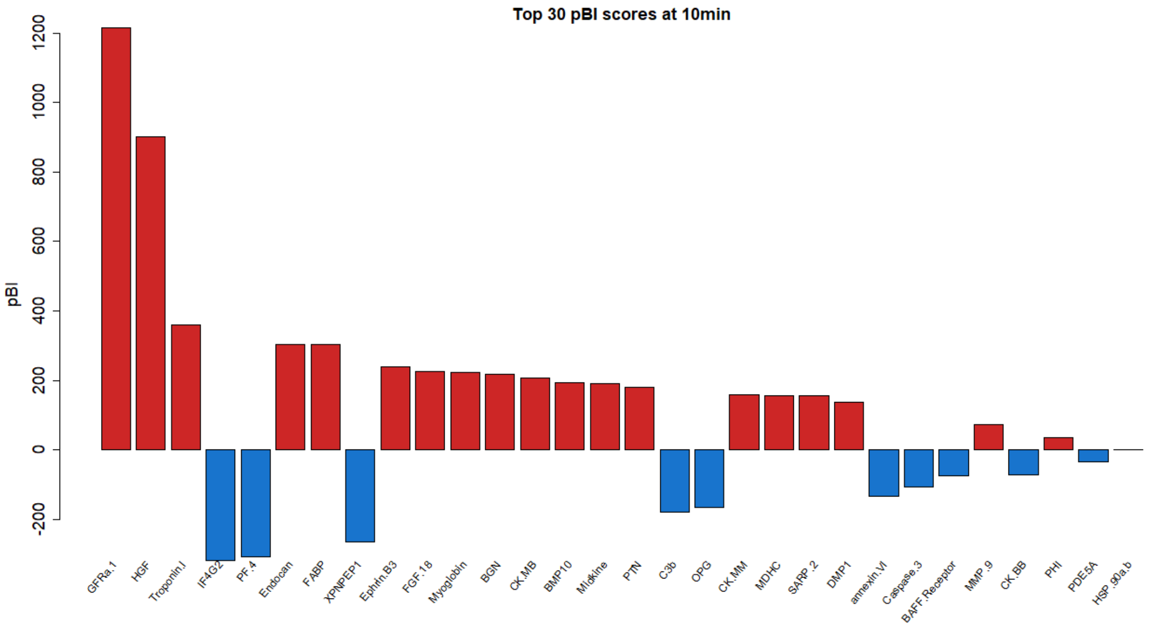
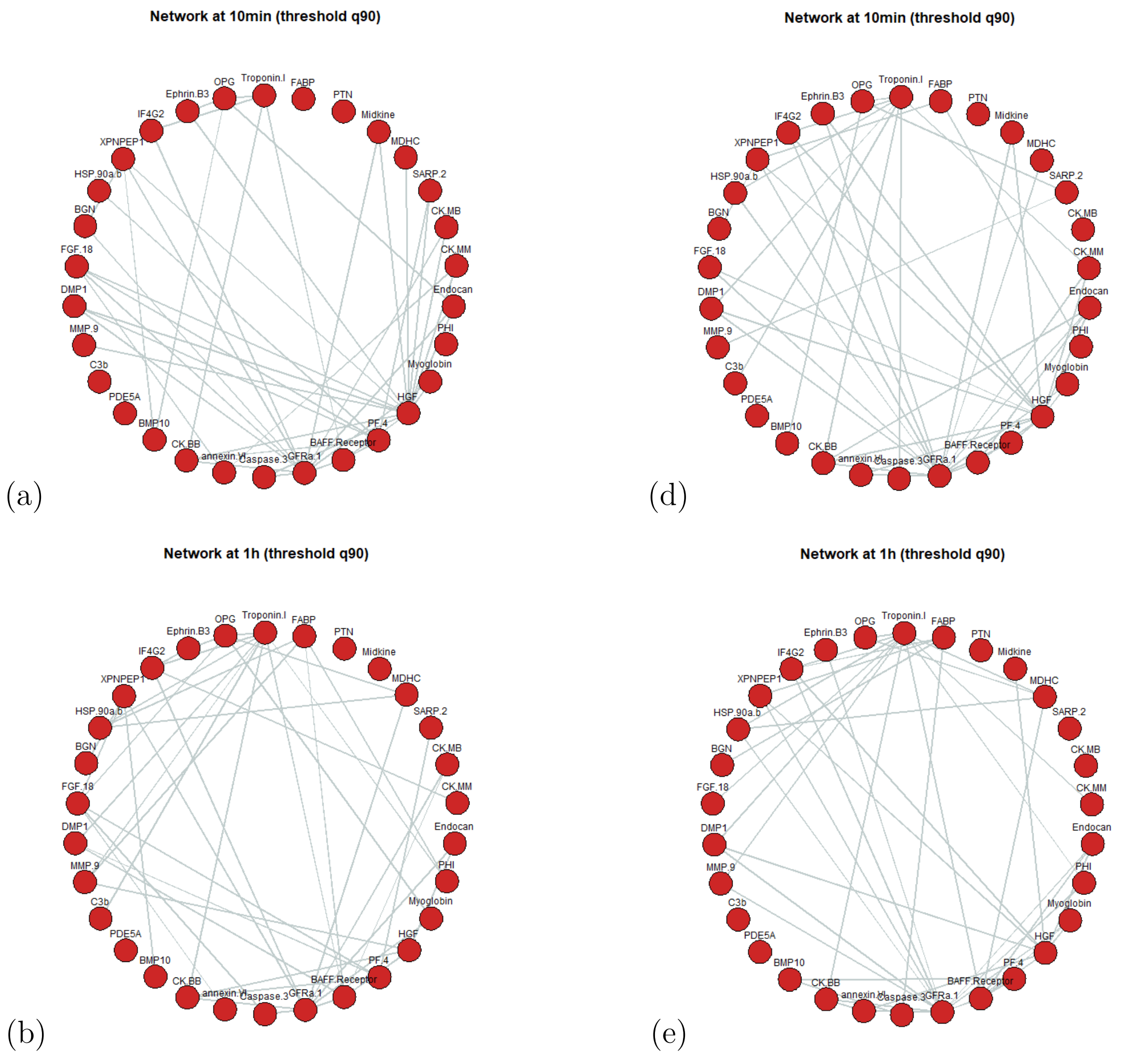


Figure 4.10.: Barplot diagram of the top 30 ranked proteins (pBI scores) of the validation cohort 2 at 10 minutes using the *targeted* approach.

4.3.2. Dynamic Network Graphs

To analyze the interactions around each protein through time for each cohort of the selected proteins, dynamic network graphs were inferred and are shown in figure 4.10 for the derivation and validation cohort 1 at each time point t_i. Figure 4.11 illustrates the networks for the validation cohort 2 at 10 minutes and 1 hour.

As well as for the *targeted* approach, additional networks with a threshold of $\tau = q75$ are shown in the supplement A.2.


Network at 10min (threshold q90)

Network at 10min (threshold q90)

Network at 1h (threshold q90)

Network at 1h (threshold q90)

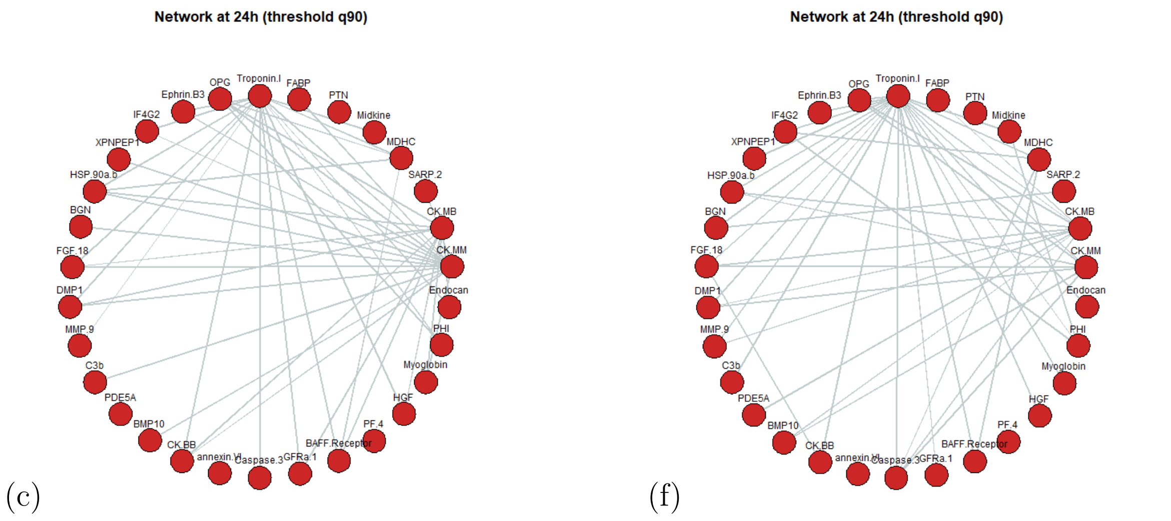


Figure 4.10.: Network representation of the selected subset of proteins of the derivation cohort (a-c) and validation cohort 1 (d-f) using the *targeted* approach at different measurement times with a threshold of $\tau = q90$. (a) derivation at 10min, (b) derivation at 1h, (c) derivation at 24h, (d) validation 1 at 10min, (e) validation 1 at 1h, (f) validation 1 at 24h

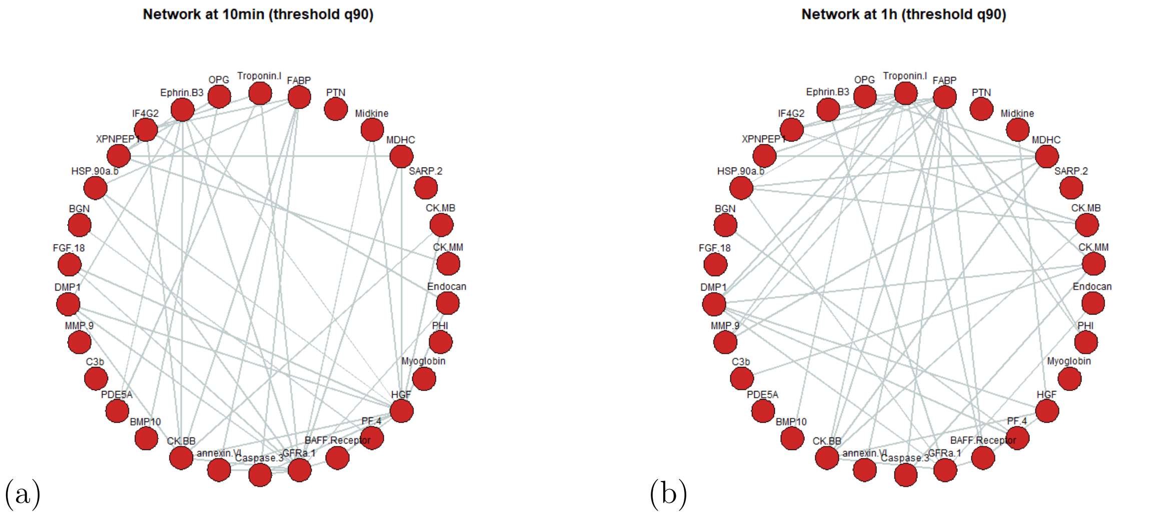
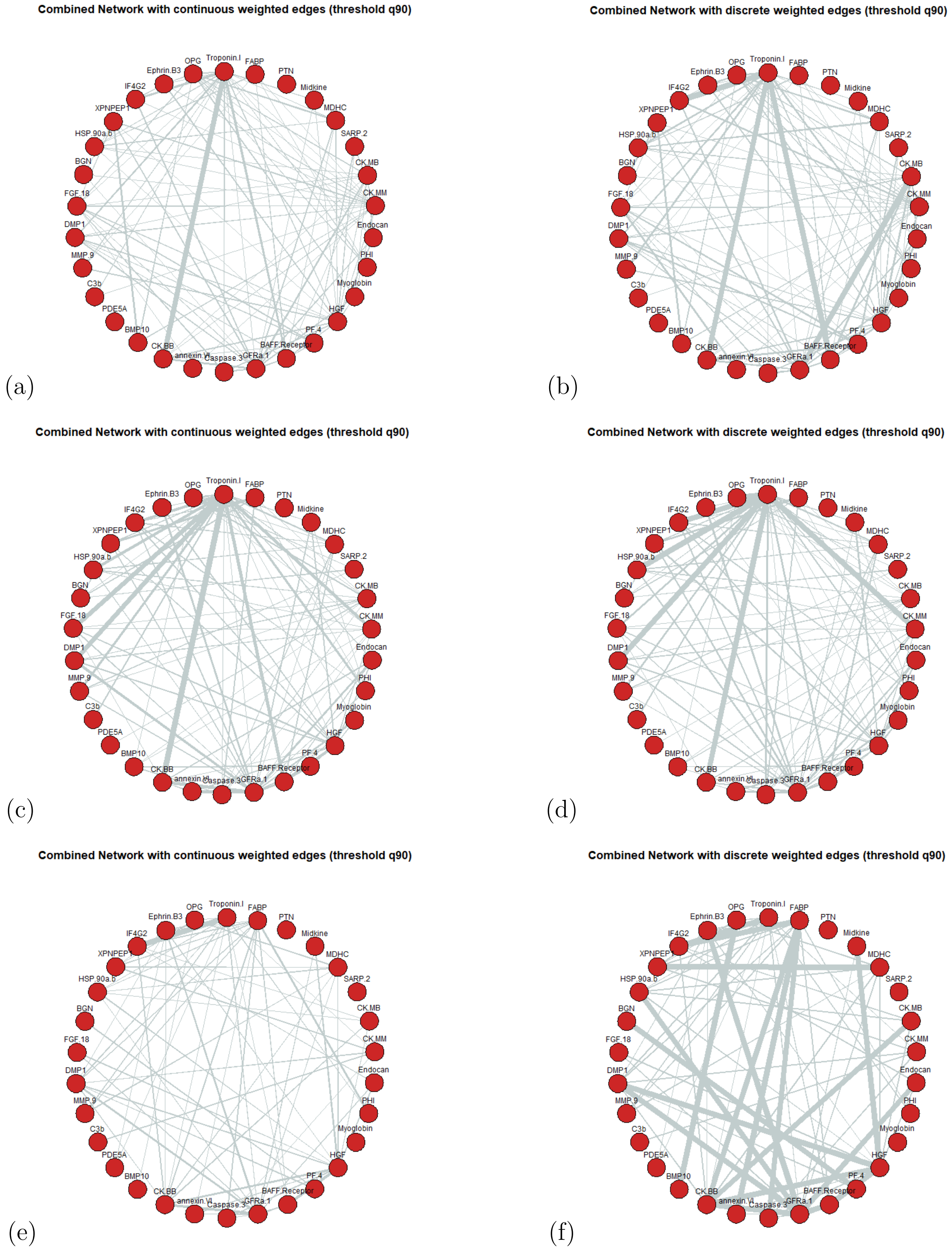


Figure 4.11.: Network representation of the selected subset of proteins of the validation cohort 2 using the *targeted* approach at different measurement times with a threshold of $\tau = q90$. (a) validation 2 at 10min, (b) validation 2 at 1h

4.3.3. Combined Dynamic Network Graphs

Like in section 4.2.3, the combined networks of the *targeted* approach represent the interactions through time in a single network graph. Figure 4.12 shows the graphs for all cohorts with the continuous and discrete weighted edges.


Combined Network with continuous weighted edges (threshold q90)

Combined Network with discrete weighted edges (threshold q90)

Combined Network with continuous weighted edges (threshold q90)

Combined Network with discrete weighted edges (threshold q90)

Combined Network with continuous weighted edges (threshold q90)

Combined Network with discrete weighted edges (threshold q90)

Figure 4.12.: Combined network representation of the selected subset of proteins with continuous (a,c,e) and discrete (b,d,f) weighted edges using the *targeted* approach with a threshold of $\tau = q90$. derivation continuous (a) and discrete (b) weighted edges, validation 1 continuous (c) and discrete (d) weighted edges, validation 2 continuous (e) and discrete (f) weighted edges

5. Discussion

The last part focuses on the discussion of the results showed in chapter 4. Thereby, the focus lies on the single proteins, which were identified using the *Biomarker Identifier* method by comparing them to the results of the SMI data, literature and databases, but also the interactions around the proteins by analyzing the networks shall be discussed.

A list of the verified proteins for the *non-targeted* and the *targeted* approach is displayed in the supplement in table A.3 with the used protein name, the full protein name, the gene name and the UniProt ID.

5.1. Protein Verification

The *paired Biomarker Identifier* score was the first step to prioritize the proteins and, therefore, gives also a first verification possibility of the proteins. The barplot diagrams are used in this step to analyze the changes of the protein scores over time in different cohorts. Figure 4.1 can be used as a comparison of any relevant and significant proteins.

The best known biomarker to detect a myocardial infarction is troponin I, which is also clinically used. However, also CK-MB and Myoglobin are known biomarkers, in this instance. Figure 5.1 shows a graphical representation of the relative changes of those three known biomarkers after the onset of an acute myocardial infarction. This illustration is supposed to give a general overview of the relative changes of these three biomarkers over the course of a few days, but does not show an exact representation.

Figure 5.1 is intended to illustrate the significance of troponin I, which can also be seen in the barplot diagram 4.1 of the SMI data. Troponin I is widely used because it has a sensitivity of 100% and a specificity of 96.3% [33]. The uBI score of troponin I exceeds the other scores, considerably. Prolonged myocardial ischemia results in myocardial cell death, which is the reason why troponin I gets spilled in the blood leading to the increase of troponin I levels in the first few hours and the level can stay increased for a few days. CK-MB and Myoglobin also appear in the top 30 proteins having a high uBI score. The second highest uBI score in figure 4.1 belongs to the fatty acid-binding protein (FABP) which is a previously identified marker of myocardial injury by various groups [22]. The next sections will focus on the verification and discussion of the barplot diagrams using the *non-targeted* and *targeted* approach.

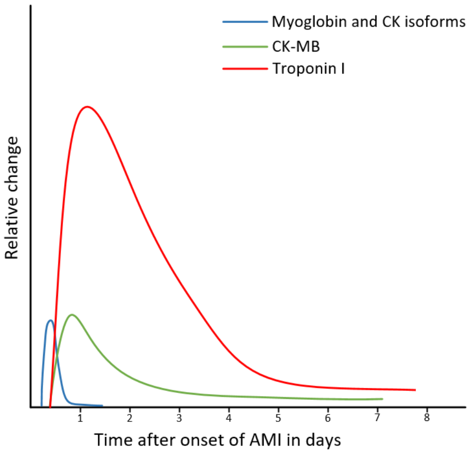


Figure 5.1.: Illustrative diagram of the relative changes of Myoglobin, CK-MB and Troponin I after the onset of an acute myocardial infarction (AMI). [33]

5.1.1. Non-targeted Approach

Table A.1 gives a general overview of the highest relative changes in each cohort. As expected, the change of troponin I levels increase over the course of time, consequently, it can be found on top of the list for the derivation and validation cohort 1, followed by CK-MB. Since the level increases in the first few hours, troponin I can be found in the first third for the validation cohort 2, because only two time points are available. Nonetheless, the rapid increase of the level can be seen by comparing it to other levels like CK-MB. Table A.1 shows, that Myoglobin just appears to play a role in the derivation cohort 2, where blood was drawn from the coronary sinus. Figure 5.1 shows that Myoglobin levels decrease fast, which might be the reason why Myoglobin is not included in the table for the derivation and validation cohort 1, since it does not play an important role anymore after 24 hours.

The increase of troponin I levels can also be observed by looking through the barplot diagrams of the derivation and validation cohort 1 over the course of time. After 10 minutes the troponin I level is rather low, seen in figure 4.2, after 1 hour in figure A.1 the level increases and reaches a peak value at 24 hours, seen in figure A.4. Thus, individual protein levels can be analyzed through time using the barplot diagrams in addition to comparing them to the other cohorts.

In contrast to the table, the barplot diagrams in section 4.2.1 show the highest pBI scores and include some features of a diagnostic test into the score and show, therefore, another way of evaluating the proteins. By comparing the barplot diagrams for all cohorts at 10 minutes in figures 4.2, 4.3 and 4.4 some proteins can be verified in all diagrams. Troponin I and CK-MB are yet to rise, therefore, in the lower part of the diagrams. Proteins like GFRa-1, CCL28, TWEAK, HGF and sFRP-3 appear in the upper half of the diagrams for every cohort. Certain proteins in the barplot diagrams are known biomarkers, and there are many novel proteins, which are not (yet) associated with myocardial infarction or atherosclerosis.

However, by checking different databases, the `IPA` database confirms that TWEAK, for instance, plays a role in the atherosclerosis signaling pathway. It is also involved in apoptosis (programmed cell death), which is increasingly observed upon plaque development, though, the exact involvement of apoptosis in the development and the course of atherosclerosis is controversial [34]. FABP appears also in all diagrams and can be confirmed by different literature, see [22, 24, 25]. Examining the diagrams at 1 hour in figures A.1,A.2 and A.3, it can be seen that the level of FABP increases over the course of time. In the `IPA` database TFPI is associated with myocardial ischemia and PECAM1 is involved in fluid shear stress and atherosclerosis according to the KEGG database and in coronary artery diseases according to `IPA`.

Leucocytes play an essential role in atherosclerotic plaque rupture and, therefore, myoeloperoxidase in leucocytes might have an impact, since myoeloperoxidase has been found in atheromatous plaques. The level of Myoeloperoxidase peaks early after an AMI and decreases over time, which can be obseverd in the different barplot diagrams. However, it does not predict heart failure, but high levels can predict MI after one year. These current findings still need further studies to analyze the actual role of Myoeloperoxidase. [25]

The decreasing levels, shown by the blue bars in the diagrams, are an interesting appearance which could not yet be further verified. However, it is noticeable, that the same proteins show a decreasing level at 10 minutes but also at 1 hour. These proteins include the Apo-B and Apo-E, and PF-4. According to `IPA` aspirin, given to patients after a MI, decreases the release of human PF-4, which might be a possible explanation for the decrease of PF-4. This decrease is very low at 24 hours for the derivation and shows a slight increase in the validation cohort 1, since the influence of aspirin has vanished after the first few hours. According to the IPA database, Apo-B and Apo-E are associated with the atherosclerosis signaling pathway and Apo-E is involved in the negative regulation of inflammatory response, according to UniProt. The decrease of those proteins needs further investigation, though.

5.1.2. Targeted Approach

The *targeted* approach was used to verify a selected subset of proteins in the PMI data. The barplot diagrams at 10 minutes shown in figures 4.8, 4.9 and 4.10 include some of the already discussed proteins for the *non-targeted* approach in section 5.1.1. Again, the increase of troponin I and CK-MB can be observed over time. Myoglobin plays a role in the *targeted* approach, since it appears in the upper half of the uBI scores in figure 4.1. The same effect of PF-4 can be seen in the barplot diagrams at the different time points. FGF-18 is a novel protein in the context of early myocardial injury [22], which also appears and shows an increasing level at 10 minutes and 1 hour and decreases again at 24 hours in the derivation and validation cohort 1.

The objective of the *targeted* approach was to select a subset of proteins, which are relevant for a SMI and verify them in the PMI data over time. Therefore, some proteins which did not appear in the *non-targeted* approach play a role here. If figure 4.1 is compared to the other barplot diagrams of the *targeted* approach, it can be seen, that proteins, which have a high uBI score do not necessarily also appear in the upper half of the barplot diagrams of the PMI data. These proteins are, for instance, PTN, MDHC and Midkine. Solely, MDHC levels seem to increase and can be seen in the barplot diagrams in figures A.11 and A.12 at 24 hours with a high score. Whereas, PTN and Midkine can mostly be found in the middle section of the barplot diagrams.

Another interesting protein is HGF, which is involved in the negative regulation of inflammatory processes and Endocan, which is involved in angiogenesis, according to UniProt. Angiogenesis, the formation of new microvessels, could possibly contribute to the salvage of ischemic myocardium at a early stage after a MI and, therefore, proangiogenic therapies have prompted numerous clinical trials to investigate this promising strategy [35]. Annexin A6 shows a decreasing level in the barplot diagrams at 10 minutes and 1 hour and according to `IPA`, annexin A6 is involved in the apoptotic signaling pathway. Also by `IPA`, BMP-10 plays a role in cardiac hypertrophy by the negative regulation of cardiac muscle hypertrophy. BMP-10 shows an increasing level in all cohorts and for all time points in the middle section of the barplot diagrams. Nevertheless, the role of these and other proteins in myocardial infarction needs to be further verified.

5.2. Dynamic Network Verification

The dynamic networks represent the second step of the analysis, after selecting a subset of proteins based on the pBI score. The creation of the networks is based on the pBI* score, calculated of the logarithmic ratio of the proteins, explained in section 3.3. Since this is a new way to analyze the relations between proteins, the represented interaction has to be interpreted in different ways as it might be an interaction in terms of change of expression and concentration, changes in function and changes of conformity. Thus, the interpretation of these networks and the interactions pose a challenge, which requires further investigation.

The single networks at each time step for the *non-targeted* and *targeted* approach show the change of the interactions through time, which can be validated in the validation cohorts. In addition to the single networks, the combined networks were created, whereas, the continuous weighted edges are a method to show the strength of the interaction through time and the discrete weighted edges, on the other hand, represent the frequency of the interaction between two proteins. For the continuous weighted edges a thick line could represent one strong interaction at one time point or a continuous, moderate strong interaction through time. The thicker the line for the discrete weighted edges, the more frequent an interaction has taken place between two proteins.

5.2.1. Non-targeted Approach

Figure 4.5 shows the networks at each time point for the derivation and validation cohort 1. By examining subfigures (a-c) and (d-f), the changes of interactions for the individual cohorts can be seen through time. The shift of the interactions to troponin I and CK-MB at 24 hours is clearly visible. While these two proteins do not show many interactions at a early stage, they dominate the networks at 24 hours. For the derivation cohort Apo-E plays a significant role, while for the validation cohort 1 other proteins like GFRa-1, HGF, sFRP-3 and TGF-b1 show the most interactions at an early stage. Figure 4.6 represents the networks for the validation cohort 2, where a similar domination of interactions can be observed as in validation cohort 1. Thus, by examining the networks carefully some common interactions can be filtered out. For instance, the interaction of GFRa-1 and Apo-E2 appears for all cohorts at 10 minutes and the interaction of HGF and TGF-b1 even appears for all cohorts at 10 minutes and 1 hour. The networks in figures 4.5 and 4.6 were constructed with a threshold of $\tau = q90$, more networks with a threshold of $\tau = q75$ are illustrated in the supplement A.1. A lower threshold includes more interactions, and therefore, enables a more detailed interpretation and verification phase.

The combined networks in figure 4.7 summarize all the interactions through time into one single network in a clear way. This way to visualize the networks brings additional information to the interpretation. By comparing subfigure (a) and (b), the thick lines in subfigure (b) show that these interactions appeared at every time point, but by looking at subfigure (a), they did not happen to be that strong over time. A stronger and more frequent interaction between GFRa-1 and TGF-b1 and between MFGM and FGF7 can be observed in subfigures (c) and (d) of the validation cohort 1, though. The connection of FGF7 and MFGM in subfigures (e) and (f) of the validation cohort 2, appears to be the strongest over time, followed by TFPI and CCL28.

5.2.2. Targeted Approach

For the *targeted* approach the single networks are depicted in figures 4.10 and 4.11. Unlike for the *non-targeted* approach, troponin I already shows some interactions at an early stage in all cohorts. Again, GFRa-1 plays a significant role at 10 minutes. In addition to that, HGF shows many interactions at 10 minutes. The networks at 24 hours are dominated by the interactions of troponin I, CK-MB and CK-MM. By taking a close look, common interactions can also be found, like GFRa-1 and Midkine, which shows an early interaction at 10 minutes. Like for the *non-targeted* approach, these networks were created with a threshold of $\tau = q90$, for a more detailed interpretation, the networks with a threshold $\tau = q75$ are illustrated in the supplement A.2.

For a better overview, the combined networks in figure 4.12 can be taken into account. It can be noticed, that the networks for the derivation in subfigures (a-b) and validation cohort 1 in subfigures (c-d) look similar between the continuous and discrete weighted edges. This can be interpreted, that the interactions over time appear to be frequent in a continuous strength for both cohorts. On the contrary, the most significant interactions for the validation cohort 2 in subfigures (e-f) appear to happen frequently but they show a low strength.

It needs to be mentioned, that all these interactions for the *non-targeted* just as for the *targeted* approach need further investigation and studies to be verified. Also, not only a strong interaction could be relevant but also single interactions could be of importance for a myocardial infarction. Like mentioned before, this poses a challenge in the interpretation and verification of these results.

5.2.3. Shared Genetics

Further research was conducted to investigate some of the interactions, which appear at an early stage. Because of the high amount of interactions, a few single ones were chosen from the *non-targeted* approach to gain a closer insight. The goal was to identify significant interactions, which appear at 10 minutes and 1 hour in all cohorts and play a role in myocardial infarction. Therefore, the networks with the threshold $\tau = q75$ was taken into account to get a list of common interactions. This list included, for example HGF/MFGM, GFRa-1/PECAM-1, TWEAK/Apo-E, Spondin-1/MFGM, Apo-E4/TFPI and 6-Ckine/Apo-B. First of all, data from a genome-wide association study (GWAS) was consulted to identify if the two proteins showing the interaction, share common SNPs (single nucleotide polymorphisms) on a gene. The above listed interactions were all identified to share some SNPs. Next, another GWAS database was downloaded, including SNPs related to cardiovascular diseases and myocardial infarction. This database is freely available online.[i] Some of the SNPs shared by two proteins could actually be found in this database. These include the proteins with the interactions of TWEAK/Apo-E, Spondin-1/MFGM, Apo-E4/TFPI and 6-Ckine/Apo-B, however they were not statistically relevant.

5.3. Annotations

Finally, there are some things which have to be considered for the interpretation. Patients undergoing a PMI are given heparin, which can have an influence on some proteins the first few hours. After 6-8 hours this influence is gone and the results, especially at 24 hours, can be seen as clinically relevant. Also, the sample size differs between the cohorts, validation cohort 2 only includes 6 patients, in contrast, the derivation and validation cohort 1 include each 15 patients. But, the interindividual variability can be reduced, since the patients could serve as their own biological control. Patients which suffered from a SMI came into the hospital at different times, which does not result in an exact baseline value, nonetheless, for convenience this value was considered as the baseline. Regarding the interpretation and verification of the proteins, the findings also depend on the databases and literature used. Most databases refer to other databases, and therefore, do not necessarily always include all the information found in another database. Finally, it should be noted, that the objective of this method is to analyze data through time, thus, it can only be used if a minimum of two time points are available.

[i] CARDIoGRAMplusC4D 1000 Genomes-based GWAS,
downloaded from *www.cardiogramplusc4d.org/data-downloads* as of 29.05.2018

6. Conclusion

The method and the results presented in this thesis show, that the *Biomarker Identifier* score and the networks, as a visualization of the interactions between the proteins, are a new promising strategy for analyzing proteomic data in a different way. It has been shown, that by taking some features of the diagnostic test into the score the proteins can be prioritized and selected to not just add information to already known biomarkers but also to select novel biomarker candidates.

Another dimension of information is added by looking at the proteins and their interactions through time by creating the network graphs of a selected subset of proteins, based on the interactions between them. These interactions, interpreted in different ways, can be analyzed around each protein at a each time point and through time in one single network. Additionally, not only the strength of the interaction but also the frequency can be visualized in specific network graphs. By using different thresholds, a more or less strict analysis can be achieved.

Proteins identified in this thesis could be validated as already known biomarkers in the context of cardiovascular diseases, but there have been many proteins which are new and are in need of further investigation and verification. Also, the networks and interactions between the proteins shown in this thesis are new in terms of proteomic analyses. A careful interpretation, verification and validation are necessary to verify and draw conclusions. The further investigation of some interactions with shared genetics has shown, that this is a promising method to discover new information, even though, so far these results were not statistically relevant. In the course of this thesis much new and interesting information was revealed, which will be pursued by further investigation.

A. Appendix: Figures and Tables

Additional tables and figures are illustrated in this section for each the *non-targeted* and the *targeted* approach, including barplot diagrams at more time points and networks with a different threshold τ.

A.1. Figures: Non-targeted Approach

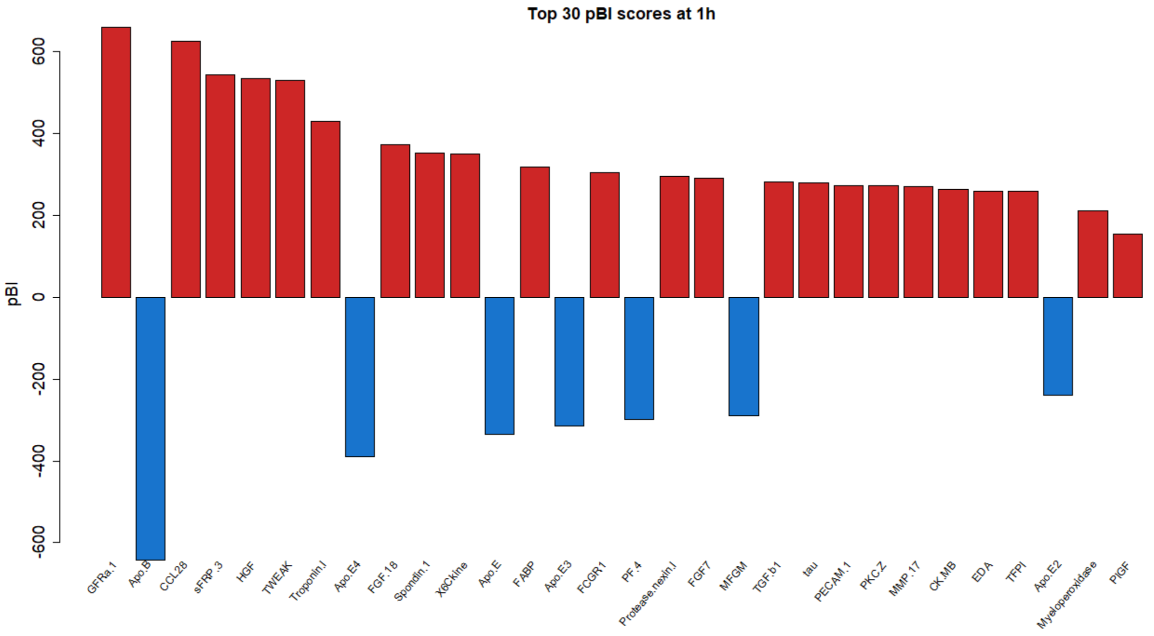


Figure A.1.: Barplot diagram of the top 30 ranked proteins (pBI scores) of the derivation cohort at 1 hour using the *non-targeted* approach.

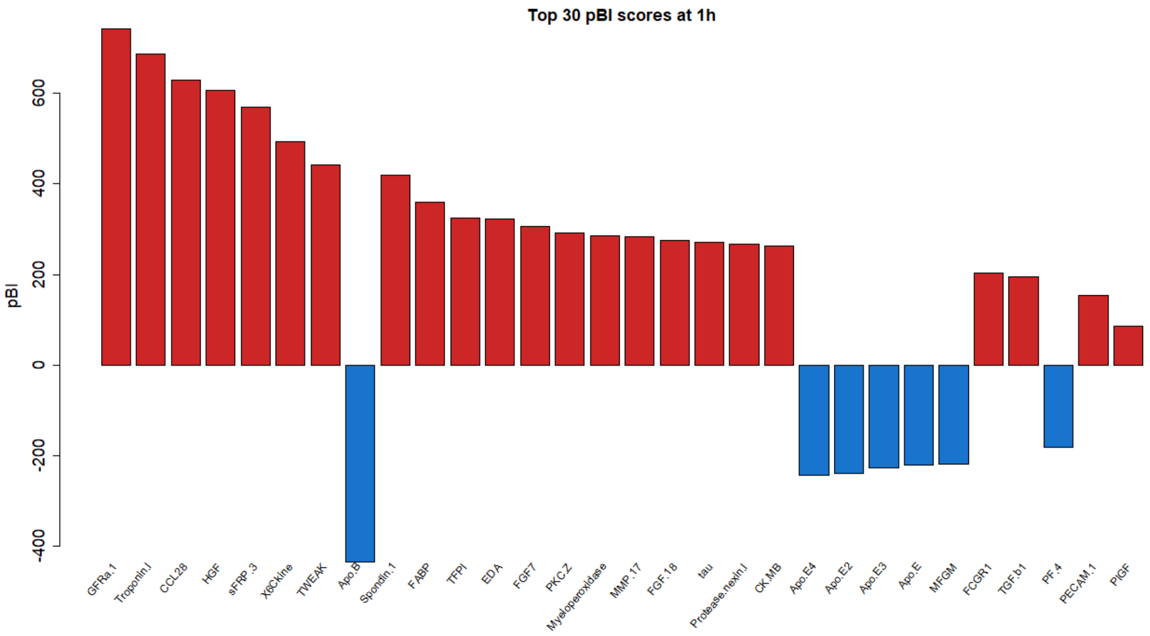


Figure A.2.: Barplot diagram of the top 30 ranked proteins (pBI scores) of the validation cohort 1 at 1 hour using the *non-targeted* approach.

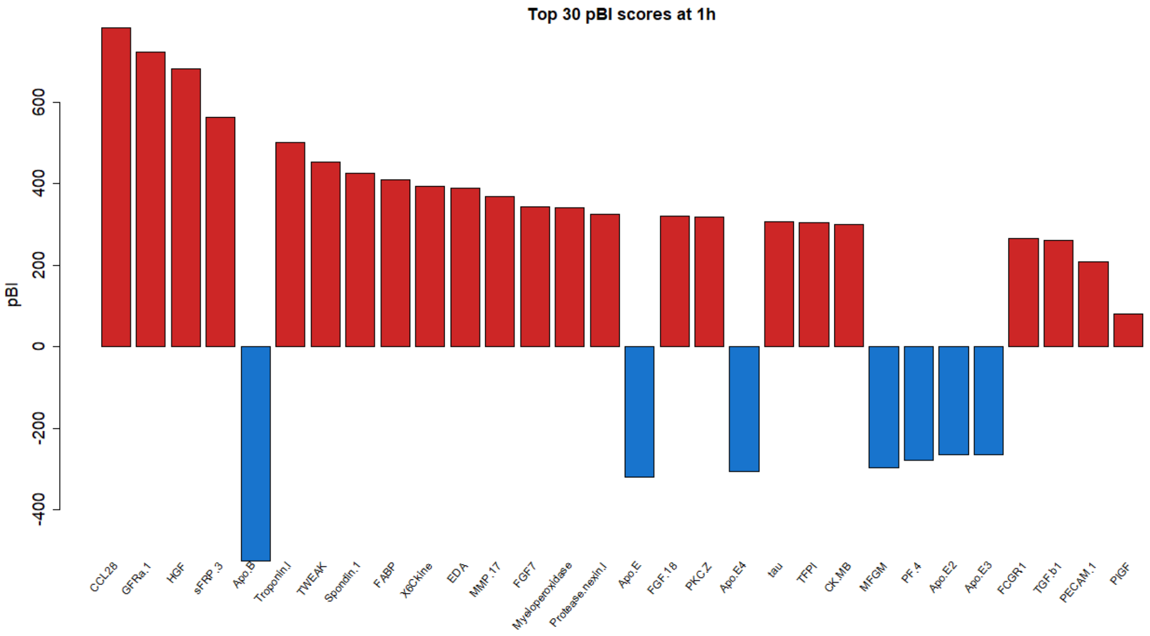


Figure A.3.: Barplot diagram of the top 30 ranked proteins (pBI scores) of the validation cohort 2 at 1 hour using the *non-targeted* approach.

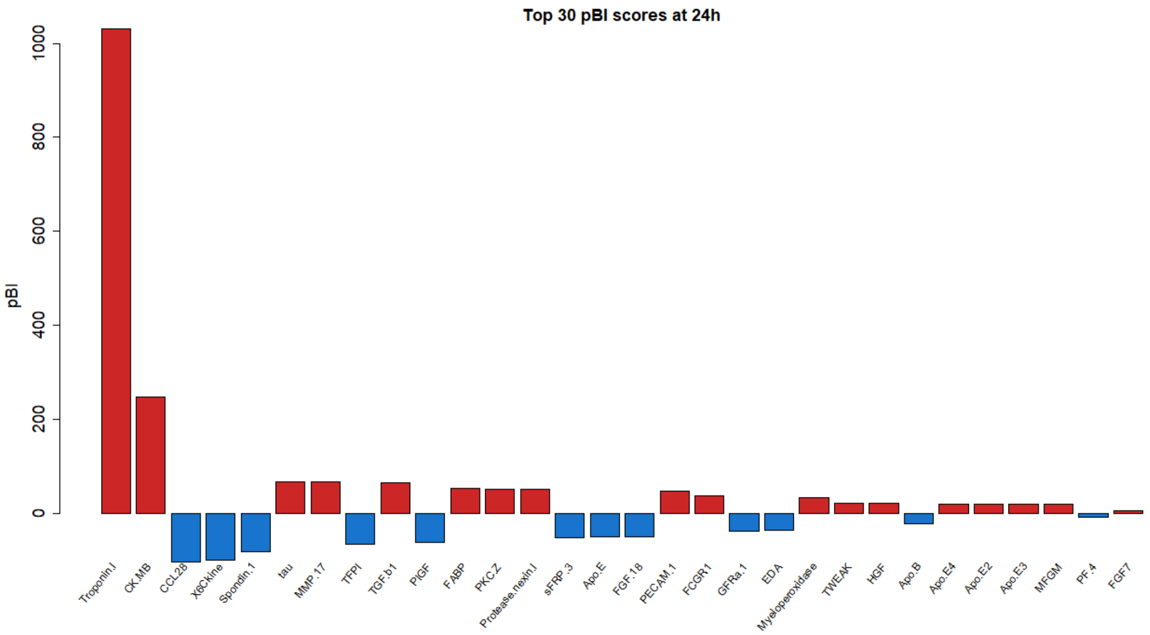


Figure A.4.: Barplot diagram of the top 30 ranked proteins (pBI scores) of the derivation cohort at 24 hours using the *non-targeted* approach.

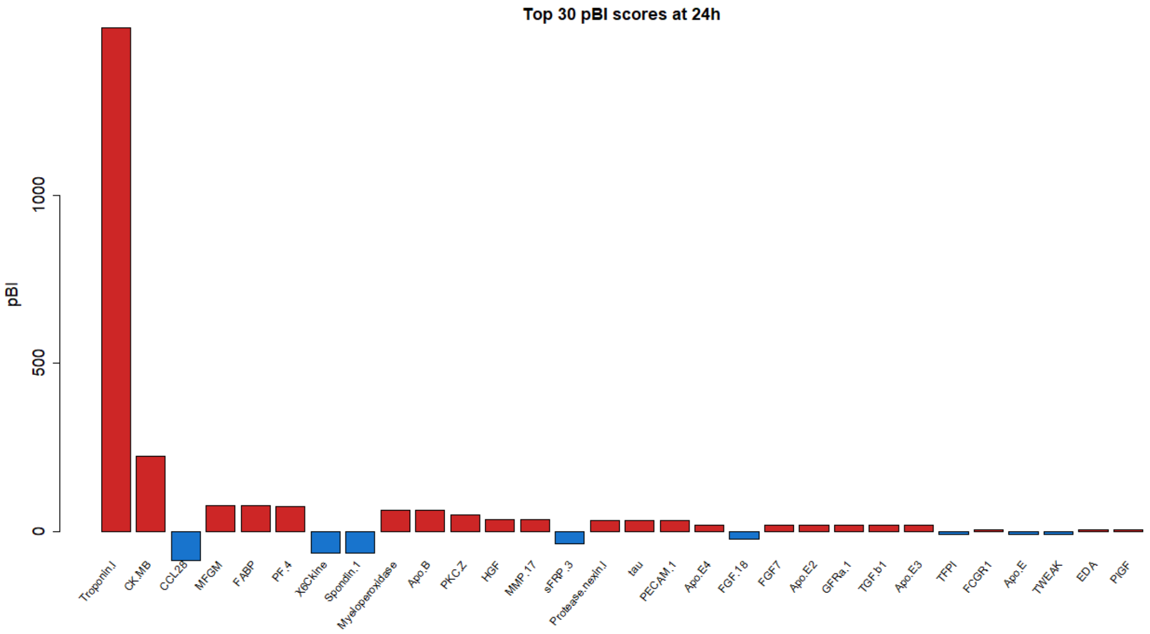


Figure A.5.: Barplot diagram of the top 30 ranked proteins (pBI scores) of the validation cohort 1 at 24 hours using the *non-targeted* approach.

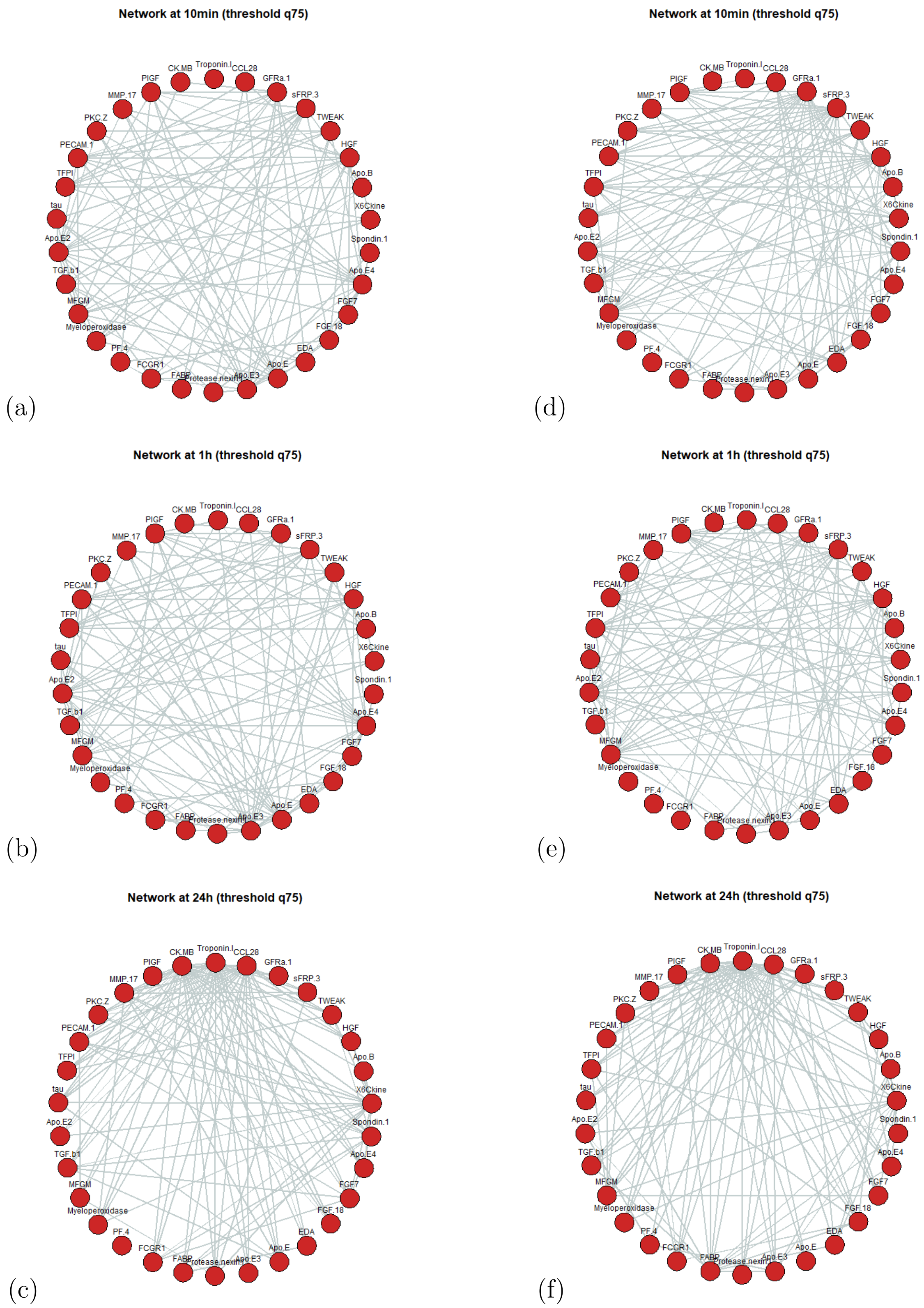


Figure A.6.: Network representation of the selected subset of proteins of the derivation cohort (a-c) and validation cohort 1 (d-f) using the *non-targeted* approach at different measurement times with a threshold of $\tau = q75$. (a) derivation at 10min, (b) derivation at 1h, (c) derivation at 24h, (d) validation 1 at 10min, (e) validation 1 at 1h, (f) validation 1 at 24h

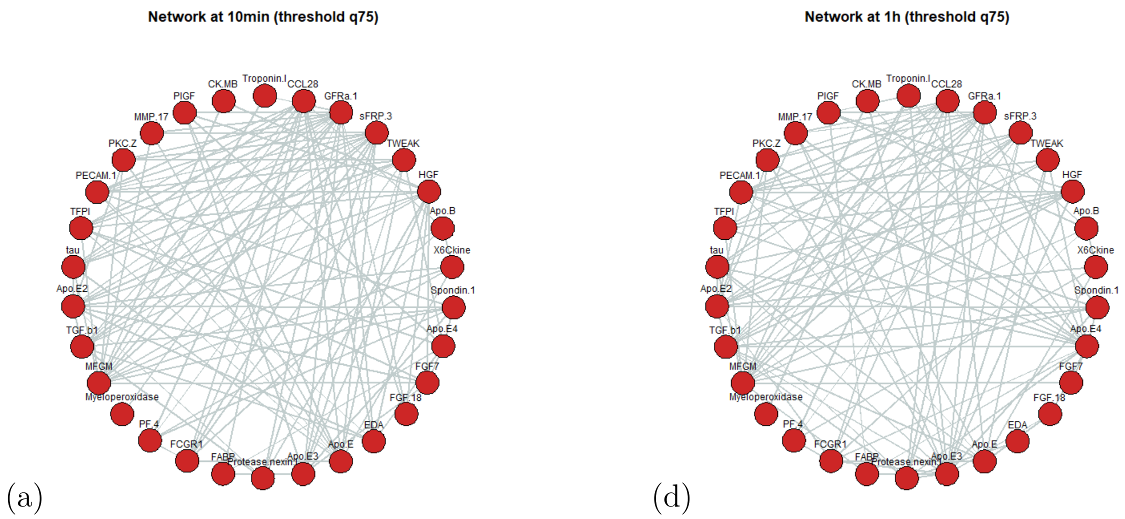


Figure A.7.: Network representation of the selected subset of proteins of the validation cohort 2 using the *non-targeted* approach at different measurement times with a threshold of $\tau = q75$. (a) validation 2 at 10min, (b) validation 2 at 1h

A.2. Figures: Targeted Approach

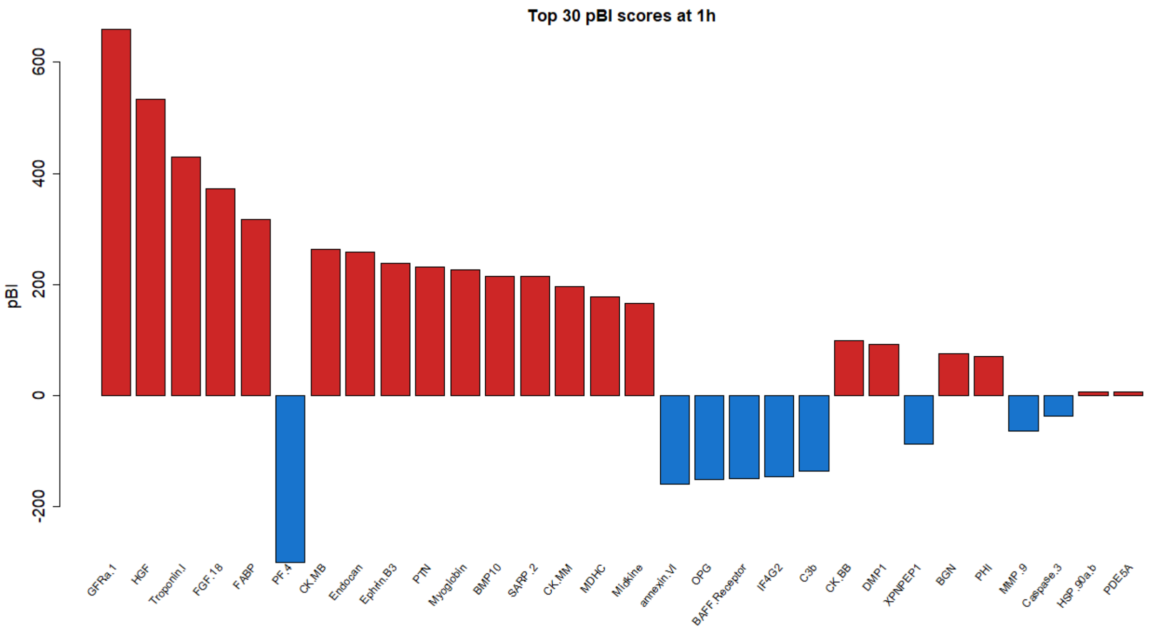


Figure A.8.: Barplot diagram of the top 30 ranked proteins (pBI scores) of the derivation cohort at 1 hour using the *targeted* approach.

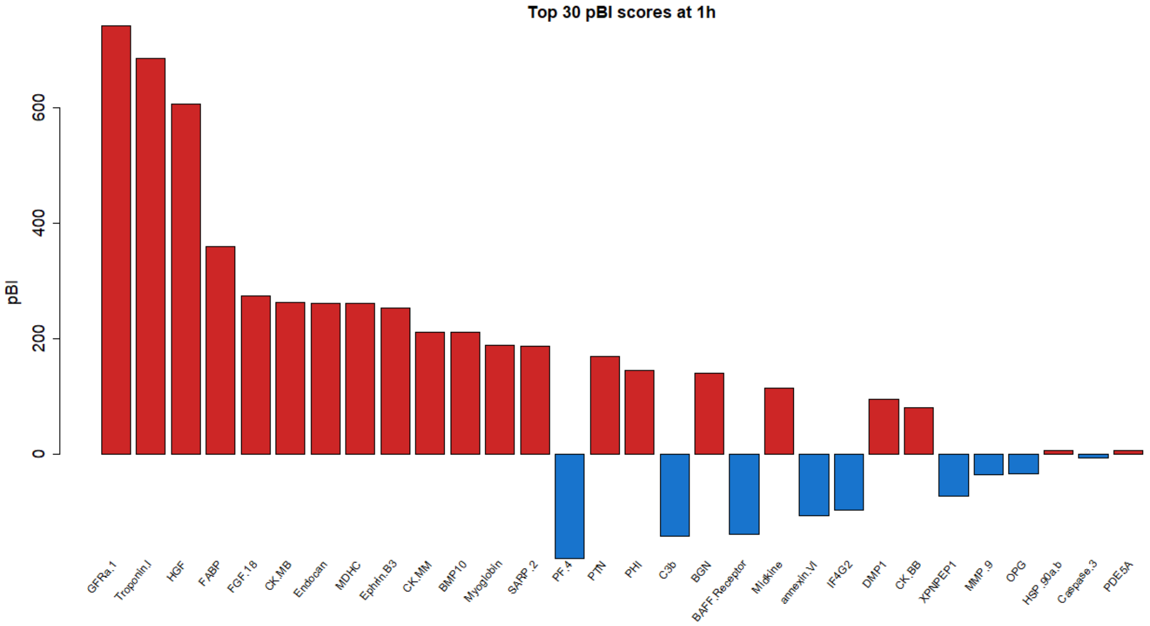


Figure A.9.: Barplot diagram of the top 30 ranked proteins (pBI scores) of the validation cohort 1 at 1 hour using the *targeted* approach.

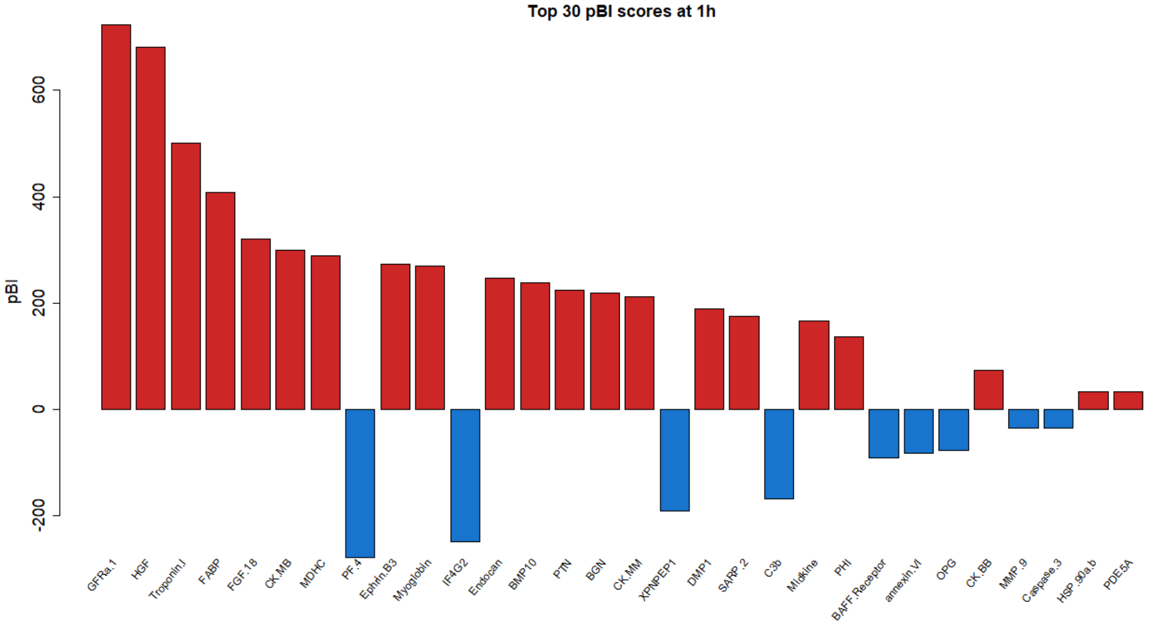


Figure A.10.: Barplot diagram of the top 30 ranked proteins (pBI scores) of the validation cohort 2 at 1 hour using the *targeted* approach.

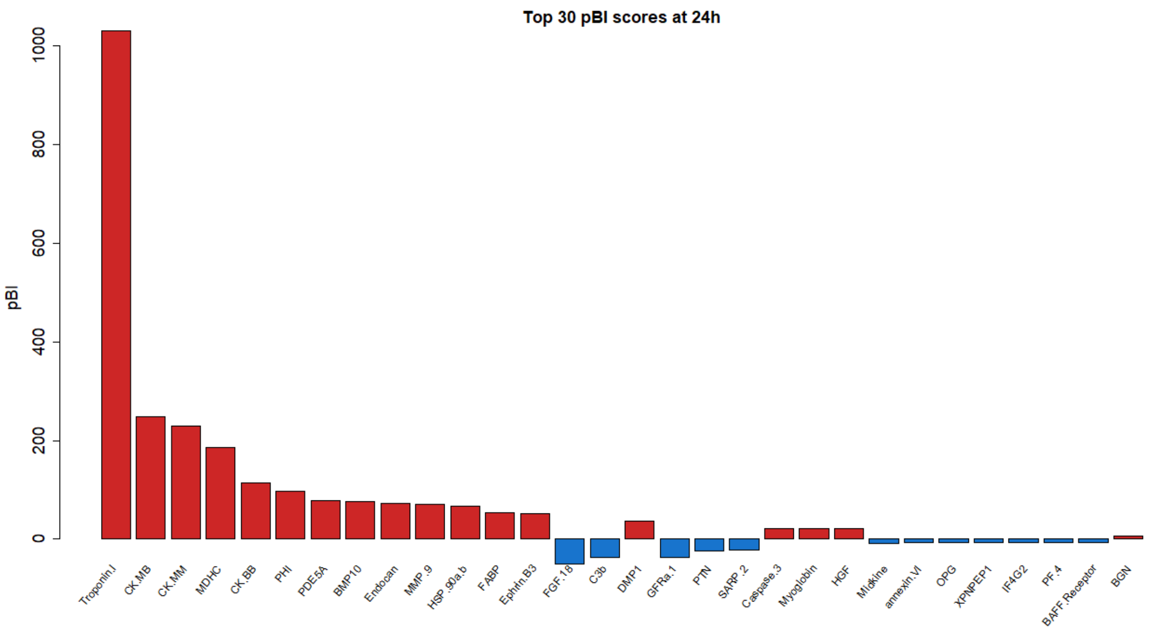


Figure A.11.: Barplot diagram of the top 30 ranked proteins (pBI scores) of the derivation cohort at 24 hours using the *targeted* approach.

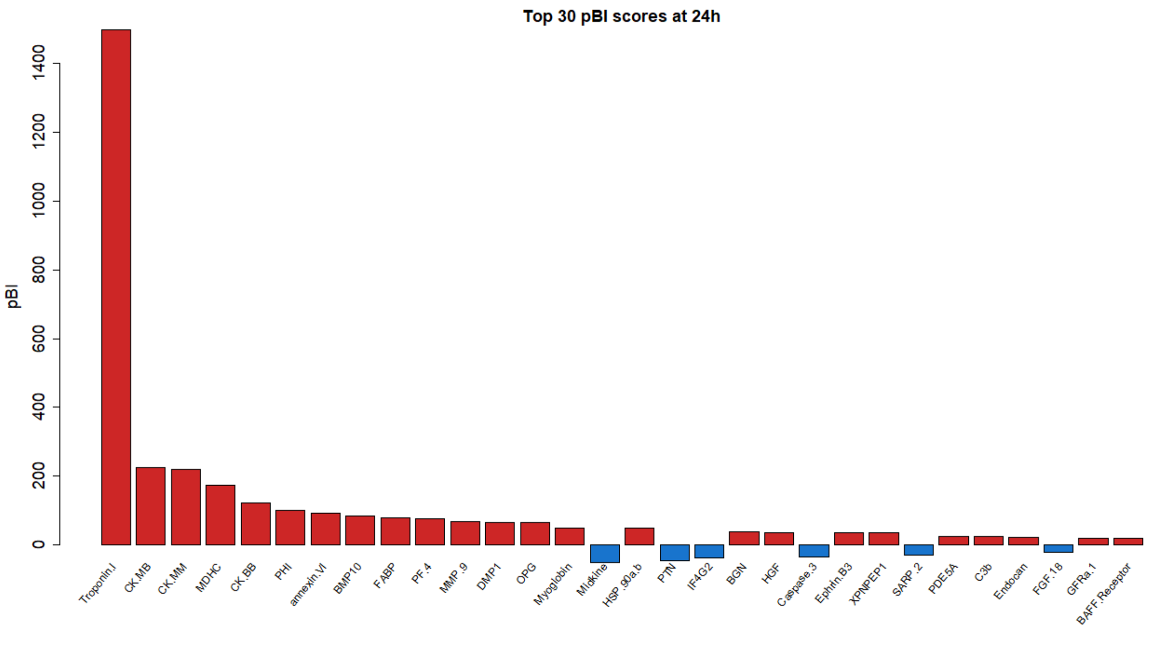


Figure A.12.: Barplot diagram of the top 30 ranked proteins (pBI scores) of the validation cohort 1 at 24 hours using the *targeted* approach.

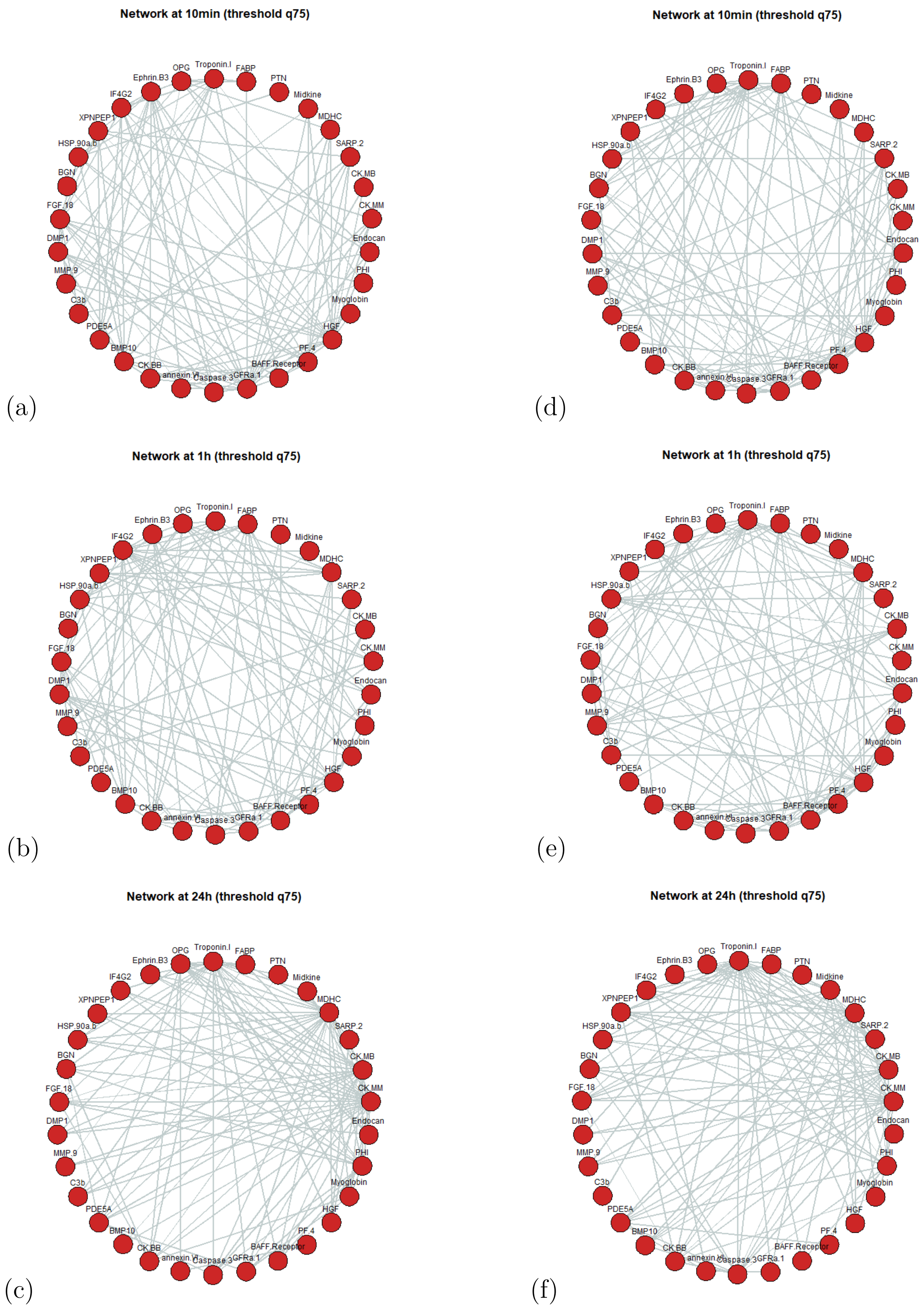


Figure A.13.: Network representation of the selected subset of proteins of the derivation cohort (a-c) and validation cohort 1 (d-f) using the *targeted* approach at different measurement times with a threshold of $\tau = q75$. (a) derivation at 10min, (b) derivation at 1h, (c) derivation at 24h, (d) validation 1 at 10min, (e) validation 1 at 1h, (f) validation 1 at 24h

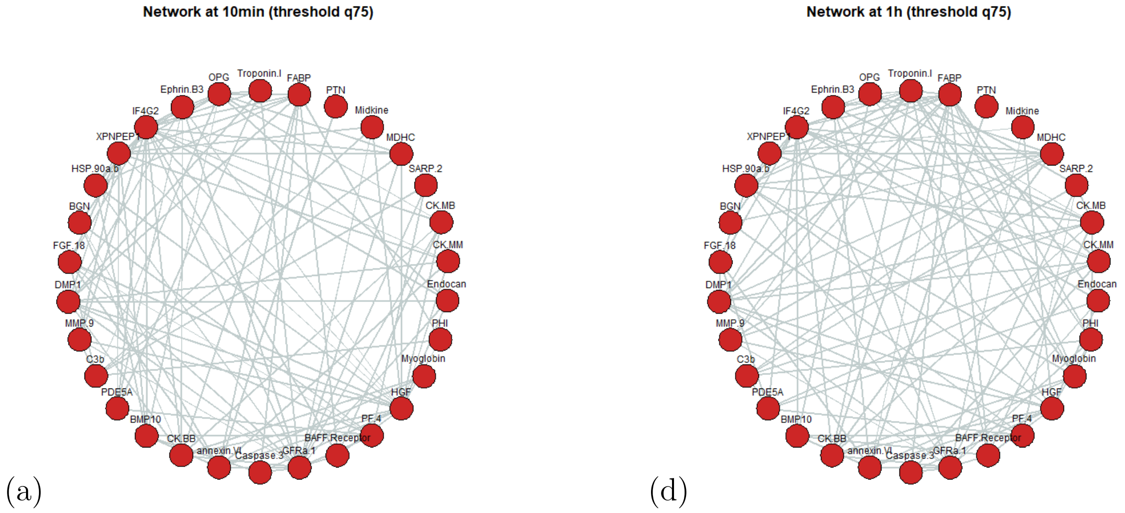


Figure A.14.: Network representation of the selected subset of proteins of the validation cohort 2 using the *targeted* approach at different measurement times with a threshold of $\tau = q75$. (a) validation 2 at 10min, (b) validation 2 at 1h

A.3. Tables

Derivation cohort				**Validation cohort 1**				**Validation cohort 2**		
Protein	10min/BL [%]	1h/BL [%]	24h/BL [%]	Protein	10min/BL [%]	1h/BL [%]	24h/BL [%]	Protein	10min/BL [%]	1h/BL [%]
Troponin I	94.53	253.31	1312.99	Troponin I	100.86	363.48	1598.21	CCL28	303.02	240.14
CK.MB	30.56	52.02	72.28	CK.MB	30.53	78.28	72.31	HGF	221.45	153.1
SP.D	18.55	12.35	61.39	C3b	11.18	10.54	68.04	GFRa.1	220.19	158.12
CK.MM	19.59	37.81	60.36	CK.MM	15.77	36.42	39.35	TWEAK	134.4	99.25
Histone.H1.2	11.06	12.1	52.57	FABP	41.86	114.54	27.75	sFRP.3	128.69	94.19
MDHC	14.74	29.88	32.3	SARP.2	49.91	55.21	27.31	X6Ckine	82.88	61.71
PHI	10.61	15.58	26.8	MDHC	14.13	33.01	26.54	Troponin.I	73.9	318.74
HMG.1	14.94	13.67	25.29	PHI	13.19	20.33	24.1	Spondin.1	69.75	50.16
Ferritin	11.5	10.52	24.5	BSP	24.18	24.4	20.03	CK.MB	68.43	144.41
PTN	57.87	82.67	22.7	Activin.A	10.98	10.15	19.72	FGF7	58.74	44.78
Cadherin.12	29.08	31.5	22.61	Midkine	43.27	38.17	19.54	FABP	53.19	122.09
LDH.H.1	10.44	11.08	22.41	LDH.H.1	10.59	11.26	19.48	TECK	48.64	22.17
Epo	14.07	15.06	22.37	GFRa.1	215.53	159.05	17.45	EDA	42.49	37.35
PGD	11.81	10.1	22	PTN	32.68	52.67	17.42	Myoglobin	42.05	52.53
vWF	14.14	16.88	21.01	SP.D	21.7	17.47	17.31	SARP.2	40.56	51.45
GAPDH..liver	15.24	17.61	20.94	Histone.H1.2	15.52	18.34	16.85	PTN	34.37	62.86
hnRNP.A2.B1	12.27	11.68	20.93	PDGF.BB	20.07	13.94	16.46	Dkk.4	31.08	20.45
Cathepsin.G	27.09	28.88	20.46	Integrin.a1b1	12.88	14.33	15.39	TFPI	30.29	27.99
tau	29.29	31.89	20.04	calgranulin.B	14.68	13.45	15.14	FGF.18	28.67	57.51
PKC.Z	34.62	36.23	19.65	BPI	13.72	12.23	14.52	PDGF.BB	28.31	19

Table A.1.: Table of the top 20 percentage changes of protein RFU levels in all cohorts at all measurement time points versus baseline for the *non-targeted* approach.

Derivation cohort				Validation cohort 1				Validation cohort 2		
Protein	10min/BL [%]	1h/BL [%]	24h/BL [%]	Protein	10min/BL [%]	1h/BL [%]	24h/BL [%]	Protein	10min/BL [%]	1h/BL [%]
Troponin.I	94.53	253.31	1312.99	Troponin.I	100.86	363.48	1598.21	HGF	221.45	153.1
CK.MB	30.56	52.02	72.28	CK.MB	30.53	78.28	72.31	GFRa.1	220.19	158.12
CK.MM	19.59	37.81	60.36	C3b	11.18	10.54	68.04	Troponin.I	73.9	318.74
MDHC	14.74	29.88	32.3	CK.MM	15.77	36.42	39.35	CK.MB	68.43	144.41
PHI	10.61	15.58	26.8	FABP	41.86	114.54	27.75	FABP	53.19	122.09
PTN	57.87	82.67	22.7	SARP.2	49.91	55.21	27.31	Myoglobin	42.05	52.53
FABP	25.39	57.82	18.75	MDHC	14.13	33.01	26.54	SARP.2	40.56	51.45
PDE5A	11.48	11.57	17.32	PHI	13.19	20.33	24.1	PTN	34.37	62.86
Ephrin.B3	25.11	26.94	16.94	Midkine	43.27	38.17	19.54	FGF.18	28.67	57.51
Endocan	21.07	23.74	16.6	GFRa.1	215.53	159.05	17.45	CK.MM	28.16	64.8
BMP10	24.81	25.99	15.33	PTN	32.68	52.67	17.42	Midkine	22.18	20.64
HSP.90a.b	10.83	10.59	15.08	HGF	214.59	154.47	14.42	BMP10	21.61	22.82
HGF	209.23	168.61	14.91	BGN	21.93	20.57	14.24	Ephrin.B3	20.19	22.67
Midkine	46.14	46.27	14.55	Endocan	19.56	21.73	14.12	Endocan	18.98	21.31
Caspase.3	10.44	10.03	12.99	PDE5A	11.19	11.41	13.66	MDHC	17.51	41.89
DMP1	13.99	14.19	12.49	BMP10	20.48	24.75	13.23	BGN	16.37	15.91
Myoglobin	22.47	29.22	11.7	Ephrin.B3	19.27	23.67	12.99	DMP1	12.65	11.9
GFRa.1	217.31	172.29	11.47	FGF.18	22.39	39.54	12.8	PHI	12.5	21.76
BGN	16.09	17.47	11.03	HSP.90a.b	10.61	11.17	12.07	MMP.9	12.02	10.05
SARP.2	42.35	52.07	10.29	DMP1	13.68	13.88	11.88	PDE5A	9.96	9.59

Table A.2.: Table of the top 20 percentage changes of protein RFU levels in all cohorts at all measurement time points versus baseline for the *targeted* approach.

Protein	Full Protein Name	Gene Name	UniProt ID
annexin.VI	Annexin A6	ANXA6	P08133
Apo.B	Apolipoprotein B	APOB	P04114
Apo.E	Apolipoprotein E	APOE	P02649
Apo.E2	Apolipoprotein E2	APOE	P02649
Apo.E3	Apolipoprotein E3	APOE	P02649
Apo.E4	Apolipoprotein E4	APOE	P02649
BAFF.Receptor	Tumor necrosis factor receptor superfamily member 13C	TNFRSF13C	Q96RJ3
BGN	Biglycan	BGN	P21810
BMP10	Bone morphogenetic protein 10	BMP10	O95393
C3b	Complement C3b	C3	P01024
Caspase.3	Caspase-3	CASP3	P42574
CCL28	C-C motif chemokine 28	P42574	Q9NRJ3
CK.BB	Creatine kinase B-type	CKB	P12277
CK.MB	Creatine kinase M-type:Creatine kinase B-type heterodimer	CKB CKM	P12277 P06732
CK.MM	Creatine kinase M-type	CKM	P06732
DMP1	Dentin matrix acidic phosphoprotein 1	DMP1	Q13316
EDA	Ectodysplasin-A	EDA	Q92838
Endocan	Endothelial cell-specific molecule 1	ESM1	Q9NQ30
Ephrin.B3	Ephrin-B3	EFNB3	Q15768
FABP	Fatty acid-binding protein, heart	FABP3	P05413
FCGR1	High affinity immunoglobulin gamma Fc receptor I	FCGR1A	P12314
FGF.18	Fibroblast growth factor 18	FGF18	O76093
FGF7	Fibroblast growth factor 7	FGF7	P21781
GFRa.1	GDNF family receptor alpha-1	GFRA1	P56159
HGF	Hepatocyte growth factor	HGF	P14210
HSP.90a.b	Heat shock protein HSP 90-alpha/beta	HSP90AA1 HSP90AB1	P07900 P08238
IF4G2	Eukaryotic translation initiation factor 4 gamma 2	EIF4G2	P78344
MDHC	Malate dehydrogenase, cytoplasmic	MDH1	P40925
MFGM	Lactadherin	MFGE8	Q08431
Midkine	Midkine	MDK	P21741
MMP.17	Matrix metalloproteinase-17	MMP17	Q9ULZ9
MMP.9	Matrix metalloproteinase-9	MMP9	P14780
Myeloperoxidase	Myeloperoxidase	MPO	P05164
Myoglobin	Myoglobin	MB	P02144

OPG	Tumor necrosis factor receptor superfamily member 11B	TNFRSF11B	O00300
PDE5A	cGMP-specific 3',5'-cyclic phosphodiesterase	PDE5A	O76074
PECAM.1	Platelet endothelial cell adhesion molecule	PECAM1	P16284
PF.4	Platelet factor 4	PF4	P02776
PHI	Glucose-6-phosphate isomerase	GPI	P06744
PKC.Z	Protein kinase C zeta type	PRKCZ	Q05513
PlGF	Placenta growth factor	PGF	P49763
Protease.nexin.I	Glia-derived nexin	SERPINE2	P07093
PTN	Pleiotrophin	PTN	P21246
SARP.2	Secreted frizzled-related protein 1	SFRP1	Q8N474
sFRP.3	Secreted frizzled-related protein 3	FRZB	Q92765
Spondin.1	Spondin 1	SPON1	Q9HCB6
tau	Microtubule-associated protein tau	MAPT	P10636
TFPI	Tissue factor pathway inhibitor	TFPI	P10646
TGF.b1	Transforming growth factor beta-1	TGFB1	P01137
Troponin.I	Troponin I, cardiac muscle	TNNI3	P19429
TWEAK	Tumor necrosis factor receptor superfamily member 12A	TNFSF12	O43508
X6Ckine	C-C motif chemokine 21	CCL21	O00585
XPNPEP1	Xaa-Pro aminopeptidase 1	XPNPEP1	Q9NQW7

Table A.3.: List of the validated proteins used for the *non-targeted* and the *targeted* approach with their full name, gene name and UniProt ID.

B. Appendix: R-Code

The following R-Code originates from [28] and was adapted and extended accordingly, whereas the package BiomarkeR can be downloaded from [27].

```
#-----------------------------------------------------------------------------
# R Script for the analysis of paired proteomic data
#-----------------------------------------------------------------------------
#                              INITIAL SETTINGS
#removing everything in working environment
rm(list=ls())

source("Functions.R")
source("pBIFunctions.R")
source("pBIGraphFunctions.R")

library(RBGL)
library(QuACN)
library(igraph)
library(graph)
library(Hmisc)
library(gplots)
library(RColorBrewer)
library(intergraph)
library(devtools)
library(pracma)
library(gridExtra)
library(grid)

today = Sys.Date()
today = format(today, format="%Y-%m-%d")

#-----------------------------------------------------------------------------
#                              READING IN DATA
cohort = "Derivation"
input = read.csv("Data/PMI_peripheral_August.csv", header = TRUE, sep = ";")
input = input[,2:ncol(input)]

cohort = "Validation"
input = read.csv("Data/PMI_peripheral_Oct.csv", header = TRUE, sep = ";")
input = input[,2:ncol(input)]

cohort = "Validation_CS"
input = read.csv("Data/PMI_coronarySinus.csv", header = TRUE, sep = ";")
input = input[,2:ncol(input)]

#-----------------------------------------------------------------------------
#                       CREATING (SUB)FOLDERS AND FILEPATHS
dir.create(paste(today, "_Plots",sep = ""))
dir.create(paste(today, "_Plots/", cohort, sep = ""))
dir.create(paste(today, "_Plots/", cohort, "/01_pBI", sep = ""))
dir.create(paste(today, "_Plots/", cohort, "/02_Graphs", sep = ""))
dir.create(paste(today, "_Plots/", cohort, "/02_Graphs/DynThresh", sep = ""))

#setting filepaths
pathGeneral = paste("./", today, "_Plots/", sep = "")
pathPlots = paste("./", today, "_Plots/", cohort, "/", sep = "")
pathpBI = paste("./", today, "_Plots/", cohort, "/01_pBI/", sep = "")
pathGraphs = paste("./", today, "_Plots/", cohort, "/02_Graphs/", sep = "")
pathGraphDynThres = paste("./", today, "_Plots/", cohort, "/02_Graphs/DynThresh/", sep = "")
```

```
#-----------------------------------------------------------------------------
#                                 PARAMETERS
if(cohort == "Validation_CS")
{
  allTimepoints = c("BL", "10min", "1h")
  timepoints = c("10min", "1h")
} else {
  allTimepoints = c("BL", "10min", "1h", "24h")
  timepoints = c("10min", "1h", "24h")
}

dynamicThreshold = c(q50 = 'q50', q75 = 'q75', q90 = 'q90')
toValidate = FALSE
topProteinsFromSMI = FALSE

#-----------------------------------------------------------------------------
#                            DATA PREPROCESSING
Data = preprocessing(input, allTimepoints, toValidate, topProteinsFromSMI, pathGeneral)
combinedData = Data$combinedData
sortedData = Data$sortedData

#-----------------------------------------------------------------------------
#                       CALCULATING PERCENTAGE CHANGES
ToInclude = 20
percIncrease = 10
plotTable = TRUE
width = 500
height = 500

proteinPercChange = calcPercentageChangeTable(sortedData, allTimepoints,
                    percIncrease = percIncrease,  numToInclude = ToInclude,
                    plotTable = plotTable, filepath = pathPlots, width = width, height = height)

#-----------------------------------------------------------------------------
#                           CALCULATING pBI SCORES
QuantileToInclude = 0.75
numberToInclude = 100

pbiCalcAll = pBIcalc(combinedData = combinedData, useMedian = FALSE, lambda = 100,
                     QuantileToInclude = QuantileToInclude,
                     numberToInclude = numberToInclude)
combinedDataTop = pbiCalcAll$combinedDataRanked
pbiScoresTop = pbiCalcAll$pbiRankedScores
pbiScoresAll = pbiCalcAll$pbiScores

if(topProteinsFromSMI == TRUE)
{
  combinedDataTop = combinedData
}

#-----------------------------------------------------------------------------
#                     BARPLOT DIAGRAMS OF pBI SCORES
numberToPlot = 30

pbiMatrix = matrix(NA, nrow = length(pbiScoresAll[[1]]), ncol = length(pbiScoresAll))
colnames(pbiMatrix) = timepoints
rownames(pbiMatrix) = names(pbiScoresAll[[1]])
scoreNames = vector("list", length = length(dynamicThreshold))

for(i in 1:length(pbiScoresAll))
{
  pbiMatrix[,i] = pbiScoresAll[[i]]
}

for(i in 1:ncol(pbiMatrix))
{
  for(j in 1:length(dynamicThreshold))
  {
    png(filename = paste(pathpBI, '30_barplot_pBI_', timepoints[i],'.png', sep = ""),
        width = 1300, height = 750)
    scoreNames[[i]] = plotScores(pbiMatrix[,i], numTopRankedToPlot = numberToPlot,
                  method = 'pBI', bars.cols = c("firebrick3", "dodgerblue3"), timepoints[i])
    dev.off()

  }
}
```

```
#-------------------------------------------------------------------------------
#                        NETWORK GRAPHS WITH DYNAMIC THRESHOLDS
numToPlotDynStatThreshold = 30
combinedDataToPlot = vector("list", length = length(combinedDataTop))

if(toValidate == TRUE)
{
  for(i in 1:length(combinedData))
  {
    combinedDataToPlot[[i]] = combinedData[[i]][,1:numToPlotDynStatThreshold]
  }
} else {
  for(i in 1:length(combinedData))
  {
    combinedDataToPlot[[i]] = combinedDataTop[[i]][,1:numToPlotDynStatThreshold]
  }
}

if(cohort == "Derivation")
{
  write.csv(names(combinedDataToPlot[[1]][3:length(combinedDataToPlot[[1]])]),
            paste(pathGeneral, "ProteinNamesFromDerivation.csv"))
}

mNamesAll = colnames(combinedDataToPlot[[1]])[3:ncol(combinedDataToPlot[[1]])]

filepattern = sprintf("*%iM.txt", numToPlotDynStatThreshold)

if(length(list.files(pattern = filepattern)) != 0)
{
  fname = list.files(pattern = filepattern)
  myCoords = read.table(file = fname, sep=';')
} else {myCoords = NULL}

for(i in 1:length(dynamicThreshold))
{
  LallGraphsDyn = calcGraph(datalist = combinedDataToPlot, time = timepoints,
                            threshold = dynamicThreshold[i])
  if(is.null(myCoords))
  {
    tempCoords = gplot(intergraph::asNetwork(LallGraphsDyn$g[[1]]),
                       mode = "circle", gmode ='graph', main = paste('Network␣at',
                       timepoints[1]), label.cex = 1, vertex.cex = 1.5)
    myCoords = tempCoords
    dev.off()
  }

  if(length(list.files(pattern = "\\.txt$")) == 0)
  {
    fname = sprintf("NodeCoordsCircle%iM.txt", numToPlotDynStatThreshold)
    write(myCoords, file = fname, sep = ";", ncolumns = 2)
  }

  dir.create(paste(today, "_Plots/", cohort, "/02_Graphs/DynThresh/", dynamicThreshold[i],
                   sep = ""))
  pathTemp = paste(pathGraphDynThres, dynamicThreshold[i],"/", sep = "")

  pbiAllRatiosDyn = LallGraphsDyn$ratios

  plotGraphs(graphList = LallGraphsDyn$g, numGraphToCombine = 5, mycoord = myCoords,
             thresh = dynamicThreshold[i], pbiAllRatios = pbiAllRatiosDyn,
             time = timepoints, filepathToStorePlots = pathTemp,
             proteinNames = mNamesAll, calcMode = 'dynamic')
}

#-------------------------------------------------------------------------------
# R Script for the analysis of unpaired proteomic data
#-------------------------------------------------------------------------------
#                                    INITIAL SETTINGS

input = read.csv("Data/SMI_peripheral.csv", header = TRUE, sep = ";")
input = input[,2:ncol(input)]

#creating folders and subfolders for plots
dir.create(paste(today, "_Plots_uBI", sep=""))
```

```
#setting filepaths
path_uBIPlots = paste("./", today, "_Plots_uBI/", sep = "")

#----------------------------------------------------------------------------
#                         PARAMETERS
allTypes = c("Ctl", "SMI")
types = c("SMI")
dynamicThreshold = c(q50 = 'q50', q75 = 'q75', q90 = 'q90')

#----------------------------------------------------------------------------
#                         DATA SORTING
dataset = outlierRemovaluBI(input, allTypes)

#----------------------------------------------------------------------------
#                      CALCULATING uBI SCORES
ubiScores = matrix("numeric", nrow = (ncol(dataset)-2), ncol = 1)

ubiScores = uBI(dataset = t(dataset[,3:ncol(dataset)]), classlabels =
                  as.vector(dataset$Type), referenceclasslabel = "Ctl")
ubiScores = ubiScores[abs(ubiScores) > 0]

#----------------------------------------------------------------------------
#                    BARPLOT DIAGRAMS OF uBI SCORES
numberToPlot = 30

png(filename = paste(path_uBIPlots, '30_barplot_uBI_SMI.png', sep = ""), width = 1000,
    height = 750)
scoresRanked = plotScores(ubiScores, numTopRankedToPlot = numberToPlot, method = 'uBI',
                          bars.cols = c("firebrick3", "dodgerblue3"), type = "SMI")
dev.off()

#----------------------------------------------------------------------------
#                          GENERAL FUNCTIONS

preprocessing = function(input, timepoints, toValidate, topProteinsFromSMI, pathGeneral)
{
  # preprocessing data: combining selected columns, removing outliers, combining data
  #                         with baseline as reference
  # Arguments:
  #   input                 raw data
  #   timepoints            vector with all time points
  #   topProteinsFromSMI    logical, if TRUE only include top 40 proteins from SMI data
  #   pathGeneral           filepath
  #
  # Return:
  #   combinedData          combined data with BL as reference

  if(topProteinsFromSMI == TRUE)
  {
    namesSMI = read.csv("RankedProteinsFromSMI.csv", header = TRUE)
    namesSMI = as.character(namesSMI[,2])

    topProteins = namesSMI[1:40]

    temp = intersect(topProteins, names(input))

    input_30 = vector('list')
    input_30 = cbind(SampleID = input$SampleID)
    for(i in 1:length(temp))
    {
      input_30 = cbind(input_30, input[names(input) == temp[i]])
    }
    input = cbind(TimePoint = input$TimePoint, input_30)
  }

  if(toValidate == TRUE)
  {
    namesToInclude = read.csv(paste(pathGeneral, "ProteinNamesFromDerivation.csv"),
                               header = TRUE, sep = ",")
    namesToInclude = namesToInclude[,2]

    temp = vector('list')
    temp = cbind(SampleID = input$SampleID)
    for(i in 1:length(namesToInclude))
    {
      temp = cbind(temp, input[names(input) == namesToInclude[i]])
    }
    input = cbind(TimePoint = input$TimePoint, temp)
  }

  sortedData = vector('list', length = length(timepoints))
  bool = vector("list", length = length(timepoints))

  for(i in 1:length(timepoints))
  {
    bool[[i]] = input$TimePoint == timepoints[i]
    sortedData[[i]] = input[bool[[i]],]
    n = nrow(sortedData[[i]])
    num = rep(i, each=n)
    sortedData[[i]]$TimePoint = num
  }
  sortedData = outlierRemoval(sortedData)
  combinedData = vector('list', length = (length(timepoints)-1))

  for(j in 1:(length(timepoints)-1))
  {
    combinedData[[j]] = rbind(sortedData[[1]], sortedData[[j+1]])
    rownames(combinedData[[j]]) = seq(1, nrow(combinedData[[j]]))
  }
  return(list(combinedData = combinedData, sortedData = sortedData))
}

#-----------------------------------------------------------------------------
outlierRemoval = function(data)
{
  # detect and replace outliers with the lower limit or the upper limit
  # Arguments:
  #   data            datalist with data of each time point
  #
  # Return:
  #   data            datalist with no outliers

  for(i in 1:length(data))
  {
    for(k in 1:(ncol(data[[i]])-2))
    {
      qnt = quantile(data[[i]][,(k+2)], probs = c(.25, .75), na.rm = TRUE)
      h = 1.5 * IQR(data[[i]][,(k+2)], na.rm = TRUE)
      data[[i]][,(k+2)][data[[i]][,(k+2)] < (qnt[1] - h)] = qnt[1] - h
      data[[i]][,(k+2)][data[[i]][,(k+2)] > (qnt[2] + h)] = qnt[2] + h
    }
  }
  return(data)
}

#-----------------------------------------------------------------------------
outlierRemovaluBI = function(dataset, types)
{
  # detect and replace outliers with the lower limit or the upper limit
  # Arguments:
  #   data            datalist with data of each time point
  #
  # Return:
  #   data            datalist with no outliers

  for(i in 1:length(types))
  {
    data = dataset[dataset[,1] == types[i],]
    for(k in 1:(ncol(data)-2))
    {
      qnt = quantile(data[,k+2], probs = c(.25, .75), na.rm = TRUE)
      h = 1.5 * IQR(data[,k+2], na.rm = TRUE)

      data[data[,k+2] < qnt[1] - h,k+2] = qnt[1] - h
      data[data[,k+2] > qnt[2] + h,k+2] = qnt[2] + h
    }
    dataset[dataset[,1] == types[i],] = data
  }
  return(dataset)
  }

#-----------------------------------------------------------------------------
calcPercentageChangeTable = function(sortedData, allTimepoints, percIncrease = 10,
                                     numToInclude = 20, plotTable = TRUE,
                                      filepath = pathPlots, width = 500, height = 500)
{
  # to check if samples come from population with normal distribution
  # Arguments:
  #   sortedData               datalist with data of each time point
  #   allTimepoints            time points of each measurement point
  #   percIncrease             minimum percentage change [%]
  #   numToInclude  number of proteins to include in table, default = 20
  #   plotTable                logical, if TRUE plot table, default = TRUE
  #   filepath                 path to save table
  #   width                    width of table plot, default = 500
  #   height                   height of table plot, default = 500
  #
  # Return:
  #   deltaIncluded        list with protein percentage changes

  delta = vector("list", length = (length(sortedData)-1))
  deltaMeans = vector("list", length = (length(sortedData)-1))
  deltaIncluded = vector("list", length = (length(sortedData)-1))
  allNames = vector("list", length = (length(sortedData)-1))
  temp = NULL

  for(i in 1:(length(sortedData)-1))
  {
    delta[[i]] = (sortedData[[i+1]][,3:ncol(sortedData[[i]])]/sortedData[[1]]
                  [,3:ncol(sortedData[[i]])])*10
    deltaMeans[[i]] = colMeans(delta[[i]])
    temp = deltaMeans[[i]] > percIncrease
    deltaIncluded[[i]] = deltaMeans[[i]][temp[TRUE]]
    deltaIncluded[[i]] = sort(deltaIncluded[[i]], decreasing = TRUE)
    allNames[[i]] = names(deltaIncluded[[i]])
  }

  uniqueNames = unique(allNames[[3]][allNames[[3]]%in%allNames[[1]]
                                      [allNames[[1]]%in%allNames[[2]]]])
  if(length(uniqueNames) < numToInclude)
  {
    numToInclude = length(uniqueNames)
  }
  for(i in 1:(length(sortedData)-1))
  {
    deltaIncluded[[i]] = deltaIncluded[[i]][uniqueNames[1:numToInclude]]
  }

  deltaMatrix = matrix("numeric", nrow = length(deltaIncluded[[1]]),
                        ncol = length(deltaIncluded))
  col_names = NULL

  for(i in 1:(length(sortedData)-1))
  {
    col_names = c(col_names, paste(allTimepoints[i+1], "/", allTimepoints[1],
                                  "␣[%]", sep = ""))
  }

  colnames(deltaMatrix) = col_names
  rownames(deltaMatrix) = names(deltaIncluded[[1]])

  for(i in 1:length(deltaIncluded))
  {
    deltaMatrix[,i] = round(deltaIncluded[[i]], 2)
  }

  if(plotTable == TRUE)
  {
    table = tableGrob(deltaMatrix)
    grid.newpage()
    h = grobHeight(table)
    w = grobWidth(table)
    main = paste("Top", numToInclude, "protein␣percentage␣changes␣after␣PMI",
                 sep = "␣")
    title = textGrob(main, y = unit(0.5,"npc") + 0.97*h,
                       vjust = 0, gp = gpar(fontsize = 13L))
    gt = gTree(children=gList(table, title))
    png(filename = paste(filepath, "/", "ProteinPercentageChangesTop", numToInclude,
                         ".png", sep = ""), width = width, height = height)
    grid.draw(gt)
    dev.off()
  }
  return(deltaIncluded)
}

#-------------------------------------------------------------------------------
pBIcalc = function(combinedData, useMedian = TRUE, lambda = 100, QuantileToInclude,
                   numberToInclude)
{
  # calculate pBI scores
  # Arguments:
  #   combinedData          datalist with each time point and reference
  #   useMedian             logical, states if median should be used, TRUE by default
  #   lambda                parameter for calculation, 100 by default
  #   QuantileToInclude     quantile in which proteins should be included
  #   numberToInclude       number of proteins which should be considered
  #
  # Return:
  #   combinedDataRanked    list of ranked proteins, included for every time point
  #   pbiScores             list of pBI scores of all proteins
  #   pbiRankedScores       list of ranked pBI scores, included for every time point
  #   pbiRankedScoresEach   list of ranked pBI scores of each time point

  column_end = ncol(combinedData[[1]])
  pbiScores = vector("list", length = length(combinedData))
  pbiRankedScores = vector("list", length = length(combinedData))
  pbiRankedScoresEach = vector("list", length = length(combinedData))
  combinedDataRanked = vector("list", length = length(combinedData))
  tempAll = NULL

  for(i in 1:length(combinedData))
  {
    pbiScores[[i]] = .pBI(dataset = t(combinedData[[i]][,3:column_end]),
                          classlabels = combinedData[[i]]$TimePoint,
                         referenceclasslabel = combinedData[[i]]$TimePoint[1],
                         ids = combinedData[[i]]$SampleID, useMedian = TRUE, lambda = 100)

    pbiTemp = pbiScores[[i]]
    quantileThreshold = quantile(abs(pbiTemp), QuantileToInclude)
    includedScores = pbiTemp[abs(pbiTemp) > quantileThreshold]

    positions = match(sort(abs(includedScores), decreasing = TRUE), abs(includedScores))
    rankedScores = includedScores[positions]
    pbiRankedScoresEach[[i]] = rankedScores

    tempAll = c(tempAll, includedScores)
  }

  scoresRanked = tempAll[match(sort(abs(tempAll), decreasing = TRUE), abs(tempAll))]
  uniqueScoresRanked = unique(names(scoresRanked))
  if(length(uniqueScoresRanked) < numberToInclude)
  {
    numberToInclude = length(uniqueScoresRanked)
  }
  proteinsToInclude = uniqueScoresRanked[1:numberToInclude]
  for(i in 1:length(combinedData))
  {
    combinedDataRanked[[i]]= cbind(combinedData[[i]][c(1,2)], combinedData[[i]]
                                  [,na.omit(proteinsToInclude)])
    pbiRankedScores[[i]]= pbiScores[[i]][na.omit(proteinsToInclude)]
  }
  return(list(combinedDataRanked = combinedDataRanked, pbiScores = pbiScores,
            pbiRankedScores = pbiRankedScores, pbiRankedScoresEach = pbiRankedScoresEach))
}

#-------------------------------------------------------------------------------
plotGraphs = function(graphList, numGraphToCombine, mycoord, thresh, time,
                      filepathToStorePlots, proteinNames, pbiAllRatios, calcMode = 'dyn')
{
  # Function to plot the graphs
  # Arguments:
  #   graphList               list containing graph elements for all measurement times
  #   numGraphToCombine       number of network graphs for the combined plot
  #   mycoord                 matrix containing the node coordinates for plotting
  #   thresh                  mector of thresholds
  #   time                    mector containing the measurement time points
  #   filepathToStorePlots    filepath for storing the network plots
  #   proteinNames            vector containg the protein names
  #   pbiAllRatios          pBI* scores
  #   calcMode              indicates if dynamic or static thresholds should be used

  combGraph = vector('list', length = 2)
  names(combGraph) = c('discrete', 'continuous')

  combGraph[[1]] = .binWeights(graphList)
  combGraph[[2]] = .dynWeights(graphList, pbiAllRatios)

  for(i in 1:length(graphList))
  {
    png(filename = paste(filepathToStorePlots, 'Network_', time[i], '_', names(thresh),
                         calcMode, '.png', sep ="␣"), width = 800, height = 750)
    gplot(intergraph::asNetwork(graphList[[i]]), mode = "circle", gmode = 'graph',
          coord = mycoord, usearrows = FALSE, main =paste('Network␣at␣', time[i],
          '␣(threshold␣', names(thresh), ')', sep = ""), label = names(V(graphList[[i]])),
          label.pos = 3, label.cex = 1, vertex.cex = 1.5, vertex.col = 'firebrick3',
          cex.main = 1.5, edge.col = 'azure3')
    dev.off()
  }
  pathAdd = c('quad', 'poly4')
  weightFac = c(2,4)

  for(j in 1:length(weightFac))
  {
    dir.create(paste(filepathToStorePlots, pathAdd[j], "/" , sep = ""))
    for(i in 1:length(combGraph))
    {
      png(filename = paste(filepathToStorePlots, pathAdd[j],'/', 'Network_',
          names(combGraph[i]), numGraphToCombine, '_graphs_', names(thresh), calcMode,
          '.png', sep = ""), width = 800, height = 750)
      gplot(intergraph::asNetwork(combGraph[[i]]), mode = "circle", gmode='graph',
            coord = mycoord, usearrows = FALSE, main = paste("Combined␣Network␣with␣",
            names(combGraph[i]), '␣weighted␣edges␣', '(threshold␣', names(thresh), ')',
            sep = ""), label = names(V(combGraph[[i]])), label.pos = 3, label.cex = 1,
            cex.main = 1.5, vertex.cex = 1.5, vertex.col = 'firebrick3', edge.lwd = 7.75*
              (E(combGraph[[i]])$weights)^weightFac[j] + 0.25, edge.col = 'azure3')
      dev.off()
    }
  }
}

#------------------------------------------------------------------------------------
.binWeights = function(graphList)
{
  # Function to calculate discrete weights for combined network plots
  # Arguments:
  #   graphList     list containing graph elements for all measurement times
  #
  # Return:
  #   gall     combined graph element with discrete weighted edges

  gall = igraph::graph.empty(n = 0, directed = FALSE)
  adjsum = 0

  for (i in 1:length(graphList))
  {
    if(length(V(gall)) == 0)
    {
      gall = graphList[[i]]
    } else {
      gall = igraph::union(gall, graphList[[i]])
    }
    adjsum = adjsum + get.adjacency(graphList[[i]], sparse = FALSE)
  }
  adjsum = adjsum/max(adjsum)

  nodeEdgeFrom = get.edgelist(gall)[,1]
  nodeEdgeTo = get.edgelist(gall)[,2]
  posx = NA
  posy = NA
  for(i in 1:length(nodeEdgeFrom))
  {
    posy[i] = which(rownames(get.adjacency(gall)) == nodeEdgeFrom[i])
    posx[i] = which(colnames(get.adjacency(gall)) == nodeEdgeTo[i])
  }
  for(i in 1:length(posx))
  {
    E(gall)$weights[i] = adjsum[posx[i],posy[i]]
  }
  return(gall)
}

#-----------------------------------------------------------------------------
.dynWeights = function(graphList, pbiAllRatios)
{
  # Function to calculate continuous weights for combined network plots
  # Arguments:
  #    graphList        list containing graph elements for all measurement times
  #    pbiAllRatios     matrix containing all pBI* scores of protein ratios
  #                       at all time points
  #
  # Return:
  #    gall     combined graph element with continous weighted edges

  weightedAdjMat = vector('list', length(graphList))
  names(weightedAdjMat) = colnames(pbiAllRatios)

  for(i in 1:ncol(pbiAllRatios))
  {
    tempGraph = graphList[[i]]
    threshold = quantile(abs(pbiAllRatios[,i]), 0.75)
    ratiosAboveThresh = abs(pbiAllRatios[which(abs(pbiAllRatios[,i]) > threshold),i])
    normRatios = vector('numeric', length = length(ratiosAboveThresh))
    names(normRatios) = names(ratiosAboveThresh)
    for(j in 1:length(ratiosAboveThresh))
    {
      normRatios[j] = ratiosAboveThresh[j]/max(ratiosAboveThresh)
    }
    adjMat = get.adjacency(tempGraph, sparse = FALSE)

    for(k in 1:length(names(normRatios)))
    {
      splitName = strsplit(names(normRatios)[k], '/')
      rowName = splitName[[1]][1]
      colName =  splitName[[1]][2]
      adjMat[rowName, colName] = normRatios[k]
      adjMat[colName, rowName] = normRatios[k]
    }
    weightedAdjMat[[i]] = adjMat
  }
  sumAdjMat = matrix(0, nrow = nrow(weightedAdjMat[[1]]), ncol=ncol(weightedAdjMat[[1]]))
  for(i in 1:length(weightedAdjMat))
  {
    sumAdjMat = sumAdjMat + weightedAdjMat[[i]]
  }
  sumAdjMat = sumAdjMat/max(sumAdjMat)
  gall = igraph::graph.empty(n = 0, directed = FALSE)

  for (i in 1:length(graphList))
  {
    if(length(V(gall)) == 0)
    {
      gall = graphList[[i]]
    } else {
      gall = igraph::union(gall, graphList[[i]])
    }
  }
  nodeEdgeFrom = get.edgelist(gall)[,1]
  nodeEdgeTo = get.edgelist(gall)[,2]
  posx = NA
  posy = NA
  for(i in 1:length(nodeEdgeFrom))
  {
    posy[i] = which(rownames(get.adjacency(gall)) == nodeEdgeFrom[i])
    posx[i] = which(colnames(get.adjacency(gall)) == nodeEdgeTo[i])
  }
  for(i in 1:length(posx))
  {
    E(gall)$weights[i] = sumAdjMat[posx[i], posy[i]]
  }
  return(gall)
}
```

```r
#------------------------------------------------------------------------------
.transferRatioVec = function(pbiall, Scores)
{
  # Function to adapt the form of the pBI* ratios for heatmap construction
  # Arguments:
  #   pbiall       list containing preprocesses data at all measurement times
  #   Scores       matrix containing score values of proteins at all times
  #
  # Return:
  #   protein ratios in adapted form

  n = ncol(pbiall[[1]][4:ncol(pbiall[[1]])])
  ProteinRatios = vector("list", n)
  pos = numeric(n)
  for (i in 1:(n-1))
  {
    pos[i+1] = pos[i]+n-i
  }
  for (j in 1:ncol(Scores))
  {
    RatioScores = Scores[,j]
    k = length(RatioScores)
        for (i in 1:length(pos))
    {
      if(i==1)
      {
        tmp = RatioScores[(pos[i]+1):pos[i+1]]
        tmpname = names(RatioScores)[(pos[i]+1):pos[i+1]]
        names(tmp) = unlist(strsplit(tmpname,"/"))[2*(1:(length(ProteinRatios)-1))]
      } else if(i==length(pos))
      {
        tmp = -RatioScores[pos[2:i]-(n-i)]
        tmpname = rownames(as.matrix(RatioScores))[pos[2:i]-(n-i)]
        names(tmp) = unlist(strsplit(tmpname,"/"))[2*(1:(length(ProteinRatios)-1))-1]
      } else {
        tmp = c(-RatioScores[pos[2:i]-(n-i)], RatioScores[(pos[i]+1):pos[i+1]])
        tmpname1 = rownames(as.matrix(RatioScores))[pos[2:i]-(n-i)]
        tmpname1 = unlist(strsplit(tmpname1,"/"))[2*(1:(i-1))-1]
        tmpname2 = rownames(as.matrix(RatioScores))[(pos[i]+1):pos[i+1]]
        tmpname2 = unlist(strsplit(tmpname2,"/"))[2*(1:(length(ProteinRatios)-i))]
        names(tmp) = c(tmpname1, tmpname2)
      }
      ProteinRatios[[i]]= cbind(ProteinRatios[[i]],as.matrix(tmp))
    }
  }
  names(ProteinRatios) = colnames(pbiall[[1]][4:ncol(pbiall[[1]])])
  return(ProteinRatios)
}
```

Bibliography

[1] Allen J. P. with an essay by D. T. Mininberg. *The Art of Medicine in Ancient Egypt.* The Metropolitan Museum of Art, 2005.

[2] Dews I. *Biomarkers are not new. In: Bleavins M. R., Carini C., Jurima-Romet M., Rahbari R.: Biomarkers in Drug Development: A Handbook of Practice, Application, and Strategy.* John Wiley & Sons, Inc., March 2010.

[3] Vasan R. S. Biomarkers of Cardiovascular Disease: Molecular Basis and Practical Considerations. *Circulation*, vol. 113, 2006.

[4] Strimbu K., Tavel J. A. What are Biomarkers? *Current Opinion in HIV and AIDS*, vol. 5, November 2011.

[5] Isaaq H. J., Veenstra T. D. *Biomarker Discovery: Study Design and Execution. In: Proteomic and Metabolomic Approaches to Biomarker Discovery.* Elsevier Inc., 2013.

[6] Aronson J. K. Biomarkers and Surrogate Endpoints. *British Hournal of Clinical Pharmacology*, vol. 59:pages 491–494, April 2005.

[7] Fleming T. R., Powers J. H. Biomarkers and Surrogate Endpoints In Clinical Trials. *Statistics in Medicine*, vol. 31:pages 2973–2984, November 2012.

[8] McDermott J. E., Wang J., Mitchell H. D., Webb-Robertson B. M., Hafen R. P., Ramey J. A., Rodland K. D. Challenges in Biomarker Discovery: Combining Expert Insights with Statistical Analysis of Complex Omics Data. *Expert Opinion on Medical Diagnostics,*, vol. 7:pages 37–51, January 2013.

[9] Aartsma-Rus A., Ferlini A., Vroom E. Biomarkers and Surrogate Endpoints in Duchenne: Meeting Report. *Neuromuscular Disorders*, August 2014.

[10] Zhou M., Conrads T. P., Veenstra T. D. Proteomics approaches to biomarker detection. *Briefings in functional genomics and proteomics*, vol. 4(no. 1):pages: 69–75, May 2005.

[11] Baumgartner C., Osl M., Netzer M., Baumgartner D. Bioinformatic-driven search for metabolic biomarkers in disease. *Journal of Clinical Bioinformatics*, January 2011.

[12] Chatterjee S. K., Zetter B. R. Cancer biomarkers: knowing the present and predicting the future. *Future Oncology*, vol. 1(no. 1), November 2005.

[13] Mattes W. B. *Biomarker Applications in the Pharmaceutical Industry. In: Goodsaid F., Mattes W. B.: The Path from Biomarker Discovery to Regulatory Qualification.* Elsevier Inc., 2013.

[14] Patel P., Kuzmanov U., Mital S. Avoiding false discovery in biomarker research. *BMC Biochemistry*, vol. 17, July 2016.

[15] Goossens N., Nakagawa S., Sun X., Hoshida Y. Cancer biomarker discovery and validation. *Translational Cancer Research*, vol. 4(no. 3), June 2015.

[16] Baumgartner C., Lewis G. D., Netzer M., Pfeifer B., Gerszten R. E. A new data mining approach for profiling and categorizing kinetic patterns of metabolic biomarkers after myocardial injury. *Bioinformatics*, vol. 26(no. 14), May 2010.

[17] Hoefkens J. Towards unbiased biomarker discovery. *Drug Discovery World*, vol. 11, Summer 2010.

[18] Lewis G. D., Wei R., Liu E., Yang E., Shi X. et al. Metabolite profiling of blood from individuals undergoing planned myocardial infarction reveals early markers of myocardial injury. *The Journal of Clinical Investigation*, vol. 118(no. 10), October 2008.

[19] Huang S. H. Supervised feature selection: A tutorial. *Artificial Intelligence Research*, vol. 4(no. 2), April 2015.

[20] Parker C. E., Borchers C. H. Mass spectrometry based biomarker discovery, verification, and validation e Quality assurance and control of protein biomarker assays. *Molecular Oncology*, vol. 8:pages 840–858, March 2014.

[21] Drabovich A. P., Pavlou M. P., Batruch I., Diamandis E. P. *Proteomic and Mass Spectrometry Technologies for Biomarker Discovery. In: Issaq H. J., Veenstra T. D.: Proteomic and Metabolomic Approaches to Biomarker Discovery.* Elsevier Inc., 2013.

[22] Ngo D., Sinha S., Shen D., Kuhn E. W., Keyes M. J. et al. Aptamer-Based Proteomic Profiling Reveals Novel Candidate Biomarkers and Pathways in Cardiovascular Disease. *Circulation*, vol. 134, July 2016.

[23] Smith J.G., Gerszten R. E. Emerging Affinity-Based Proteomic Technologies for Large-Scale Plasma Profiling in Cardiovascular Disease. *Circulation*, vol. 135, April 2017.

[24] Mythili S., Malathi N. Diagnostic markers of acute myocardial infarction. *Biomedical Reports*, vol. 3, July 2015.

[25] Chang D., Leong L. Ng. Biomarkers in acute myocardial infarction. *BMC Medicine*, vo. 8, June 2010.

[26] Jacob J., Ngo D., Finkel N., Pitts R., Gleim S. et al. Application of Large-Scale Aptamer-Based Proteomic Profiling to Planned Myocardial Infarctions. *Circulation*, vol. 137, March 2017.

[27] Netzer M., Weinberger K. M., Handler M., Seger M., Fang X., Kugler K. G., Graber A., Baumgartner C. Profiling the human response to physical exercise: a computational strategy for the identification and kinetic analysis of metabolic biomarkers. *Journal of Clinical Bioinformatics*, vol. 1(no. 1), December 2011.

[28] Spath V. A dynamic network based approach for discovering new metabolic biomarkers in myocardial injury. Master's thesis, Graz University of Technology, March 2017.

[29] Baumgartner C., Spath-Blass V., Niederkofler V., Bergmoser K., Langthaler S., Lassnig A., Baumgartner D., Asnani A., Gerszten R. E. A novel network-based approach for discovering kinetic biomarkers in cardiovascular disease using metabolic time-series data. *Bioinformatics*, 2018, in preparation.

[30] SomaLogic, Inc., 2945 Wilderness Place, Boulder, CO 80301. *SOMAmer Reagent Specificity - Technical White Paper*, 2015.

[31] SomaLogic, Inc., 2945 Wilderness Place, Boulder, CO 80301. *SOMAscan Proteomic Assay - Technical White Paper*, 2016.

[32] Li F., Zhan M. Data-driven Biomarker and Drug Discovery using Network-based Approach. *Journal of Genetics and Genome Research*, November 2015.

[33] Peela J. R., Jarari A. M., Hai A., Rawal A. K., Kolla S. D., Sreekumar S., Khurana L., Sidhanathi N. R. Cardiac Biomarkers: The Troponins and CK- MB. *Ibnosina Journal of Medicine and Biomedical Sciences*, September 2010.

[34] Stoneman V. E., Bennett M. R. Role of apoptosis in atherosclerosis and its therapeutic implications. *Clinical Science*, October 2004.

[35] Cochain C., Channon K. M., Silvestre J. Angiogenesis in the Infarcted Myocardium. *Antioxidants and Redox Signaling*, vol. 18(no. 9), March 2013.