Newcastle University

School of Computing Science

Knowledge extraction from biomedical data using machine learning

Nicola Lazzarini

Submitted for the degree of Doctor of Philosophy in the School of Computing Science, Newcastle University

June 2017

Declaration

I declare that this thesis is my own work unless otherwise stated. No part of this thesis has previously been submitted for a degree or any other qualification at Newcastle University or any other institution.

Date.....

Signature.....

“You should make something. You should bring something into the world that wasn't in the world before. It doesn't matter what it is. It doesn't matter if it's a table or a film or gardening - everyone should create. You should do something, then sit back and say: “I did that.” ”

Ricky Gervais

Abstract

Thanks to the breakthroughs in biotechnologies that have occurred during the recent years, biomedical data is accumulating at a previously unseen pace. In the field of biomedicine, decades-old statistical methods are still commonly used to analyse such data. However, the simplicity of these approaches often limits the amount of useful information that can be extracted from the data. Machine learning methods represent an important alternative due to their ability to capture complex patterns, within the data, likely missed by simpler methods.

This thesis focuses on the extraction of useful knowledge from biomedical data using machine learning. Within the biomedical context, the vast majority of machine learning applications focus their effort on the generation and validation of prediction models. Rarely the inferred models are used to discover meaningful biomedical knowledge. The work presented in this thesis goes beyond this scenario and devises new methodologies to mine machine learning models for the extraction of useful knowledge.

The thesis targets two important and challenging biomedical analytic tasks: (1) the inference of biological networks and (2) the discovery of biomarkers. The first task aims to identify associations between different biological entities, while the second one tries to discover sets of variables that are relevant for specific biomedical conditions. Successful solutions for both problems rely on the ability to recognise complex interactions within the data, hence the use of multivariate machine learning methods. The network inference problem is addressed with FuNeL: a protocol to generate networks based on the analysis of rule-based machine learning models. The second task, the biomarker discovery, is studied with RGIFE, a heuristic that exploits the information extracted from machine learning models to guide its search for minimal subsets of variables.

The extensive analysis conducted for this dissertation shows that the networks inferred with FuNeL capture relevant knowledge complementary to that extracted by standard inference methods. Furthermore, the associations defined by FuNeL are discovered

more pertinent in a disease context. The biomarkers selected by RGIFE are found to be disease-relevant and to have a high predictive power. When applied to osteoarthritis data, RGIFE confirmed the importance of previously identified biomarkers, whilst also extracting novel biomarkers with possible future clinical applications.

Overall, the thesis shows new effective methods to leverage the information, often remaining buried, encapsulated within machine learning models and discover useful biomedical knowledge.

Publications

Part of the work within this thesis have been documented in the following publications:

J. Bacardit, P. Widera, **N. Lazzarini**, and N. Krasnogor. "Hard data analytics problems make for better data analysis algorithms: bioinformatics as an example" *Big data*, vol. 2, no. 3, pp. 164-176, sep 2014.

F Eduati et. al. "Prediction of human population responses to toxic compounds by a collaborative competition". *Nature Biotechnology*, vol. 33, pp. 933-940, aug 2015.

N. Lazzarini, P. Widera, S. Williamson, R. Heer, N. Krasnogor, and J. Bacardit. "Functional networks inference from rule-based machine learning models" *BioData Mining*, vol. 9, sep 2016.

S. Baron, **N. Lazzarini** and J. Bacardit. "Characterising the influence of rule-based knowledge representations in biological knowledge extraction from transcriptomics data" *EvoBio 2017, Evolutionary Computation, Machine Learning and Data Mining for Biology and Medicine*, in press, apr 2017.

N. Lazzarini and J. Bacardit. "RGIFE: a ranked guided iterative feature elimination heuristic for the identification of biomarkers" *BMC Bioinformatics*, in press 2017.

Under review:

N. Lazzarini, J. Runhaar, A.C. Bay-Jensen, C.S. Thudium, S.M.A. Bierma-Zeinstra, Y.Henrotin and J. Bacardit. "A machine learning approach for the identification of reduced panels of biomarkers and its application to knee osteoarthritis incidence in overweight and obese women" *Osteoarthritis and Cartilage*

Acknowledgements

First and foremost I wish to thank my supervisors, Dr. Jaume Bacardit and Prof. Natalio Krasnogor for giving me the chance to pursue a Ph.D. at Newcastle University. Your advice and guidance during these years have been truly invaluable and helped me in enhancing the quality of my work.

A special thank goes to Dr. Paweł Widera. Your support has been fundamental during my Ph.D. adventure since the first day. All your help, advice and suggestions made me a better scientist.

I would also like to thank all the colleagues from the ICOS (Interdisciplinary Computing and Complex BioSystems) research group. I am grateful to the people in 922 that had to hear me ranting and complaining during the last four years. They helped me to see the brighter side of things and always gave me hope. I was lucky to complete my Ph.D. together with supporting people, especially Dr. Jurek Kozyra, Dr. Annunziata Lopiccolo and Dr. Göksel Misirli. Finally, I need to mention Fedor Shmarov, Jens Geyti and Dr. Joseph Mullen that made my Ph.D. experience a bit lighter.

I wish to thank the School of Computing Science and Newcastle University. Also, the Engineering and Physical Sciences Research Council (EPSRC) that supported my work [EP/L001489/2, EP/J004111/2, EP/I031642/2, EP/N031962/1]. Furthermore, I am thankful to the European Union Seventh Framework Programme (FP7/2007-2013) that funded part of this work under the "D-BOARD" project (grant agreement number 305815). I am also thankful to my examiners Prof. John H. Holmes and Prof. Anil Wipat for their time dedicated to my dissertation and for the valuable feedback received.

Last but not least, a very special thanks must go to my parents and my brother. Your continuous support and encouragement helped me towards my whole decennial academic path until the finish line. Thank you.

Contents

List of Figures

List of Tables

Acronyms

ACR American College of Rheumatology

AUC Area Under the ROC Curve

DB-SCV Distributed Balanced-Stratified Cross Validation

G-D Gene-Disease association

GO Gene Ontology

GNB Gaussian Naive Bayes

JSN Joint Space Narrowing

KL Kellgren & Lawrence

KNN K-nearest neighbour

LOOCV Leave One Out Cross Validation

MIC Maximal Information Coefficient

OA Osteoarthritis

PCA Principal Component Analysis

PCC Pearson Correlation Coefficient

PPI Protein Protein Interaction

SPL Shortest Path Length

SVM Support Vector Machine

RF Random Forest

ROC Receiver Operating Characteristics

1

Introduction

Contents

1.1 Background and motivation

During the last few decades, the advances in high-throughput technologies have led to an explosion in the availability of biomedical data, which subsequently increased the understanding of how those data can be used to improve human life. The analysis of such a large amount of data can help us in revealing and explaining the complex mechanisms that characterise biological and medical conditions. However, this goal can only be achieved if appropriate analytical tools are designed to fully exploit the large quantity of available information and extract relevant knowledge.

Statistical-based and computational methodologies have been extensively applied for data analysis in the field of biomedicine, trying to underline difficult biological and medical processes [1–3]. However, due to the simplicity of these approaches (e.g. linear models or univariate techniques), the amount and the kind of information that can be extracted from the data is limited [4]. Machine learning represents a powerful alternative that can offer better, more robust and flexible solutions and is currently rising in the field of biomedicine [5]. The advantageous position of machine learning methods is given by the use of complex multivariate knowledge representations that allow, when mining the data, to discover interesting patterns that are often missed by simpler approaches. Thanks to such a rich and diverse knowledge representation, machine learning approaches are well suited for the analysis of biomedical data that often are characterised by: large dimensionality (high number of variables), class imbalance distribution (e.g. many more healthy patients than sick), vast number of samples, information collected from different sources (e.g. clinical examination, gene expression levels, protein abundances), etc. Hence, over the years, the use of machine learning methods has proven successful in many different biosciences: medicine [6], biology [7], chemistry [8], etc.

Machine learning is defined as the set of methods that automatically learn from experience [9]. Machine learning algorithms analyse the data and generate solutions (models) to address a large variety of complex problems. Now, once the model is inferred, if we understand why the algorithm performed certain choices or we interpret the structure of the solutions, we have the possibility to learn something. The experienced gained

by the algorithms when analysing the data can help us in improving the understanding of biomedical questions. For example, the presence of certain features or the relationships between specific components of the computational models are information that can reveal new unexpected insights. The challenges in identifying and extract this information provide the primary motivation for this thesis. Inspired by the possibility to gain new interesting and relevant knowledge, novel methodologies are presented for the mining of machine learning models generated from biomedical data.

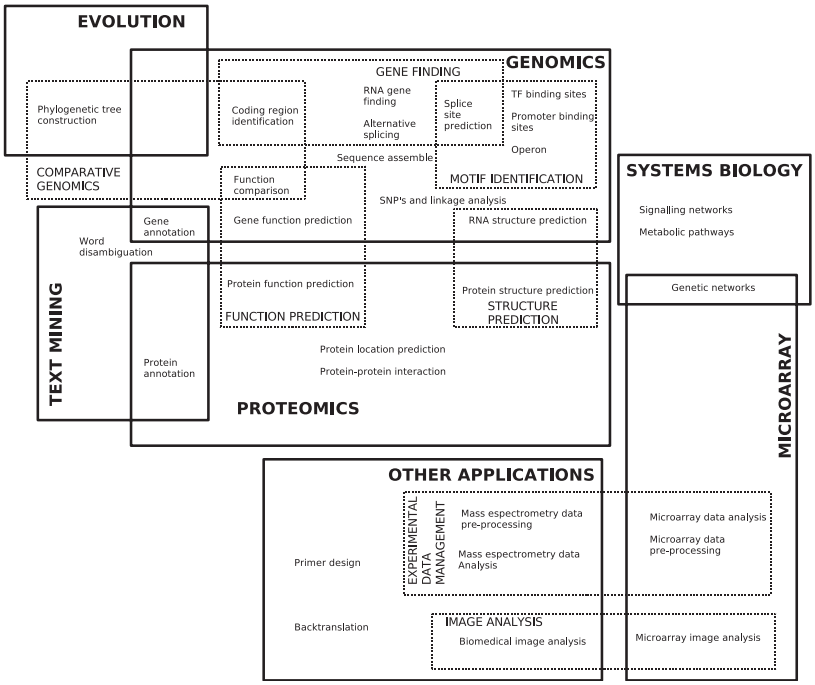


Fig 1.1: Example of the possible applications of machine learning in biomedicine as illustrated in [10].

In biomedicine, machine learning methods have been used to solve many different tasks [10]. As presented in Figure 1.1, machine learning applications can be useful within diverse domains: gene network inference, microarray analysis, pathways investigation, protein function prediction, phylogenetic tree construction, etc. Among them, this thesis focuses on proposing new methodologies that can tackle two main analytic tasks: (1) the inference of biological networks and (2) the discovery of biomarkers (short for biological markers, that is a measure of a biological state). Both are highly relevant and challenging tasks in the field of biomedicine [11, 12]. Their successful resolution requires the ability to capture and exploit the relationships between the entities of

data. Machine learning, with its rich knowledge representations, can contribute in proposing alternative solutions to what is currently existing.

As a consequence of the growth of public biomedical data, the scientific community has been able to disprove theories and beliefs formulated in the past. For example, in 1941 Beadle and Tatum proposed the "one-gene/one-enzyme/one-function" paradigm [13]. Over the years, with the improvement of technologies and the analysis of relevant data, we learnt how the picture is far more complex. It has been established that biological processes and diseases are rarely caused by a single molecule, but they are instead the result of many interactions between several factors. The complicated mechanisms behind those biological processes and diseases can be modelled by *complex networks* to facilitate their comprehension. When coupled with large-scale data, networks have been proven to provide a useful conceptual framework. Machine learning, with its ability to discover hidden and relevant patterns, can contribute towards filling the gap created by traditional methods based on simple and sometimes limiting approaches.

The large amount of information associated with biomedical data motivates the other research task tackled in this thesis. Modern high-throughput experiments allow the analysis of the relationships and the properties of many biological entities at once. Therefore, the observations included in the data result defined in a high dimensional space. Unfortunately, a vast abundance of irrelevant and sometimes misleading data is encapsulated within those dimensions. Therefore, there is a need for adequate computational approaches to recognise and filter out insignificant information. The identification of factors that are important, and potentially can drive a specific condition or disease, assumes the name of *biomarkers discovery*. Machine learning methods in this context become important, as they can efficiently mine the data and, taking into account possible dependencies among the variables, discard irrelevant information.

1.2 Overview of the problem

When coming to the use of machine learning in biomedicine, most of the research effort tends to focus exclusively on the core data mining tasks of building and applying models [14]. Typical examples in which machine learning is employed are: classification

problems where the goal is the discrimination of patients that belong to different categories such as controls vs. cases [15], regression problems where the aim is to predict the values of a continuous variable such as the chemical level of a compound [16], clustering problems where different samples (patients) need to be grouped together based on common characteristics [17], etc. Quite often the success of the proposed solution is purely based on performance metrics such as the classification accuracy (i.e. how many patients can the model correctly classify?). Far less interest and effort have gone into the knowledge discovery and the hypothesis generation from the analysis of the data. In this context, machine learning models are simply treated as a "black box" that, given some data as input provide somehow a "magical" solution as output.

The difficulty in interpreting the machine learning solutions is generating a gap with the bioscience experts and is preventing a wider adoption of machine learning techniques. Currently, the proposed methods do not always entirely fulfil the needs and the expectations of the bioscientists. As mentioned earlier, mere "black box" solutions are not enough anymore. For example, an oncologist is not interested in a model that can only slightly outperform his ability in identifying cancer cells. On the contrary, he would be fascinated to discover how the classifier recognises cancer cells and which criteria it uses to discriminate them from healthy cells. Machine learning models, if exploited with appropriate techniques, have the potential to fulfil the expressed needs. Thus, as it is currently a common practice, the usage of machine learning narrowed to solve core data mining tasks (e.g. predict the category of the samples) is limiting the advance in the understanding of many biomedical problems, far more can be achieved.

Besides, generic computational algorithms, including machine learning, not always provide the best solution for biomedical problems. Sometimes they cannot adapt to better address the problem in hand. For example, in biomarker discovery, the number of candidates is crucial as the fewer they are, the more likely is to have them experimentally validated. Generic machine learning methods which simply aim to maximise the predictive performance of the candidate sets, regardless their size, are not always the best choice. Thus, in this context, better methods are needed. For instance, an algorithm that can trade small drops in predictive performance in favour of a smaller set of biomarker candidates, so that is more likely to have them experimentally tested. Fur-

thermore, many methods, based on specific types of knowledge representation, might be able only to capture a limited kind of information. An intrinsic bias is associated to each knowledge representation [18], this narrows down the overall information that each method can extract. Therefore, flexible approaches that can use different knowledge representations and ideally, can identify the best type based on the data being analysed, can improve the provided solution.

Overall, there is an increasing need for methodologies that are designed to solve biomedical problems and can bridge the gap between the generation of computational models and their interpretability for the gaining of new research insights. The work proposed in this thesis is intended to fill this gap and tackle the mentioned problems.

1.3 Aims and scope

Overall, this dissertation tries to verify the following research hypothesis:

Research hypothesis

Can we extract relevant knowledge from the analysis of machine learning models generated from biomedical data?

To test this research hypothesis, the thesis concentrates on the mining and the analysis of the structure of various machine learning models generated from different biomedical data. The aim is to move a step further from the inference of computational models and verify whether their structure can be used to discover new knowledge. Using the DIKW (Data, Information, Knowledge, Wisdom) model [19] as a reference, represented in Figure 1.2, a typical application of machine learning methods would stop at the model generation (*information*). Conversely, the research performed through this dissertation, using biomedical *data*, explores the output of the model generation step (*information*) to discover new insights (*knowledge*) that potentially can help to understand complex biomedical problems better (*wisdom*).

The general process used to extract knowledge from biomedical data is depicted in Figure 1.3. As evident from the figure, the process of exploiting machine learning models can result in different outputs: from biological networks to the identification

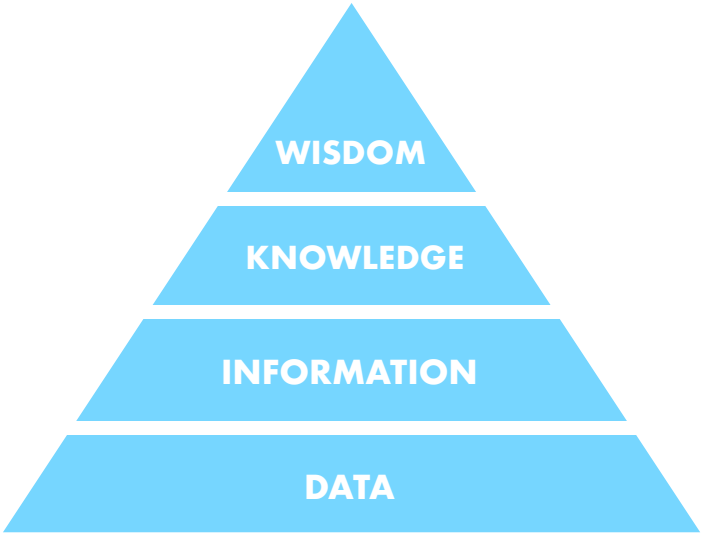


Fig 1.2: The DIKW (Data, Information, Knowledge, Wisdom) pyramid model.

of biomarkers, from patient stratification to phylogenetic tree inference, etc. For example, by checking which elements characterise the generated models and how they are used to perform computational tasks, it is possible to deduce if biological entities interact with each other or if they are evolutionarily related. Following the steps illustrated in Figure 1.3, this dissertation proposes solutions for two important biomedical problems: (1) the inference of biological networks and (2) the identification of small sets of predictive biomarkers. The presented methods aim to: discover relevant knowledge and be generic, that is not tailored to handle a specific type of data but instead capable of dealing with a wide variety of biomedical data.

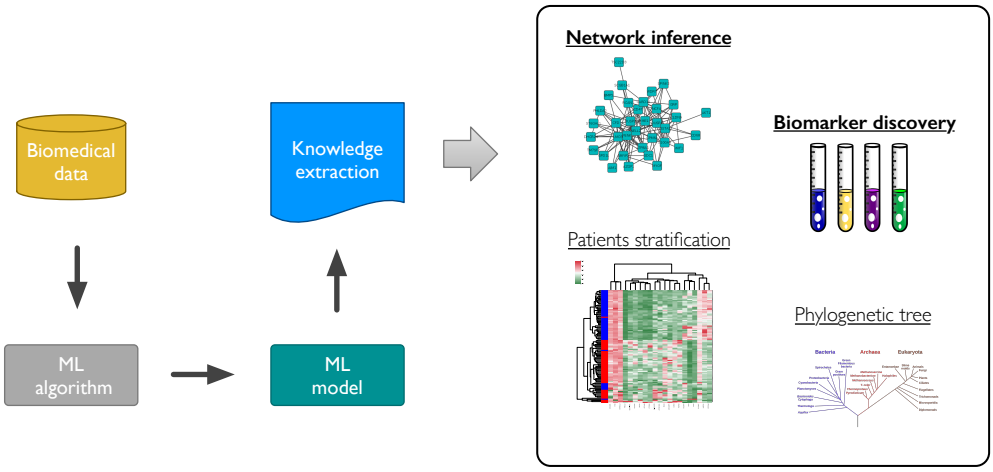


Fig 1.3: The process of knowledge extraction through the analysis of machine learning (ML) models generated from biomedical data. In bold are highlighted the research topics on which this thesis is focused.

1.4 Thesis Structure

This thesis is organised into six chapters: the introduction, a preliminary chapter that presents the context and the background material on which the dissertation is built on, three content chapters that describe the contributions of this work to the bio-data mining field and one final chapter that underlines the conclusions and further work. The overall structure of the thesis, excluding this introductory chapter, is the following:

Chapter 2 - Background research introduces the key concepts necessary to understand the content of the dissertation fully. The different types of biological data are described and is given an introduction to machine learning. This is followed by the presentation of the state-of-the-art approaches employed for (1) the inference of biological networks and (2) the identification of biomarkers. Afterwards, is included a description of the common approaches employed for the biomedical validation of the knowledge extraction process.

Chapter 3 - FuNeL: a protocol for the inference of functional networks from machine learning models describes FuNeL, a protocol for the inference of functional networks generated from rule-based machine learning models. The chapter provides an extensive analysis of the networks inferred with FuNeL using both synthetic and real-world data. In addition, FuNeL is contrasted with the state-of-the-art methods for network inference. The comparison is performed from both a biomedical and a topological point of view.

Chapter 4 - RGIFE: a ranked guided iterative feature elimination heuristic for biomarkers identification introduces, improves and evaluates RGIFE: a heuristic for the identification of small sets of biomarkers. The analysis consists of a thorough validation of the new features implemented in the heuristic. Furthermore, RGIFE is contrasted with classic methods used for biomarker discovery employing both real-world and synthetic data. The comparison is done in terms of predictive performance and biomedical relevance of the selected biomarkers.

Chapter 5 - Identification of biomarkers for knee osteoarthritis describes the application of machine learning techniques to a variety of biomedical data (from

lipids abundance to clinical measurements) obtained from a knee osteoarthritis study. A machine learning-based pipeline, using RGIFE at its core, is used to generate predictive models for the presence of knee osteoarthritis. The proposed models are extensively analysed and contrasted with literature findings. In addition, FuNeL is used to infer networks from a subset of the available data.

Chapter 6 - Conclusions summarises the results and the main findings of the dissertation. The chapter also includes a discussion of the limitations of the proposed methodologies and possible future work.

1.5 Main contributions

The main contribution of this dissertation is the introduction of new methods for the discovery of relevant knowledge, from machine learning models, for biomedicine. The research performed for this thesis resulted in the:

- definition and evaluation of a protocol, called *FuNeL*, for the inference of functional networks from the analysis of rule-based machine learning models, in Chapter 3
- characterisation of a systematic approach to evaluate biological networks based on biomedical knowledge (gene-disease associations), in Chapter 3
- computational improvements and biomedical validation of *RGIFE*, a heuristic for the identification of small sets of biomarkers guided, in its search for the optimal solutions, by the information extracted from machine learning models, in Chapter 4
- identification of knee osteoarthritis biomarkers via the application of machine learning methods to biomedical data and their characterisation in a network context, in Chapter 5.

2

Background Research

Contents

Abstract

This chapter introduces the main concepts behind each step of the knowledge extraction process employed in this dissertation. The chapter covers the data and the methodologies used to generate both models and hypothesis. Afterwards, the approaches available to evaluate the extracted knowledge are presented. More in details, the chapter offers an overview of the type of biological data that can be generated nowadays. In addition, the state-of-the-art approaches for both the inference of biological networks and the discovery biomarkers are described. Finally, the chapter provides an introduction to the methodologies commonly employed to validate the output of the knowledge extraction process in biomedicine.

2.1 Types of biological data

In 1958, Francis Crick proposed a concept that is believed to provide the underpinning of all biology: *the central dogma of molecular biology* [20]. The dogma describes the flow of genetic information within a biological system. A simplified version of the dogma is illustrated in Figure 2.1. The DNA is transcribed into RNA strands, messenger RNA strands are then translated into proteins that are virtually involved in all the cell functions. Current technologies can provide measurements made on different tiers of the central dogma and beyond. Those measurements lead to the generation of the so called *-omics* data. The suffix -omics refers to the collective technologies used to explore the roles, the relationships and the actions of the various types of molecules that make up the cellular activity of an organism.

As suggested in [22], -omics fields can be grouped as:

- **Genomics**: probably represents the most mature of the different -omics fields. It is defined as the study of the whole genome sequence (the complete DNA sequence of an organism's genome) and the information contained therein. A GWAS, also known as Genome Wide Association Study, provides an examination of a genome-wide set of genetic variants in different individuals to check if any variant is associated with a trait. GWASs mainly focus on the association

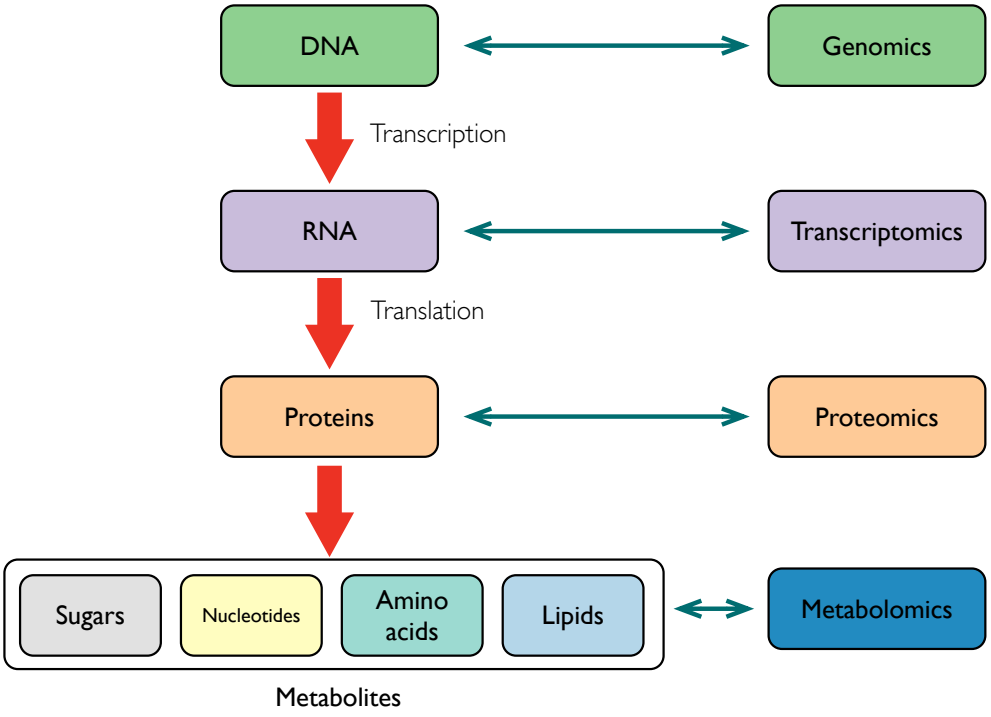


Fig 2.1: The central dogma of molecular biology and the connections with the type of -omics data obtained from each tier (based on diagram by Lmaps [21]).

between SNPs (Single Nucleotide Polymorphism) and human traits. A SNP is a variation at a single DNA site, they are the most frequent type of variation that can be found in the genome and they have been extensively studied to identify diseases susceptibility and for assessing the efficacy of drug therapies.

- **Transcriptomics**: contain information about both the presence and the relative abundance of RNA transcripts, by that illustrating the active components within the cell. Microarrays are the most well-established approaches and have been extensively used in many fields of bioinformatics over the years.

- **Proteomics**: identify and quantify the cellular levels of each protein being encoded by the genome. Proteomics data can be used for different purposes such as: biomarkers discovery, analysis of functional pathways and quantification of proteins [23].

- **Metabolomics**: seek to analyse the set of metabolites (also known as the metabolome) of the cell. The metabolome is the output that results from the cellular integration of the transcriptome and proteome, so it offers both a list

of metabolite components and functional readout of the cellular state. Among the metabolomics data, lipidomics are recently receiving much interest, they have been found to major an important role in many metabolic diseases such as obesity, atherosclerosis, stroke, hypertension and diabetes [24].

The analysis performed for this dissertation involved only the use of transcriptomics and lipidomics data. However, the methodologies presented are generic enough to be applied to other types of biological data. Prior to every kind of analysis on -omics data, several pre-processing steps need to be performed (e.g. background correction, normalisation, summarisation, etc.) Different types of -omics data require different pre-processing approaches [25]. All the data used for this dissertation were either taken from public repositories or provided by clinicians. In both cases, the data were already pre-processed, so there will be no mention of such techniques in this dissertation.

2.2 Introduction to machine learning

2.2.1 Machine learning paradigms

Many different definitions have been proposed, over the years, for the term *machine learning*. In 1959 Arthur Samuel [26] stated that:

> “Machine learning is the subfield of computer science that gives computers the ability to learn without being explicitly programmed”

It means that machine learning algorithms are able to perform a specific task without being directly told how to do it. Let’s assume we would like to create a program that can distinguish between spam and valid email messages. We can define a set of rules that highlights the messages that contain certain features such as specific words (e.g. viagra) or explicitly fake adverts. Unfortunately, the generation of an efficient set of rules can be difficult because spammers tend to use strategies to avoid spam filters (e.g vi@gr@ instead of viagra). In this context, machine learning is the solution because, given a set of manually labelled good and bad email examples, an algorithm can automatically learn a set of rules that distinguish them.

Another famous machine learning definition was proposed by Tom Mitchell [9]:

> “Machine Learning as the set of computer algorithms that automatically learn from experience”

Following this definition, we can define the learning as:

Definition 2.2.1. *A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.*

According to this definition, we can reformulate the email problem as the task of identifying spam messages (task T) using the data of previously labelled email messages (experience E) through a machine learning algorithm with the goal of improving the future email spam labelling (measure P).

According to how E, P and T are defined, we can identify different machine learning paradigms. A classical division of the learning paradigms includes:

- **Supervised learning** is defined as a learning process where the system is guided (either automatically or by human interaction) and receives feedback about the correctness of its performance. In this type of paradigm, the performance measure P allows the system to improve its learning process continuously.
- **Unsupervised learning** is characterised by the absence of the performance feedback P. The machine learning system needs to infer the hidden structure of the data without any information about the potential solution. It is important, for the learning system, to avoid the regularities existing in E in order to generate a well-performing solution.
- **Semi-supervised learning** is a middle point between the two previous paradigms where some of the input data are labelled, while some are not.
- **Reinforcement learning** is a paradigm where the system receives an indirect feedback about the appropriateness of its response. Different than in supervised learning, in reinforcement learning the system only knows that the behaviour was inappropriate and (usually) how inappropriate it was.

The next sections will provide more detailed information about the two most used types of learning: supervised and unsupervised. However, the work presented in this thesis only employs supervised learning approaches. Therefore, most of the focus of the chapter will be on this paradigm.

2.2.2 Supervised learning

In supervised learning, the system receives feedback about the correctness of its solution using the information available in the data. More specifically, in supervised learning, the system tries to solve a problem known as *classification*. The next sections will describe the classification problem and the different types of knowledge representations that can be used to solve it.

2.2.2.1 The classification problem

In machine learning *classification* is defined as the problem of identifying the category to which a new observation belongs based on the similarities with previously analysed data, for which, the category membership is known. A more formal definition of classification is:

Definition 2.2.2. *Given a set of data points $X = \{x_1, ..., x_n\}$, each of them belonging to a finite set of classes $Y = \{y_1, ..., y_m\}$, the task of classification is to generate a function $f : X \rightarrow Y$ which maps elements of X to Y.*

Each data point x_i is commonly called *instance* (or sample) and is characterised by a finite set of features $F = \{f_1, ..., f_l\}$ that can be either categorical or numerical. Often, the *features* are known as *attributes* or *variables*, in this dissertation the three names will be used interchangeably. Each data point x_i is also associated with a label y_i which indicates its class from a finite set Y. The goal is to define a model that given a data point x_i can determine its label y_i.

The classification process is summarised in Figure 2.2, it can be split into two phases: *model construction* and *model usage*. In the first phase, given a set of data representing a target concept, the goal is to build a model that can "explain" the concept. The next phase consists in using the inferred model to classify future unlabelled samples. It is crucial to generate a system able to model the concept represented by the instances

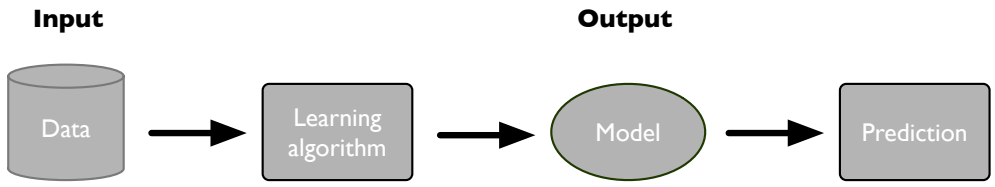


Fig 2.2: Classification as the task of generating a model to map the input features into class labels.

rather than just reproduce the instances themselves. If the system has not been able to capture and generalise the concept of the data, there will be a large generalisation error, that is the future unseen instances will likely be incorrectly classified. However, when developing a learning system, future instances are not available; therefore it is necessary to simulate the model usage phase. By simulating the future behaviour of the model it is possible to sense whether the learning part was successful and we can estimate the future generalisation error rate. The simulation can be done by splitting the available data into two non-overlapping sets called: *training* and *test set*. First, the model is generated by learning from the training set, then the test set is used to assess if the concept represented by the input data was correctly identified. If the learning algorithm has inferred an accurate model, then the instances of the test set will be correctly classified. An overview of the model usage simulation is illustrated in Figure 2.3.

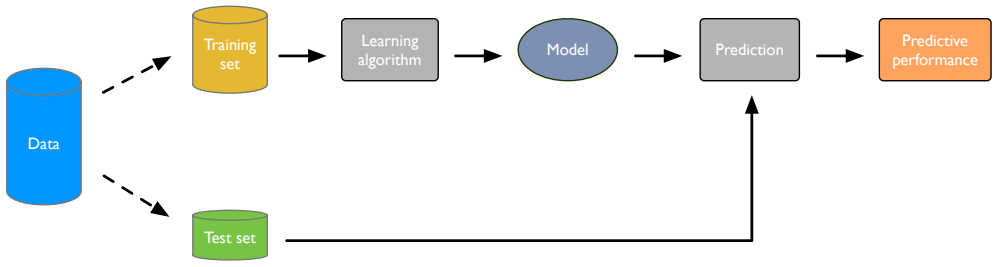


Fig 2.3: The general approach to build and validate a predictive model.

The validation of a model consists of assessing how well the labels of the test instances (unseen during the learning phase) can be predicted. For binary classification problems, where the samples belong to **only** two classes (positive and negative), the performance of a model are commonly visualised using 2×2 table called *confusion matrix* (or contingency table). In biomedicine and bioinformatics, the positive class usually represents individuals affected by a medical condition (case) or treated with a

drug, while the negatives represent the controls or healthy patients. As can be seen in Table 2.1, the confusion matrix summarises the correct and incorrect prediction for each class.

		Real class	
		Positive	**Negative**
Predicted class	**Positive**	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

Table 2.1: Example of a confusion matrix

A variety of metrics is defined from the confusion matrix. Each performance metric might be more or less informative based on the task T that the classifier is expected to solve. Some of the most common metrics are:

- **accuracy**: is probably the most simple and adopted metric. It is defined as the rate of the correctly classified instances over the total number of instances in the test set:
$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
- **balanced accuracy**: can be used in the presence of imbalanced datasets, where the samples of one class outnumber the samples of the other one. It equally weights the correct number of classified instances for each class:
$$balanced\,accuracy = \frac{1}{2}\left(\frac{TP}{TP + FP} + \frac{TN}{TN + FN}\right)$$
- **sensitivity, specificity**: they respectively measure the proportion of positives and negatives that are recognised as such. The sensitivity is also known as *recall*:
$$sensitivity\,(recall) = \frac{TP}{TP + FN} \quad specificity = \frac{TN}{TN + FP}$$
- **Gmean**: is the geometric mean of *sensitivity* and *specificity* [27]. It is commonly used when the performance of both classes are expected to be considered:
$$gmean = \sqrt{sensitivity \times specificity}$$

- **Precision**: is also called positive predictive value (PPV) and calculates the ability of not labelling as positive a sample that is negative:

$$precision = \frac{TP}{TP + FP}$$

Many other metrics (e.g. F1-score, Matthews correlation coefficient, etc.) exist to assess the predictive performance of machine learning models. In this section, only the metrics most relevant to this dissertation have been listed.

In a clinical context, the performance of a model is often evaluated via the analysis of the Receiver Operating Characteristics (ROC) curve. A ROC curve is a plot that illustrates the performance of the model based on the true and the false positive rate [28], an example of ROC curve is provided in Figure 2.4. The true positive rate is equivalent to the *sensitivity* and represents the ratio of positive instances that are correctly classified (e.g. percentage of sick people that have been accurately recognised as a case). The false positive rate indicates the proportion of negative samples that are incorrectly labelled as cases (e.g. percentage of healthy people diagnosed with a disease). A ROC curve can be generated **only** when the classifier can compute a "score" (real value) for each instance. This score, typically in the range between 0 and 1, is often intended to indicate the probability that an instance has to belong to the a specific class. The ROC curve is plotted by varying the threshold setting at which the instances are assigned to a specific class. For example, if the threshold is set to 0.2, all the samples that received a score (predicted output), equal or higher than 0.2 are predicted as positive. Once all the test samples have been classified, the *sensitivity* and *specificity* values are calculated and added as data points in the space of the true and false positive rate. ROC curves are typically used to determine the threshold that best suits the goal of the research question (e.g. at which value the sensitivity is maximised while having at least a false positive rate of 0.7). The ROC curve can also be summarised into a single value by calculating the area under it, called Area Under the ROC Curve (AUC). The AUC represents the probability that the classifier will rank a positive instance, randomly picked, higher than a randomly selected negative one (when assuming that positive samples rank higher than negatives). For a binary classification problem, a perfect classifier generates an AUC of 1, while a random

classifier (that assigns with 50% chance one of the two class labels) obtains an AUC of 0.5. Every model providing an AUC lower than 0.5 is considered to perform worse than a random one. Similar to the ROC curve, the Precision-Recall (PR) curve shows the performance of a classifier based on precision and recall. The AUPRC (area under the PR curve) summarises, as the AUC, the performance with a single value that ranges between 0 and 1. The PR curve represents a valid alternative to the ROC curve, that, on the other hand, is widely used and adopted in the biomedical field.

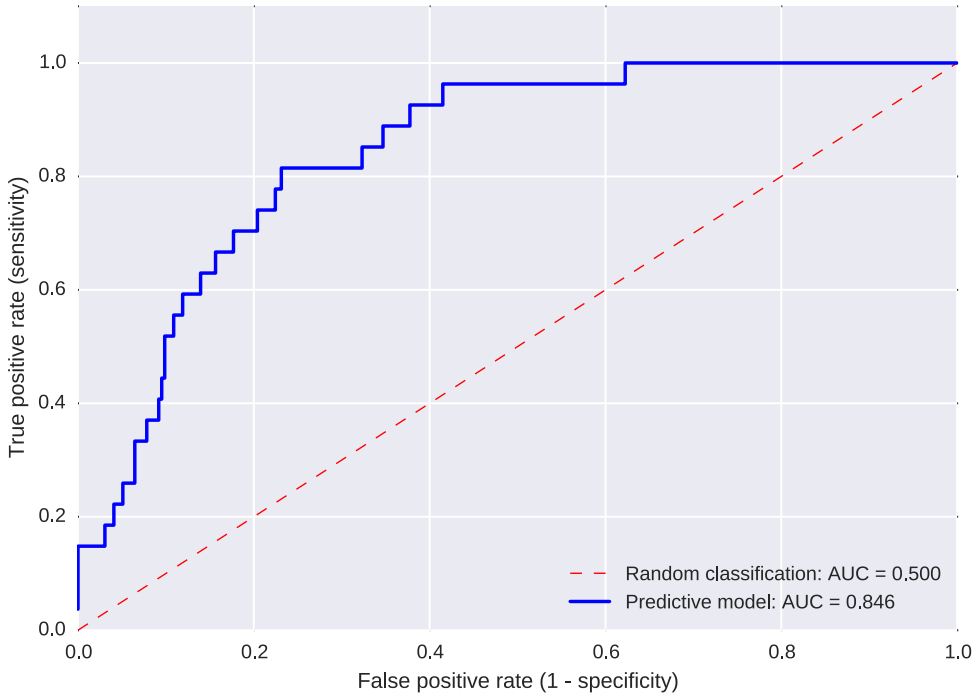


Fig 2.4: Example of a ROC curve summarising the performance of a predictive model.

Data are usually heterogeneous, therefore when dividing, often randomly, the instances into training and test set, one of the two sets might not properly represent the concept of the data. A standard approach used to reduce the bias of the data being split into training and test set is the *cross-validation*. A typical n-fold cross-validation scheme randomly divides the dataset D in n equally-sized disjoint subsets $D_1, D_2, ..., D_n$. In turn, each fold is used as test set while the remaining $n-1$ are used as training set. A *stratified* cross-validation aims to partition the dataset into folds where the original distribution of the classes is preserved [29]. The drawback of the stratified cross-validation is that it does not take into account the presence of clusters (similar samples) within each class. As observed in [30], this might lead to a distorted measure of the perfor-

mance. When dealing with datasets having a small number of observations, as typical in a biomedical context, such distortion in performances can be amplified. To solve this problem, Zeng and Martinez proposed the Distributed Balanced-Stratified Cross Validation (DB-SCV) scheme [31]. The DB-SCV is designed to assign close-by samples to different folds so that each fold will end up with enough representatives of every possible cluster. When n is equal to the total number of samples, the cross-validation is known as Leave One Out Cross Validation (LOOCV). Each instance is in turn used as test set while all the remaining are used for the training phase. Two reasons make the leave-one-out attractive, first it maximises the number of samples used for the training phase and therefore increases the chance to have and accurate model. Secondly, it does not involve random sampling (bias) as the procedure is deterministic (only one way to divide the dataset with a LOOCV) and there is no need in repeating it multiple times. On the other hand, it has been demonstrated that such approach tends to overestimate the performance of the models [30], mainly because no stratification can be applied. Nevertheless, when dealing with small datasets having few samples, perhaps concentrated in one class, the leave-one-out is one of the few available options. Regardless the type of cross-validation chosen, the overall performance of the model are assessed by averaging the performance values obtained in each test set. Overall, with the cross-validation, it is possible to simulate the model usage and better estimate how the learning algorithm was able to generalise the concept represented by the input data. In addition, this process provides a hint on how the model will perform when dealing with future instances.

2.2.2.2 The knowledge representation in supervised learning

The learning can also be classified according to the knowledge representation used to reproduce the output [32]. Russell and Norvig [33] stated that:

> “The object of the knowledge representation is to express knowledge in computer-tractable form, such that it can be used to help agents perform well.”

The main different knowledge representations that can be found in the supervised learning are:

- **Decision Trees** are tree-like graphs that define a series of questions about the attributes to predict the label of the data samples. An example of a decision tree, for the classification of stroke risk (low or high), can be seen in Figure 2.5. Each node of the tree divides the instances according to a test over an attribute; the leaves correspond to the final predicted label. When decision trees are used to predict numeric values they are called regression trees. C4.5 is in absolute the most representative algorithm based on decision trees [34].

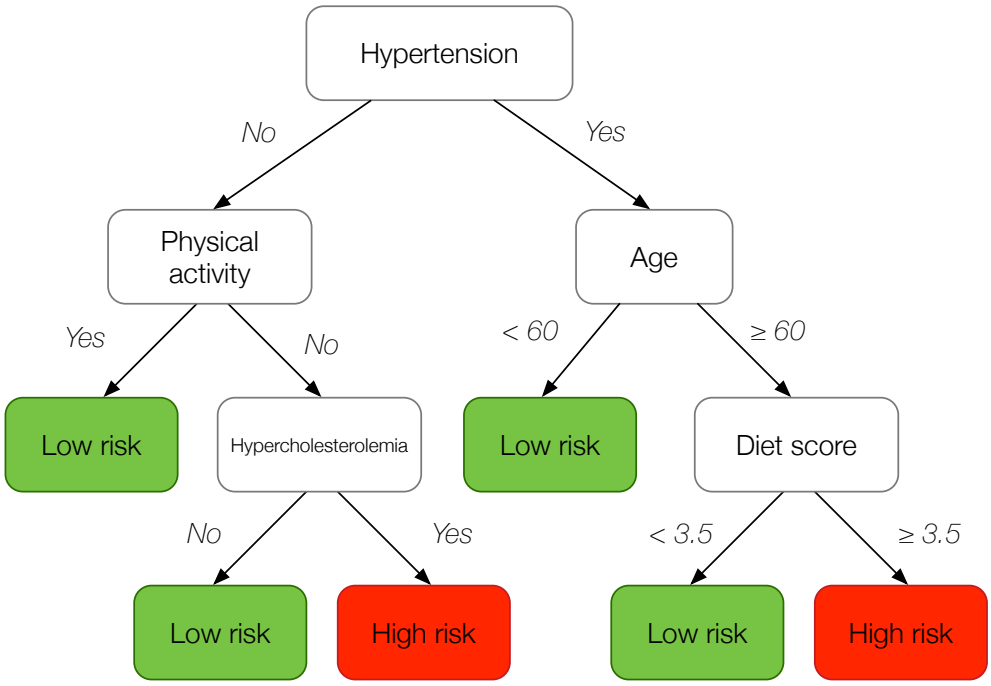


Fig 2.5: Example of a decision tree for a stroke risk classification problem.

- **Classification rules** consist of a series of rules that assign each instance to a class if a condition is met. The rule is usually represented using an *If-Then* form: "**IF** condition C **Then** Class A". The condition C is called rule antecedent or precondition and is defined by one or more attribute tests logically combined. The *then* part is called rule consequent and consists of the class prediction. A learning system that employs rules typically produces a set of different classification rules, each one matching a different area of the input space. An example

of a classification rule set is represented in Figure 2.6. RIPPER [35] and PART [36] are two of the most well known rule-based algorithms.

IF *age* ≤ *30* AND *student* = *"no"* ⟶ *buy computer* = *"no"*
IF *age* ≤ *30* AND *student* = *"yes"* ⟶ *buy computer* = *"yes"*
IF *age* > *40* AND *credit rating* = *"excellent"* ⟶ *buy computer* = *"yes"*
IF *age* ≤ *30* AND *credit rating* = *"fair"* ⟶ *buy computer* = *"no"*

Fig 2.6: Example of a classification rule set for a computer purchase problem.

The Learning Classifier System (LCS) is a machine learning paradigm introduced by Holland [37] that exploits evolutionary computation to develop a set of conditional rules (classifiers). LCSs have been extensively used in the biomedical domain as a powerful tool for knowledge discovery given their elevated interpretability [38]. There exist two main distinct types of LCSs: Michigan-Style and Pittsburgh-Style. In the Pittsburgh approach each individual is a complete solution for the classification problem, traditionally an individual is a variable-length set of rules. Conversely, in the Michigan approach, each individual is a single rule and the whole population cooperates to solve the classification problem. Although different, both approaches share the goal of finding sets of classifiers that provide a solution for the analysed task.

BioHEL (Bioinformatics-Oriented Hierarchical Learning) [39] is a rule-based evolutionary machine learning system designed to handle large-scale biological datasets. BioHEL generates sets of classification rules using an approach different than the Michigan and Pittsburgh style: the iterative rule learning (IRL) principle. The IRL creates the classification rules sequentially using a standard genetic algorithm (GA). After every rule is generated (the best individual of the GA population), the samples from the training set covering that rule are removed. This learning process is repeated until there are no more examples in the training set. BioHEL uses an explicit default rule in each rule set, the IRL process also terminates if the system cannot generate rules that are better than the default one. The fitness function used by BioHEL is based on the Minimum Description Length (MDL) principle and is defined to promote accurate, general

and compact (simple) rules. BioHEL employs a rule representation called Attribute List Knowledge Representation (ALKR), illustrated in Figure 2.7. ALKR has been designed to cope with biomedical data that often are large-scale, noisy, ambiguous and usually described by a large number of attributes [39]. Each classifier condition is defined by five structures: (1) the number of represented attributes (2) a list of the identifiers of the represented attributes, (3) a list of values for the represented attributes, (4) a list of the positions where each attribute can be found in the classifier and (5) the class of the classifier. The rationale behind this design is that most of the successful rules obtained from biomedical datasets contain only a few key attributes (from the large set of available ones). Hence, automatically discovering these key attributes and only keeping track of them, contributes to a substantial speed-up of the learning phase as it avoids useless match operations with irrelevant attributes. ALKR, rather than coding all the domain attributes, uses a list containing only the expressed ones, this avoids irrelevant match operations (computationally expensive) with non-expressed attributes. In addition, the ALKR structure facilitates important operations during the learning process such as *specialisation* and *generalisation* that add and remove attributes from the list with a certain probability. Overall, the ALKR provides competent learning performance and manages to reduce the system run-time considerably.

Fig 2.7: Representation of a classifier using ALKR.

- **SVM** namely Support Vector Machine, belongs to the family of the linear models, a set of model-based learning approaches that expresses the output as a linear combination of the input attributes. The SVM is based on the concept of decision planes that define decision boundaries [40]. A decision plane, or hyperplane,

tries to separate a set of objects that belong to different classes. Thus, the SVM attempts to generate hyperplanes that separate the samples while maximising the margin, that is the distance between data points from distinct classes. An SVM example is represented in Figure 2.8 (a). More formally, having $\{(x_i, y_i)\}$ with $i = 1, ...l \;\; x_i \in \Re^d$ and $y_i, \in \{-1, 1\}$, where x_i are data points and y_i are the corresponding labels, an hyperplane that separates the objects can be defined as:

$$f(x) = (w^\top \cdot x) + b$$

where w is a d-dimensional coefficient vector that is normal to the hyperplane and b is the offset from the origin. A linear SVM tries to maximise the margin (the distance of the points from the hyperplane) by solving the following optimisation task:

$$\min_w \frac{\|\mathbf{w}\|^2}{2}$$

subject to:

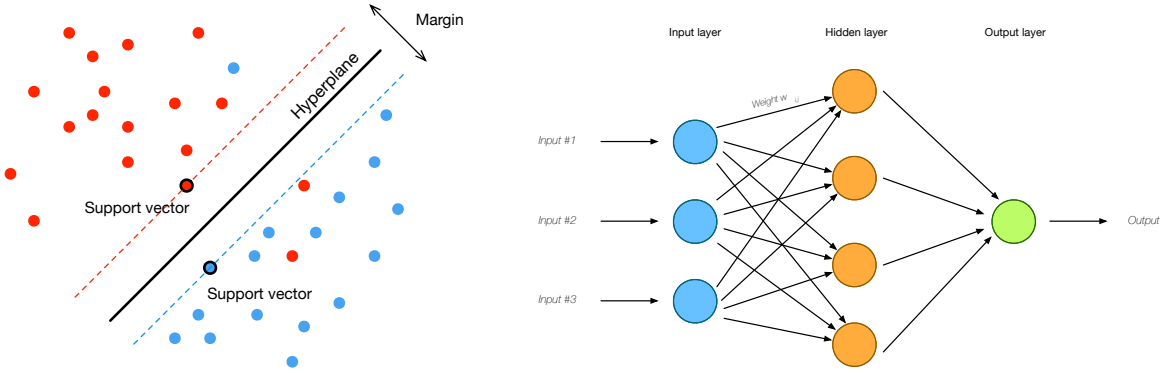
$$y_i(w^\top \cdot x) + b \geq 1 \;\; i = 1, .., l$$

This approach works well only when dealing with data that are linearly separable. If the data are non-linear, SVM, but also other linear classifiers, provides an easy and efficient way to overcome the problem. This is known as the "the kernel trick" [41] and it consists of defining a mathematical function $\Phi : K^n \rightarrow H$ that maps the data into a higher dimensional space where is possible to generate an hyperplane that separates objects from different classes. The most common kernel functions, for two data points x_1 and x_2 are:

- RBF (Radial basis function): $\exp\left(\gamma \left\|x_1 - x_2\right\|^2\right)$
- Polynomial: $\left(x_1 \cdot x_2 + 1\right)^d$
- Sigmoid: $\tanh\left(x_1 \cdot x_2\right)$

where d and γ are user-defined parameters (common default values are 3 and 1/(number of features), respectively).

- **ANN** namely Artificial Neural Network [42], is inspired by the natural neurons and is another example of a linear model. A perceptron represents an artificial neuron, an ANN simply consists of a set of perceptrons connected to each other. The output of the ANN is generated as the weighted sum (strength) of the connections between perceptrons. The set of perceptrons that connects the input nodes (input layer) with the output nodes is defined as the *hidden layer*. A typical ANN with one hidden layer is illustrated in Figure 2.8. The *back propagation* algorithm is commonly used to train the ANN and identify the best set of weights for a particular problem [43]. When having multiple levels of representation, such as an ANN with many hidden layers, we fall into the class of techniques called *deep learning*, nowadays one of the most studied field of machine learning. Deep learning methods are defined by multiple levels of representation (i.e. layers) that are generated by composing simple (non-linear) modules (i.e. neurons). Each module transforms the representation at one level (starting with the input) into a representation at a higher, slightly more abstract level [44]. With such composition of layers, very complex functions can be learned, thus very complex problems can be addressed. Different types of deep neural networks exist, each one better suited for a specific task. For example, convolutional networks (neural networks where the connectivity pattern between the neurons is inspired by the organisation of the animal visual cortex) are ideal for the analysis of data with structured variables such as images, text and audio, recurrent networks (neural networks that contain connection within neurons of the same layer) perform well in the analysis of sequential data such as text and speech, autoencoders are special types of neural networks that receive unlabelled data (unsupervised learning) and aim to transform the input into the output with the least possible amount of distortion (typically used for dimensionality reduction), etc. Deep learning is currently the fastest growing field in machine learning, new successful approaches are continuously proposed to tackle a wide range of problems from predicting the potential of drug molecules to the analysis of particle accelerator data. An interesting overview of deep learning can be found in [44], more detailed information are outside the scope of this thesis.



(a) Support Vector Machine (b) Artificial Neural network

Fig 2.8: Example of linear models: an SVM classifier (left) and a simple artificial neural network (right).

- **Bayesian networks** are acyclic directed graphs in which each node represents a random variable and the edges define probabilistic dependencies among the corresponding random variables. Bayesian networks can be used as models to represent the probability that a certain sample belongs to one class:

$$P(lung\,cancer = yes | smoking = no,\, positiveXray = yes) = ?$$

The probability of the event to occur can be calculated by applying the Bayes' theorem:

$$P\left(A \mid B\right) = \frac{P\left(B \mid A\right) P\left(A\right)}{P(B)}$$

where are $P\left(A\right)$ and $P\left(B\right)$ are the probabilities of observing A and B without regard to each other, $P\left(A \mid B\right)$ and $P\left(B \mid A\right)$ are the probabilities of observing the event A given that the event B is true and viceversa. Naive Bayes [45] is the most simplistic Bayesian network classifier. It has been shown to perform well on many classification problems despite its simplicity and strong assumptions [46]. More complex classifiers have been presented based the Bayesian approach. For example, ABC-Miner learns the structure of a Bayesian network Augmented Naive-Bayes (a Bayesian network graph with no restrictions on the number of parents of a node) using Ant Colony Optimisation [47]. Its extension has also been used for the hierarchical classification of ageing-related proteins [48].

- **Ensembles** are the combination of multiple simple models together. The goal of the ensemble approaches is to obtain better performance than what can be achieved by the single models alone. In 1994, Leo Breiman proposed an ensemble technique called *bagging* (**b**ootstrap **agg**regat**ing**) [49]. Bagging is based on the idea of improving the classification by combining classifier trained on randomly generated training sets. The overall classification is performed by weighting the prediction of the single component of the ensemble. The *random forest* [50], one of the most widely used methods in the recent years, couples a bagging technique with a random selection of features. A random forest is created as an ensemble of decision trees. Assuming a training set with N samples, each one defined by M features, each decision tree is generated as follows:

  1. Randomly select n samples with replacement, this set will be the training set for growing the tree.
  2. In each node of the tree, given a number $m < M$, select m variables at random out of the M. Use the best split on these m to split the node.
  3. Grow each tree to the largest extent possible without pruning.

  Finally, aggregate the predictions of the trees to obtain the classification for the test set. The multitude of positive characteristics of the random forest (i.e. excellent classification performance, an efficient run time with large datasets, variable importance estimation, the ability to work with missing values, etc.) made the method extremely popular not only in the machine learning community. Another ensemble technique is *boosting*: the models are iteratively created based on the samples that were badly classified in the previous iterations. The idea is to generate complementary models that focus on different parts of the input space. Adaboost [51] is one of the earliest most successful examples of ensemble-classifier based on the boosting approach.

2.2.3 Unsupervised learning

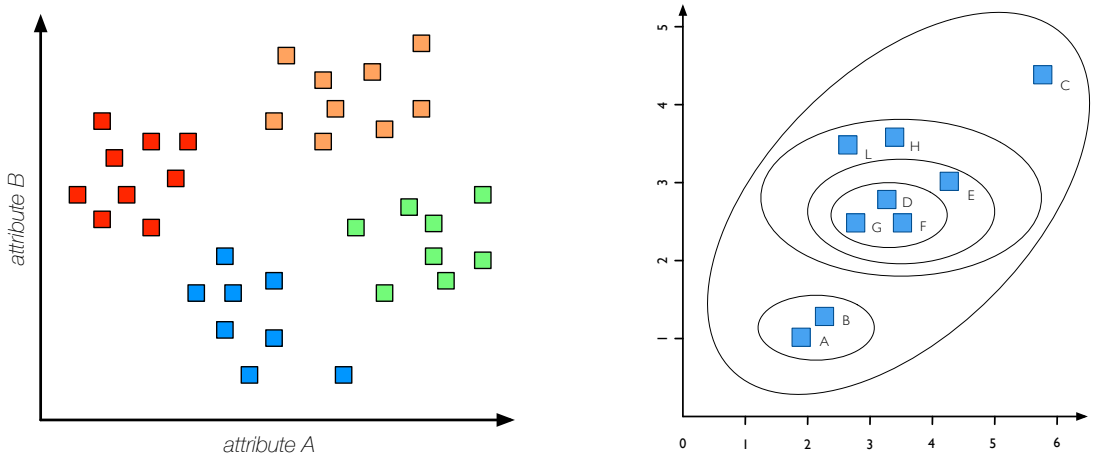
In contrast with supervised learning, unsupervised learning is characterised by the absence of performance feedback. The system tries to automatically identify patterns

within the data without any information about the correctness of the solution. The next paragraphs will describe the most common unsupervised methods employed in a biomedical context.

Clustering is the task of grouping a set of observations so that the participants of the same group, called *cluster*, are more similar to each other than to those in other clusters, see Figure 2.9 (a). If the clusters are allowed to have sub-clusters, then it becomes *hierarchical clustering* where commonly the nested clusters are organised and visualised as a tree, see Figure 2.9 (b). Clustering is an operation that can be performed by different algorithms, using different definitions of a cluster and different measures to assess the similarity between objects. K-means clustering is one of the simplest and most used clustering algorithms [52]. First, k initial centroids are chosen, where k is a parameter that indicates the number of desired clusters. Then, each data point is assigned to the closest centroid, a set of data points assigned to the same centroid is defined as a cluster. Different metrics can be used to determine the distance between clusters, the most common are the Euclidean and the Manhattan distance. The centroids of the cluster are then updated based on the data points contained in the cluster. The process is repeated until either no points change clusters, or the centroids remain the same. The tricky part of K-means is to assign the correct value to k, the number of clusters should match the data. The Fuzzy K-means is a variance of the original K-means algorithm where each data point can belong to more than one cluster with certain probabilities [53]. In biomedicine clustering represents an important tool, it has been used for many different problems such as: create a taxonomy of living things, identify groups of genes with similar biological functions, stratify patients with similar clinical characteristics, etc.

Association rule mining aims to find frequent and interesting patterns or associations among the observations in a dataset. Association rules are defined as an implication of the form:

$$X \Rightarrow Y$$



(a) K-means Clustering (b) Hierarchical clustering

Fig 2.9: Example of clustering: K-means with K=4 and hierarchical clustering.

where $X, Y \subseteq I$ and $I = i_1, i_2, ..., i_n$ is a set of attributes called *item*. Association rules do not differ much from the classification rules presented in Section 2.2.1, except that they predict any attribute, not simply the class attribute, including a combination of them. A famous association rule, which emerged from the analysis of supermarket shoppers, is *diaper* $\Rightarrow$ *beer*. A study showed that customers (presumably young men) who buy diapers also tend to buy beer. Apriori is a classic algorithm for learning association rules [54]. The algorithm tries to find subsets of attributes which are in common to at least a minimum number instances. Using a "bottom up" approach, frequent subsets are extended one item at a time and groups of candidates are tested against the data. Apriori terminates when no further extensions can be found. Many other algorithms have been proposed after the Apriori algorithm [55].

PCA namely Principal Component Analysis, is a process that aims to summarise the original set of variables into a smaller set that collectively explains most of the variability in the original variables [56]. PCA uses an orthogonal transformation to convert a set of data points, of possibly correlated variables, in a set of values of linearly uncorrelated variables called principal components. Thus, the total number of components is less or equal than the number of original variables. PCA is well known

as it helps in visualising high dimensional data in 2D plots (when considering only the first two components), as illustrated in Figure 2.10.

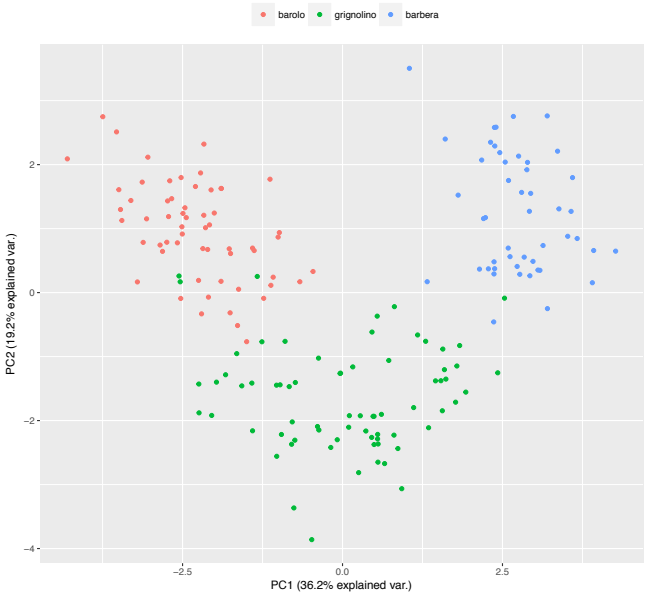


Fig 2.10: Plot of the first two components from a PCA generated with a 3 class datasets.

One-class classification also known as unary classification, aims to identify objects that belong to a specific class [57]. By learning from a training set containing only data points of a particular class, the model can recognise whether an unseen test point belongs or not to that class. The one-class classification is commonly adopted for anomaly (outlier) detection where the goal is to recognise objects that do not conform to an expected pattern. One-class SVM is probably the most used classifier based on this unary approach [58].

2.3 Machine learning for the inference of biological networks

The inference of biological networks is a highly relevant and challenging task in systems biology and bioinformatics [59]. The analysis of biological phenomena through a network representation nowadays has become a common approach. Biological networks are graphs in which nodes represent a biological entity (such as genes or proteins), and

a connection between them indicates some kind of biological relationship, e.g. regulatory or functional. Network inference is, in an essence, an attempt to reverse engineer the biological relationships from the data [60]. Networks, generated from biomedical data, represent a tool to investigate complex biological systems, not only as individual components but as a whole. The analysis and study of biological networks have extended our understanding in many biomedical contexts: from the discovery of the involvement of gene-gene interactions in diseases [61] to the analysis of therapeutic drugs and their targets [62], from the prediction of protein functions [63] to gene regulations [64], etc.

Over the years, the adoption of machine learning techniques to address the challenging tasks of identifying gene-gene associations, and more generally, to infer biological networks, has continuously gained popularity. This is mainly due to the wide range of knowledge representations that can be used within machine learning methods (e.g. classification rules, decision trees, artificial neural networks, SVM kernels, etc.) and that can lead to a discovery of more complex and diverse relationships. Having such a large variety of knowledge representation, within machine learning models, the attributes are associated not because they are similar (e.g. have similar expression profiles), but because together they detect strong patterns.

Different types of machine learning algorithms have been successfully applied to infer networks and associations from biomedical data. In the next paragraphs are reported some examples, grouped based on the machine learning algorithm used to generate/infer the networks:

Association rules Martinez-Ballesteros et al. abstracted genetic associations from association rules [65]. Their approach defines an edge between the elements of the antecedent and the consequence. The Apriori algorithm, a well-known method for the inference of association rules, was also employed to resolve KIR gene patterns associated with haematological malignancies [66].

Decision trees have been successfully used to extract gene-gene dependencies. For example, in [67, 68] each gene is in turn set as a target gene and a model tree is built

to predict its expression values using the other genes. Once the tree is constructed, a linear regression function is generated for every node of the tree. Finally, the edges are defined among the genes involved in the same linear models.

Random Forest Yoshida and Koike [69] presented a modification of the classical random forest classifier with the capacity to identify multiple interactions simultaneously. Different than the original approach, multiple attributes are selected at each node. The possible genetic associations, between SNPs in the presented study, are represented by the branches of the trees. Each branch accounts for a possible SNP interaction, if a certain SNP combination appears quite often on a branch, then those SNPs are likely to interact more strongly. Another approach is known as *permuted random forest* (pRF) and selects the top interacting SNP pairs by assessing how much the removal of an attribute pair influences the random forest classification [70]. One of the recent network inference DREAM challenge, a series of competitions organised to foster collaborations and propose new solutions for many questions in biology and medicine, was won by a team proposing GENIE3 [71], a method based on an ensemble of decision trees. In GENIE3 a model is iteratively created for each gene with its expression levels set as output values and the expression levels of the other genes as input values. A rank is then extracted for each model to guide the discovery of gene-gene interactions.

SVM Chen et al. [72] used an SVM approach that when combined with search algorithms generates different models to detect gene-gene interactions. SVM, using a quadratic kernel, was able to show that multiple SNP sites from several genes, located in zones of the genome that are far away, are better at predicting patients with breast cancer than single SNP [73].

Bayesian Networks Most of the molecular measurements are continuous, so they can be naturally described using continuous Bayesian networks. In the literature, there have been many examples of inference methods from Gaussian Bayesian networks where each node represents a continuous variable and it is modelled as a function of its parents plus and added Gaussian noise [74–76]. Recently GEBN (Grammatical

Evolution Bayesian Network) was presented [77]. GEBN employs Bayesian Networks to infer interactions from biological data, the novelty is that at the same time, it also uses an evolutionary algorithm to reduce the computational cost due to the network optimisation.

2.4 Similarity-based approaches for the inference of biological networks

Together with machine learning approaches, other methodologies have been presented to tackle the problem of inferring meaningful biological networks. One of the earliest (but still widely used) proposed approaches is based on the "guilt-by-association" principle [78]. That is, if two genes show *similar* expression profiles, it is assumed they are also biologically related (via a direct or indirect interaction). Initially, this paradigm was applied to infer networks from transcriptomics data, and this is why in most of the literature it is known as the *co-expression* network inference principle. Nevertheless, it is abstract enough to be applied to all kinds of biological data. This thesis will refer to this approach simply as "co-expression" and to its application to transcriptomics (gene expression) data.

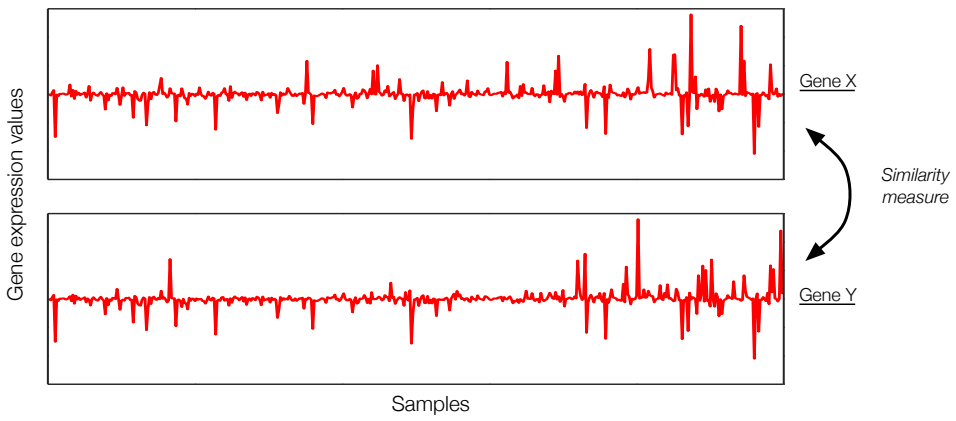


Fig 2.11: The co-expression paradigm identifies associations between genes that have similar expression profiles across different samples.

The co-expression paradigm identifies similarity of gene expression patterns under different experimental conditions. Two genes are considered to be co-expressed, therefore biologically related, if their transcript levels are similar across a set of samples,

see Figure 2.11. By considering two genes as two random variables X and Y, there are multiple approaches to measure the relationship between them [79]. Most of the methods based on the co-expression paradigm utilise two main association measures: correlation and mutual information (MI).

2.4.1 Correlation-based methods

Correlation is a commonly used association measure. Pearson Correlation Coefficient (PCC) is probably the best-known correlation measure of linear dependence between two variables. When applied to gene expression profiles, it measures the similarity in the direction of the gene responses across samples. Its main disadvantages are the lack of distributional robustness (it assumes data normality) and the sensitivity to outliers. Given a pair of genes x and y, the PCC is calculated as:

$$PCC(x,y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \times \sqrt{\sum_i (y_i - \bar{y})^2}}$$

where x_i and y_i are the expression values for the genes in a given sample i and $\bar{x}$ and $\bar{y}$ are their mean expressions values. The PCC measures how much a pair of genes tends to respond in the same (or in the opposite) direction across different samples. The value ranges from -1, revealing that the genes respond in totally opposite direction, to $+1$, indicating a similar way to behave across the samples. Alternatively, Spearman correlation, based on ranks, can be used. It measures the extent of a monotonic relationship between two genes X and Y. Spearman correlation offers more stability to the outliers, however, the outlier effect was shown to be small in large-scale microarray data [80], furthermore, in the same study, the two measures performed overall similarly. A biological network can be created by ranking all the possible gene pairs based on their correlation value and by finding a threshold that defines which edges to include. However, the selection of the optimal threshold is a non-trivial problem [81]. A well-established method for the generation of correlation networks is WCGNA [82], over the years it has been employed, with successful results, in many different fields: from system biology to neuroscience.

2.4.2 Mutual information-based methods

Complementary to the correlation, the information theory and the mutual information can be employed to estimate the association measure between two variables. The mutual information $MI(X;Y)$ determines the entropy to quantify the amount of information that Y contains about X (measured in bits):

$$MI(X,Y) = \sum_{x_i \in X} \sum_{y_j \in Y} p\left(x_i,\, y_j\right) \log_2 \left(\frac{p\left(x_i,\, y_j\right)}{p\left(x_i\right) p\left(y_j\right)} \right)$$

In contrast to the PCC, mutual information can detect non-linear dependencies. In its original definition, the mutual information measure was defined for discrete or categorical variables. Therefore, its application to continuous variables (e.g. gene expression variables) is challenging. Several strategies have been proposed in the literature. The most common approach consists on discretising the numeric vector X by using the equal width method. This approach divides the interval $[min(X), max(X)]$ into equal-width bins, the resulting discretised vector has the same length as X but its i-th component provides the bin number in which X_i falls in. The only parameter necessary is the number of bins of the equal-width.

ARACNE [83] is a method to generate biological networks by measuring the dependence between two gene expression profiles via the mutual information. ARACNE calculates $MI(X;Y)$ for every pair of gene expression profiles X and Y, and applies the data processing inequality to remove the majority of indirect dependencies. For each triplet X, Y and Z, the weakest link is removed, that is the edge between X and Y is removed if:

$$MI(X;Y) \leq \min(MI(X;Z),\ MI(Z;Y)) - \epsilon.$$

The tolerance threshold ϵ is used to adjust for the variance of the mutual information estimator. Other approaches were also defined for identifying biological networks using the mutual information theory: CLR [84], MRNET [85] and RELNET [86].

The Maximal Information Coefficient (MIC) [87] is yet another proposed measure for the strength of association between two variables that is closely related to mutual

information. MIC, in contrast with the classical approaches, based on a single discretisation strategy to bin the compared variables, chooses individual bins for each variable, such that the value of mutual information $MI(X;Y)$ is maximised. Compared to the standard estimation of $MI(X;Y)$ value used in ARACNE, the optimised estimation provided by MIC can detect a wider range of non-linear associations. In addition, MIC has been shown to identify a more diverse variety of association between variables when compared to PCC, see Figure 2.12. MIC was capable to identify known and novel relationships from a wide range of datasets such as global health, gene expression, baseball and human gut microbiota [87].

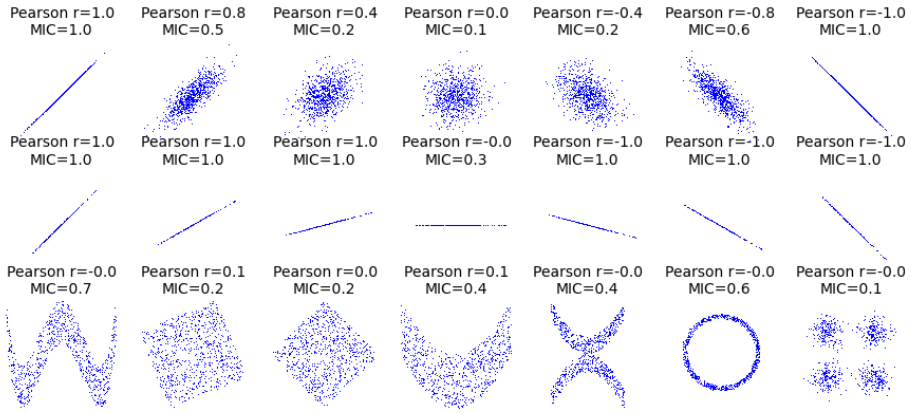


Fig 2.12: Comparison of the similarity measures calculated with PCC and MIC for different type of associations between two variables [88].

2.5 Network inference via the integration of multiple data

Datasets often present a limited overview about a specific biomedical problem. This is mainly because some types of experiment can only provide information about a specific aspect of the cell's behaviour [89]. Furthermore, different computational approaches, when dealing with the same set of data, can lead to different and complementary results. Hence, the integration of multiple computational models, biological networks (or associations) in this instance, can produce more robust solutions. A common approach is to score each model using a gold standard. For example, Lee et al. assess the goodness of the proposed (integrated) associations against a gold standard and

calculates a log likelihood score that generates a network with weighted edges (the weight represents the sum over all the data sources) [90]. An extension of this approach uses instead multiple gold standards to infer the overall network and compute an existence probability for each association [91]. Those type of networks are often termed as *probabilistic functional networks* as each edge is weighted with a confidence score (probability) representing the likelihood of the association. The bias presents in the experimental data is exploited, rather then be eliminated, in [92], where the authors integrate information for multiple sources to optimise network predictions relevant for a specific biological process. An alternative paradigm is the semantic data integration where all the multiple types of associations (i.e. generated with different approaches) are kept and separately identified. The main challenge of semantic networks is to assess the confidence score for each relationship because each type must be scored separately. To tackle this problem, Weile et al. [93] proposed a generic solution that, based on a fully Bayesian method, calculates the probability that each statement is true in the semantic graph without using gold standard but completely relying on experimental data.

2.6 Statistical approaches for biomarkers identification

Modern technologies allow one to sense the state of large quantities of biological entities at once. Thus, biomedical data are often characterised by samples that are defined in a high-dimensional space (thousands of features). Unfortunately, most of those dimensions do not contain any relevant information that describes the phenomena analysed in the experiment. In addition, several studies revealed that most of the biological measurements available in the experiments are not helpful when it comes to the classification of data points into different categories [94]. Therefore, it is fundamental to identify, among all the available information, driving factors that can be relevant for a biological/medical condition. Such a process is widely known as **biomarkers discovery**, where a biomarker is defined 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*" [95].

When analysing biomedical data, statistical approaches have been extensively exploited to select important biomarkers. Univariate methods are the most straightforward and commonly adopted statistical approach for the biomarker discovery [96]. The idea is to verify the possible association between each factor (variable of the dataset) and the outcome variable (e.g. presence of cancer, effect of a drug, etc.) via a statistical test. Traditional approaches utilise statistics measurements, such as mean or median, to evaluate the difference between groups of individuals, like healthy vs. unhealthy patients. One of the most basic and used methods is the *Student's t-test*, it verifies whether the means of two population are equal or not. In the context of biomedical data, the population are represented by samples that belong to different categories. The formula of the Student's t-statistic, for the analysis of a single variable, is given by:

$$t = \frac{\bar{x_A} - \bar{x_B}}{\sqrt{\frac{s_A^2}{n_A} - \frac{s_B^2}{n_B}}}$$

where $\bar{x}$ and s represent the mean and the standard deviation of the populations A and B, while n indicates the size of each population. From the t-statistic, using appropriate tables, a p-value can be extracted representing the strength of association between the variable (factors) and the outcome variable. The Student's t-test assumes that the values of the variables are normally distributed and requires a minimum of 20 samples per category (rule of thumb), and if this is not verified, wrong associations can be inferred. When the normality assumption does not hold, and the distributions of the variables are skewed, alternative tests offer better statistical power, for example, the Mann-Whitney U test [97]. If the biomedical data are defined by categorical variables (e.g. SNPs data), the χ^2 test can replace the Student's t-test. ANOVA is a well-known test used to detect significant factors in a multi-factor model. A multi-factor model is characterised by a response (dependent) variable (e.g. healthy state) and one or more factor (independent) variables (e.g. age, BMI, gene expressions, etc.). ANOVA tests use variances to know whether the means of different groups are equal or not. Commonly, ANOVA is employed to evaluate the influence of groups (≥ 2) of variables on the response variable; this approach allows to consider interactions and main effects between factors. Given this characteristic, ANOVA is also viewed as a special case of linear regression [98].

When analysing large-scale biomedical datasets, the number of variables can go up to several thousands, thus a common practice is to couple a univariate standard method with a more complex multivariate approach [99]. By using a multivariate method, (e.g. ANOVA or linear regression), potential relationships between factors, missed by univariate methods, are taken into account. A well-established pipeline for the identification of biomarkers, used in many clinical studies, includes three main steps:

1. Filtering of irrelevant factors using a univariate analysis
2. Generation of a (risk) prediction model using a multivariate approach
3. Identification of the most important biomarkers within the model

First, the large set of variables included in a biomedical dataset is filtered using one of the cited univariate approaches. Then, considering only the significant variables identified at step 1), a multivariate predictive model is generated. Finally, the most relevant variables within the predictive model are contemplated as possible biomarkers for the analysed condition. The prediction model is commonly generated using a logistic regression analysis. Logistic regression provides methods to model binary response variables, for example, presence of a medical condition [100]. What makes the logistic regression different from the linear regression is that it doesn't measure the outcome variable directly, but instead it calculates the probability of obtaining a particular value for the outcome variable. The equation that provides the probability for an event to occur, in logistic regression, is:

$$p = \frac{e^{\alpha+\beta x}}{(1 + e^{\alpha+\beta x})}$$

where $\beta x = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_n x_n$ with β_i representing the coefficients (parameters) for the variable x_i and α representing a model with no predictor variables. Typically the values of α and β are determined using the maximum likelihood estimation [100]. Once the model is defined, it is possible to test the statistical significance of each coefficient, that is to assess the importance of prediction of each variable (predictor). The Wald test is a standard choice and it tests the hypothesis that each $\beta = 0$.

Finally, the p-values calculated by this test are used to determine which variables can be considered as a possible biomarker and can be granted with further investigations.

Given the high dimensionality of biomedical data, often multiple tests are necessary to identify relevant variables. However, as the number of performed tests increases, the chance of encountering false positive results (non-significant factors recognised as significant) rises in parallel. Therefore, a correction for multiple testing is fundamental to obtain meaningful results. If for each test the significance level α is set to 0.05, there is 5% chance to accept not significant results. Thus, if 100 tests are performed together, 5 of them could be found significant by chance, the value increases to 500 when performing 10 000 tests (e.g. checking gene expression values). Multiple testing is used to address this problem by adjusting the individual p-value (after which is also called q-value) to keep the overall error rate at the desired level. Different methods can be applied [101]. The Bonferroni correction aims to control the family-wise error rate; it sets the significance cut-off to α/n where n corresponds to the total number of performed tests. So, if all the null hypotheses are true, the probability that the set (family) of tests includes one or more false positives by chance is 0.05. The Bonferroni approach can be too conservative (many hypothesis rejected) when the number of tests is high. A less stringent method that controls the family-wise error rate at an α level is the Bonferroni-Holm correction. The p-values are sorted from smallest to largest, iteratively the p-values are multiplied (adjusted) by $(n - i)$ where n is the total number of tests and i indicates the rank of the current test. The procedure stops when no test is found to be significant. A different approach is to control the false discovery rate (FDR), that is the proportion of significant results that are actually false positives. In other words, the test sets the percentage of false positives results that you are willing to accept among the significant ones. The Benjamini-Hochberg is a well-established technique to control the FDR. Given the sorted set of p-values (from smallest to largest), each p-value is compared to its Benjamini-Hochberg critical value of $(i/n) \times \alpha$ where i is the rank, n is the total number of tests, and α is the false discovery rate chosen. The largest p-value lower than $(i/n) \times \alpha$ is significant together with all the p-values smaller than it.

2.7 Machine learning for biomarkers identification

Along with the traditional statistical methods described in the previous section, machine learning methods have started to make a great impact in the field of biomarker discovery [10, 102, 103]. One of the main characteristics of machine learning approaches is their ability to identify complex patterns (usually with a multivariate approach) within the data, this become fundamental when analysing phenomena that are the product of complex chains of interactions among many factors. Given the complexity and the vast variety of knowledge representations that can be adopted by machine learning methods, potentially they can overcome the limitations represented by the traditional statistical methods. The process of selecting relevant features, employed in machine learning, is called *feature selection* (or extraction). Thus, the discovery of biomarkers from biomedical data can be modelled as a feature selection problem. The leverage due to the dimensionality reduction in machine learning is two-fold: (1) new driving factors for complex diseases can be easier to identify and (2) the learning algorithm can obtain better performances at a reduced computational cost when working in a smaller dimensional space. Traditionally, feature selection approaches can be summarised within three main groups: filter, wrapper and embedded methods (see Figure 2.13):

Filter methods evaluate the relevance of feature subsets by analysing the intrinsic properties of the data. Typically a single attribute or a subset of attributes is evaluated against the class label. Often, for each feature an importance score is calculated, then the features with the lowest scores are discarded. Filter techniques offer high scalability, they are computationally efficient and they are independent of the classification algorithm used in the later stages of the analysis. On the other hand, the latter can be a drawback as the interactions with the classifier are ignored (i.e. the search in the feature subset space and the search in the hypothesis space are separated). CFS is an example of a multivariate filter-based feature selection method [104]. By exploiting a best-first search, it assigns high scores to subsets of features highly correlated to the class attribute but with low correlation between each other. The information theory can guide the selection of the best subset of features as in [105], where the search

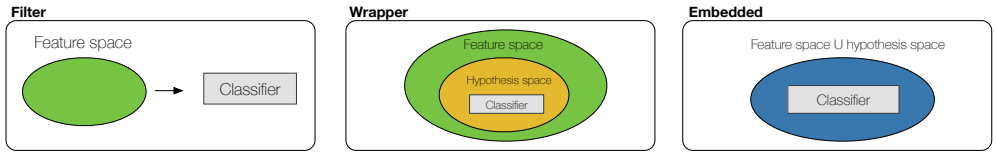


Fig 2.13: Taxonomy of feature selection methods: filter, wrapper and embedded (adapted from [102]).

process is based on a maximum weight and minimum redundancy (MWMR) criterion. Relief is another renowned filter-type algorithm [106]. First, it randomly samples an instance from the data, then it picks the nearest neighbour from both the same and the opposite class. By comparing the values of the attributes of the nearest neighbours to the sampled instance, Relief defines a relevance scores for each attribute that is used for the final attribute selection. Relief is based on the idea that relevant attributes should have similar values for instances of the same class while differentiating between instances of different classes.

Wrapper methods different than the filter approaches, include the model hypothesis search within the feature subset search. The idea is to use a classifier to determine if a subset of features perform well in the classification task. Many wrapper methods are coupled with heuristics to help the search for an optimal set of features as the space of possible feature subsets grows exponentially [107–109]. GA-KDE-Bayes is a fairly recent evolutionary wrapper method that joins a non-parametric density estimation method with a Bayesian classifier [110].

Embedded methods provide the search for an optimal subset of features embedded within the classifier construction. This approach can be seen as a search in the combined space of hypotheses and feature subsets. SVMs have been successfully used to guide the discovery of feature subsets within the classification task. SVM-RFE (Support Vector Machine - Recursive Feature Elimination) is probably the most famous example, an iterative feature reduction method that was designed to deal specifically with genetic data but that nowadays is employed in many fields [111]. The weight vector from the SVM classifier is used as ranking criteria, iteratively the features with the lowest rank are discarded until a small set is obtained. Another approach, named

kernel-penalized SVM (KP-SVM) is used to optimise the shape of an anisotropic RBF Kernel by removing the features that have low importance for the classifier [112]. In [113], to achieve efficient gene selection from thousands of candidate genes, particle swarm optimisation was combined with a decision tree classifier. BioHEL [39], presented in Section 2.2.1, performs an *embedded* feature selection during the learning phase. The ALKR automatically identifies the relevant attributes and discards the irrelevant ones. Hence, BioHEL employs a fine-grained embedded feature selection as only the most relevant attributes are identified and considered in each rule. In addition, a feature importance rank, that can drive a further selection phase, can be generated by counting how many times each attribute has been used (importance) in the BioHEL classification rules.

2.8 Knowledge integration in machine learning methods for biomarkers identification

A research path that is emerging involves the integration of prior biological knowledge into the model inference process [114]. The prior knowledge can assume different forms: from cellular pathways to biological and molecular networks. The reason behind this emerging approach is that by using patterns extracted from prior knowledge, deceptive information embedded within the data can be identified (e.g. spurious structures) and help the learning model to be mainly focused on predictive features that are coherent with the knowledge depicted in pathways or molecular networks. Vlassis and Glaab presented GenePEN, an algorithm for the identification of gene (or protein) sets that are both predictive for disease-control datasets and form connected components within a provided biological network (such as a protein-protein interaction network) [115]. GenePEN obtained not only similar performing results, compared with other feature selection methods, when working with in-silico datasets, but also was able to identify, from Parkinson's disease data, a subset of genes enriched for that disease. In [116], driver condition factors were identified by applying a greedy search algorithm to find subsets of genes, members of the same biological pathway, that maximise a t-statistic score for the discrimination of control-case samples. The selected genes had higher discriminative power, across multiple datasets, when compared to gene-based

classifiers. Kim et al. [117] created a biomarker model for transcriptomics data at two levels (gene and pathway level) by using a hierarchical feature structure. Finally, in [118] biomarkers are identified via a two-step procedure. First, a random walk is performed on a molecular network where the weight of each node corresponds to the fold change of that gene in the cancer-related expression. Then, network modules (gene sets) are ranked based on their score (the square of the average weighted expression for all the members). Overall, this research strategy is showing to be promising and is leading towards more reliable biomarker discoveries, limiting the risk of overfitting that can affect pure data-driven approaches.

2.9 Biomedical evaluation of the results

In many fields researchers have the luxury to use ground truth data to verify, in a simple way, the validity of their computational methods. Ground truth is referred to a well establish set of data and results that a method is expected to reproduce to be considered correctly working. Examples are face recognition problems, text mining, etc. In bioinformatics and system biology this is rarely the case. Problems such as the identification of transcription factors or the discovery of regulation among genes are few exceptions. In literature, there exist transcriptional networks (especially for simple organisms such as E.coli [119]) that the researchers can employ as ground truth. To assess the performance of a regulatory network inference method, for example, it is possible to count how many known interactions were identified and how many false positives (unconfirmed associations) were generated. An alternative is to create synthetic biological data (and networks) that resemble the real ground truth. The DREAM challenges, for examples, offer some in-silico benchmarks commonly employed for the analysis of new algorithms [120].

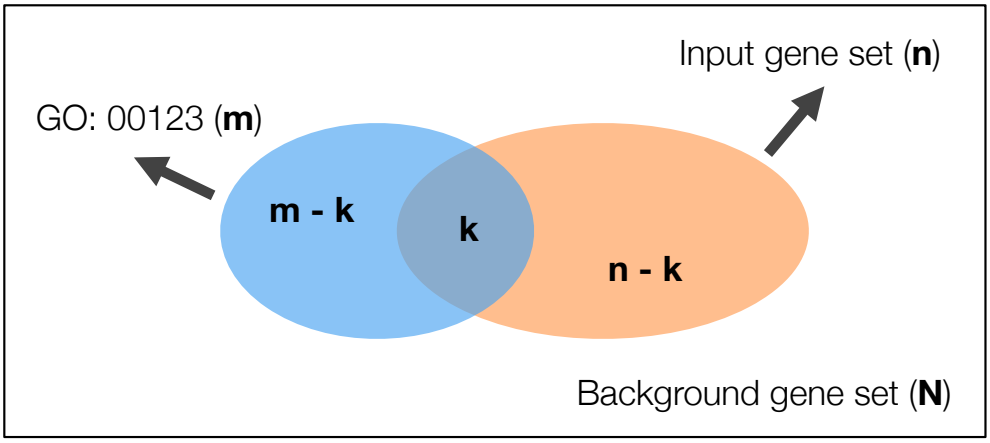
Unfortunately, when it comes to the evaluation of functional networks, which are one of the main topics of this dissertation, no ground truth network is available. The ideal solution is to assess the relationship and the role of biological entities via experimental tests. However, this solution is time-consuming and expensive, and thus rarely feasible. Therefore, other techniques are necessary to estimate the goodness of new methods.

Two main approaches are used to tackle this problem: (a) enrichment analysis and (b) the use of established and confirmed literature knowledge. Both will be briefly described in the following subsections, their application is not limited to the analysis of biological networks, but it can be extended to the validation of newly identified biomarkers. The section about the confirmed literature knowledge will focus on the description of known associations between gene and biomedical phenomenon (disease).

2.9.1 Enrichment analysis

The enrichment analysis is a method for checking whether a set of biological entities (mainly genes) have common characteristics based on a statistical approach. The goal of this process is to assign a biological meaning to groups of genes and provide a tool for the interpretation of biological results. The enrichment analysis exploits statistical methods to identify biological features (annotations) that are represented in a particular gene set more than it would be expected by chance. Typically, the annotations of the input set are contrasted with the annotations of elements that belong to a background set (e.g. human genome). The biological features that statistically appear more in the input set are called *enriched* (overrepresented). The most common source of biological knowledge employed for the overrepresentation analysis is the Gene Ontology (GO) [121]. The Gene Ontology is a public annotation database which provides descriptions of molecular functions, biological processes and sub-cellular locations attributed to gene products. In addition to the Gene Ontology, other sources offer biological information: KEGG, MeSH, PubMed, OMIM, etc. KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies [122]. PubMed is instead a search service that provides access to over 11 million of scientific references and abstracts related to biomedical topics.

The general approach for an enrichment analysis is illustrated in Figure 2.14. Many different tests can determine the statistical association between a term and a gene set. Figure 2.14 shows the Fisher Exact test implemented by DAVID [123], a well-known



	Input gene set	Back. gene set
GO: 00123	k	m
Not GO: 00123	n-k	N-m

$$P(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$

Fig 2.14: Example of a gene set enrichment analysis performed using a Fisher exact test.

enrichment tool. By counting the number of genes, of both the background and the input set, that are associated with a term (GO:00123 in this case), the Fisher Exact test provides a p-value that indicates the likelihood of obtaining the gene set-term association by chance. Other statistical approaches include: binomial test, chi-square test, hyper-geometric test, etc. [124].

Plenty of tools are publicly available to perform gene set enrichment analysis and visualise its results. g:Profiler is a public web server for characterising and manipulating gene lists of high-throughput genomics; it includes data for more than 80 species including mammals, plants, fungi, insects, etc. [125]. The analysis is based on multiple sources of functional evidence, including Gene Ontology terms, biological pathways, regulatory motifs of transcription factors and microRNAs, human disease annotations and protein-protein interactions. The PANTHER (Protein Annotation THrough Evolutionary Relationship) classification system combines gene function, ontology, pathways and statistical analysis tools to help the analysis of large-scale -omics experiments [126]. The PANTHER website includes a suite of tools that enable the evaluation of gene sets according to their function in many different ways: families and subfam-

ilies are annotated with GO terms, PANTHER protein class and pathways. Most importantly, PANTHER is developed by Gene Ontology consortium and is constantly maintained up to date. EnrichR [127] is yet another enrichment analysis web-based service that associates prior knowledge (pathways, ontologies, diseases, drugs, transcription) to gene lists. The strength of this tool relies on the visualisation of the results into different forms including networks and grid of terms that allow a simpler and better interpretation of the inferred knowledge.

In the last few years a new approach for enrichment analysis has emerged, similar to what is described in Section 2.9. It consists of using other sources of knowledge to provide better and more relevant results. The information encapsulated within biological networks (such as protein-protein interaction, molecular networks, gene regulatory networks, etc.) is exploited to better assess common characteristics among a set of biological entities. Two main classes of approaches can be identified: (a) methods that exploit the topology of the networks to check how similar set of genes or protein are and (b) methods that first identify functionally-related modules within the networks and then define the biological role of genes or protein from such modules. EnrichNet [128] belongs to the first class and uses the knowledge associated with biological networks to obtain stronger enriched results. EnrichNet maps the input gene set onto an interaction network and using a random walk, scores distances between the genes and pathways (taken from a reference database). The XD-score is a network-based association score and is relative to the average distance to all pathways; it also represents a deviation from the average distance. PINA (Protein Interaction Network Analysis) [129] belongs instead to the second category and is based on the integration of 6 different PPI databases. The core of PINA consists in identifying clusters of densely connected nodes which are likely to represent sets of proteins that are functionally related. The input gene/protein set is then mapped on the clusters and an hypergeometric enrichment test identifies overrepresented clusters. Finally, the input set is characterised by the annotations of the enriched clusters. TopAnat is instead a tool to identify enriched anatomical terms from the expression patterns of a list of genes. Different than common enrichment tools, it discovers where genes are preferentially expressed, as compared to a background set, represented by default by all expression

data in Bgee [130] for the requested species (e.g. human). TopAnat is similar to a standard Gene Ontology enrichment test, except that it analyses the anatomical structures where genes are expressed, rather than their GO functional annotations.

When it comes to the interpretation of networks, this can be harder than explaining the commonalities between sets of biological entities. A trivial adaptation for the biological interpretation of the networks can be to perform the enrichment analysis on either the whole set of nodes or on a subset of important nodes. In the latter case, a solution is to apply a clustering algorithm (such as MCODE [131]) to the network and select a subset of relevant nodes (e.g. highly interconnected within each other). Network algorithms have also been developed to identify sets of interconnected nodes sharing a common phenotype or a consistent response across experimental conditions [132]. SANTA [133] represents a general approach for the extension of functional annotations from gene lists to biological networks. The input of SANTA are a gene set and a network, its goal is to verify the statistical significance of their association. Based on the guilt-by-association principle and using spatial statistics techniques, the functional information content of any biological network can be assessed with respect to a given set of seed genes (for example the set of genes annotated with a specific GO term). A gene set is called "associated" with the network if it shows a surprising degree of clustering on the network, otherwise, if it is randomly distributed over the network, is defined as "non-associated".

After having performed an enrichment analysis, the final step is to verify whether the list of overrepresented terms is meaningful in the studied biological context. However, it is crucial that researchers pay attention to the selection of the enrichment tools. As shown in a recent publication [134], many enrichment tools provide outdated gene annotations (a summary of the results is reported in Figure 2.15). An outstanding example is DAVID [123], a tool referenced by more than 4000 publications in 2015 (and in 80% of the publications considered by Wadi et al. [134]), that until October 2016 was still employing a knowledge base dated 2009. By using tools that are not up to date, not only the significance and the interpretation of the analysis becomes unreliable, but new wrong hypothesis can be generated and negatively impact follow-up studies.

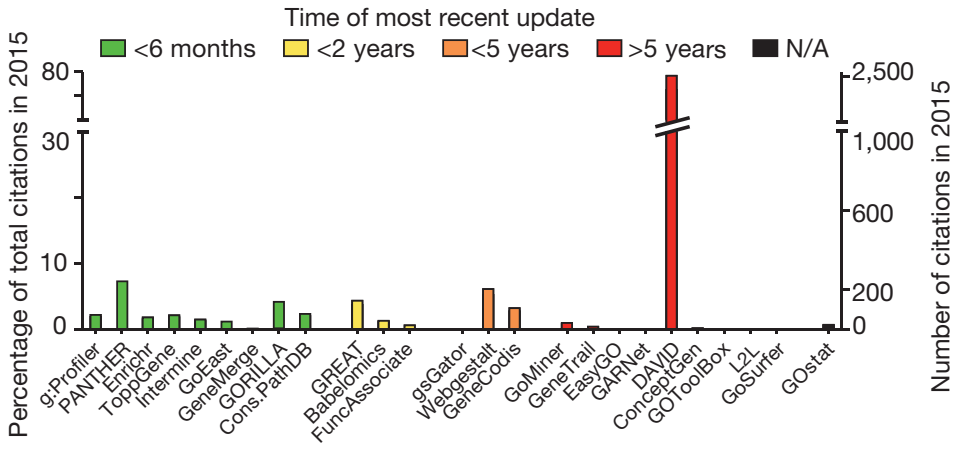


Fig 2.15: Time of the last update for the most commonly employed and cited enrichment analysis tools. Image taken from [134].

2.9.2 Gene-disease associations

Most of the large-scale datasets that are continuously generated, contain molecular measures (e.g. gene expression values) extracted from patients or tissues affected by a particular condition, often a specific disease. Those experiments are performed to understand the complex mechanism behind diseases and syndromes. Thus, a way to evaluate the knowledge extracted from those type of data is to check whether the inferred output can be confirmed by existing information about that specific disease. A Gene-Disease association (G-D) represents a circumstance where a gene is directly or indirectly responsible for disease risk via one or more mechanisms [135]. G-D associations can be identified via experimental techniques, however, many diseases are multigenic, that is caused and influenced by several genes. Due to this complexity, finding causal links between a gene and a disease with experimental techniques is expensive and time-consuming. A solution is offered by the advance of high-throughput techniques that allow to probe thousands of genes and can return hundred of candidates genes. Associations between genes and diseases can, for example, be extracted from techniques such as genome-wide association studies (GWAS). Numerous databases nowadays gather and integrate these data to provide reliable associations [136]. Entries can be obtained either via manual curation of the specialised literature [137], or through automated text mining approaches [138]. Other portals, such as Malacards [139] or OpenTargets [140], use data integration techniques to gather G-D associations

from many different sources (64 in Malacards) and rank them according to some importance/reliability criteria. In addition, mouse and rat models have also been used to predict G-D association in humans [141, 142].

Several public databases contain G-D associations, often their knowledge is extracted directly from a manual curation of specialised literature. OMIM (Online Mendelian Inheritance in Man) [143] represents a well-established source for associations. CTD (Comparative Toxigenomics Database) [144] and UniProtKB [145] are sources for G-D relations as well. Orphanet [146] is yet another source that targets mainly rare diseases and orphan drugs. As mentioned early, a common approach is to use text mining techniques to retrieve new G-D associations, examples are BeFree [138] and SemRep [147]. However, text mining techniques do not guarantee high accuracy as the manually created data, they are more likely to include true positives associations together with a large number of false positives. Once retrieved a set of G-D, researchers can assess their relations within the inferred networks or can evaluate their presence within a set of biomarkers generated from a disease-associated dataset.

2.10 Summary

This chapter provided an introduction to the concepts that will be used in this dissertation. Each section described the information that characterises each step of the knowledge extraction process employed in this PhD project. The overall (generic) pipeline is illustrated in Figure 2.16.

The input of this process can be potentially any type of biomedical data described in Section 2.1, as long that they are suitable for a classification problem (supervised learning). A large variety of machine learning algorithms can analyse the data to generate predictive models (decision trees, linear models, ensemble classifiers, etc.). Section 2.2.2.2 described several of them, the work presented in this dissertation was mainly produced using the BioHEL [39] and the random forest [50] classifiers. The generated models are then mined and analysed so that their structure can provide information about the processed data. The extracted knowledge can be exploited for many different research problems, this dissertation is focused on (a) the inference

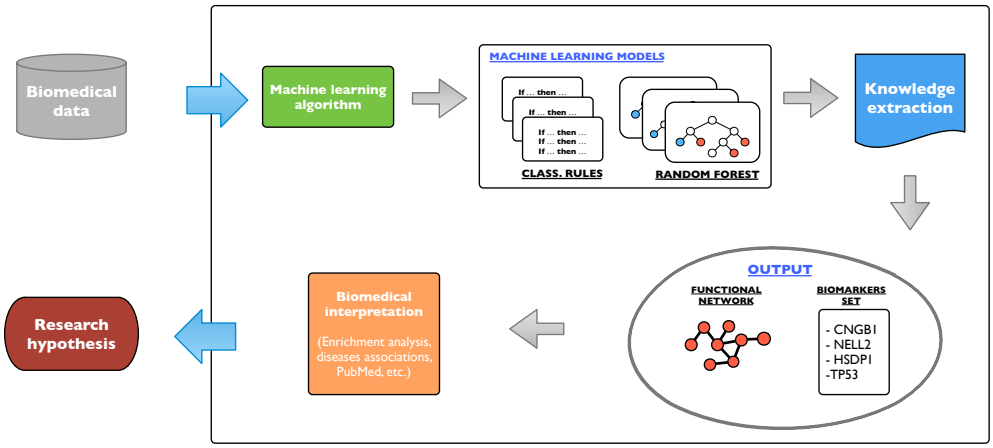


Fig 2.16: A generic pipeline describing the knowledge extraction by the mining of machine learning models inferred from biomedical data.

and the analysis of biological networks and (b) the discovery of small sets of highly predictive biomarkers. The typical approaches used to solve this research questions are presented in Section 2.4–2.3 (networks) and Section 2.6–2.7 (biomarkers). Afterwards, the output of the knowledge discovery process needs to be validated, it is fundamental to assess whether the proposed solutions are relevant in a biomedical context. The enrichment analysis is a traditional method to characterise biological models with established knowledge, the general approach is described in Section 2.9.1. In addition, in this thesis, the proposed models were studied and analysed using both the specialised literature and the disease associations, the latter one is covered in Section 2.9.2. Once the biological validation step (partially) confirms the validity of the new proposed computational solution, new research hypothesis can be formulated and potentially bring to new insights.

3

FuNeL: a protocol for the inference of functional networks from machine learning models

Contents

Abstract

This chapter presents FuNeL, a protocol for the inference of functional networks based on the analysis of rule-based machine learning models. Associations are generated from attributes that collaborate in solving a classification problem. FuNeL is one of the main contributions of this thesis and it represents the first example of how, a smart exploitation of machine learning models, such as classification rules, can generate new knowledge, in this instance in the form of functional networks.

3.1 Introduction

The behaviour of complex biological systems arises from the cooperation of a large number of components. The understanding of how biological events occur at a molecular level is one of the main goals of System Biology and an important effort has been devoted to determine the chain of interactions that controls and mediates biological processes. Networks are the main tool used to characterise and study these complex processes and systems. A biological network is a graph in which nodes represent biological entities such as genes or proteins, and a connection between them indicates a biological relationship, e.g. regulation or common functions. The inference of these networks from biomedical and especially from high-throughput (-omics) data, is an area of intense research.

Most network inference methods focus on the definition of gene regulatory networks, in which edges represent direct regulatory interactions between genes [71, 83, 148]. Far less effort has been put into the design of methods to build functional networks where a connection indicates a functional relationship, e.g. membership in the same pathway, protein complex or sharing the same function. One of the primary uses of functional networks is the identification of functional modules based on the nodes connectivity (subsets of genes with multiple internal connections and a few connections with genes outside the module that describe, explain or predict a biological process or phenotype) [131]. Functional networks are also often employed to identify genes that play a major

role in a specific biomedical context, such as a disease, based on their position in the networks (e.g. hubs).

A commonly adopted approach is to generate functional networks based on the "co-expression" principle [78]. A functional relation (via a direct or indirect interaction) is assumed between two genes when they have similar expression profiles across data (from here comes the name similarity-based methods). It has been demonstrated that co-expression networks can effectively identify pathways and candidate biomarkers [149], or reveal gene modules representing a biological process perturbed in a disease [150], just to name a few examples. The similarity-based approach remains the dominant method of functional network inference today, with many recent examples of successful applications [151–154].

Although similarity-based inference methods have been extensively and successfully used, they detect relationships among genes **only** when similar expression patterns emerge. This limits the range of functional relationships that they reveal [4, 155]. A different approach, to infer biological networks, that is recently gaining popularity, involves the use of machine learning techniques. Due to the wide range of knowledge representations used within machine learning methods (e.g. classification rules, decision trees, artificial neural networks, SVM kernels, etc.), they can discover more complex and diverse relationships and overcome the limitations of the similarity-based methods. This is possible because within machine learning models the attributes are associated not because they are similar (e.g. have similar expression profiles), but because together they detect strong patterns. Moreover, if the learning is supervised, it can take advantage of the additional phenotype information (class labels of the samples, such as case and control) available with the data.

Machine learning can be employed in different ways and forms to solve the task of network inference. One approach is to train machine learning models that directly predict network edges [156]. However, this process requires an experimentally verified "ground truth" of known interactions and suitable controls that represent a challenging task on its own. A different approach is to generate machine learning models from the biological data and then mine the structure of the models to infer networks. Attributes co-operate in machine learning models not only when they are "similar" but when

together meaningful patterns can be extracted. Therefore, such an approach based on the mining of complex machine learning models could possibly lead one to uncover new and different (biological) knowledge, that is likely to escape the traditional similarity-based approaches. Figure 3.1 aims to illustrate the differences between these two approaches: on one hand the similarity-based methods, on the other hand, the methods founded on the knowledge extraction from the machine learning models. The figure highlights how the two approaches differently analyse the same data and how the relationships between the entities are extracted.

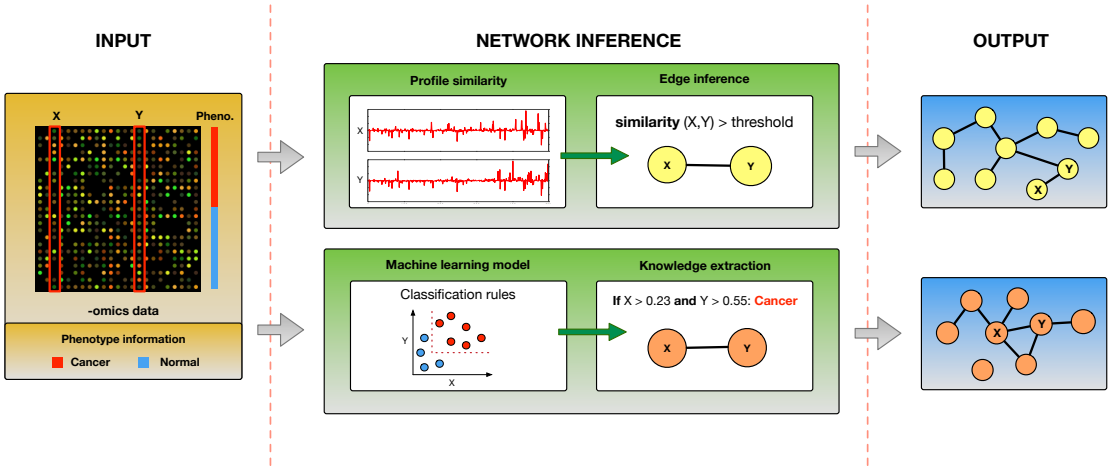


Fig 3.1: Two approaches for the functional network inference: one based on the expression profile similarity and the other based on the extraction of knowledge from machine learning models. The similarity-based methods construct an network edge $X \leftrightarrow Y$ when the similarity between the expressions of genes X and Y across the samples is above a threshold. Methods based on machine learning first build a predictive model, in this example a rule-based model, using the samples phenotype (class labels) information and then construct a network edge $X \leftrightarrow Y$, when genes X and Y are used together within that model to classify the samples. As these two approaches lead to different functional networks, it is possible that they capture complementary knowledge.

As described in Section 2.3, several types of machine learning have been successfully applied to accomplish this task: unsupervised learning in the form of association rules [65], supervised learning using regression (model trees [67, 68]) or classification (random forest [69, 70]).

This chapter proposes, describes and analyses a protocol, called **FuNeL**, for the inference of functional networks based on rule-based machine learning models. FuNeL generates functional networks using: an optional feature selection process to control

the size of the networks, a statistical filtering of the predicted associations between genes using permutation tests and a multi-stage rule-based network inference. The different options available within FuNeL, illustrated in Figure 3.2, generate a total of four protocol configurations.

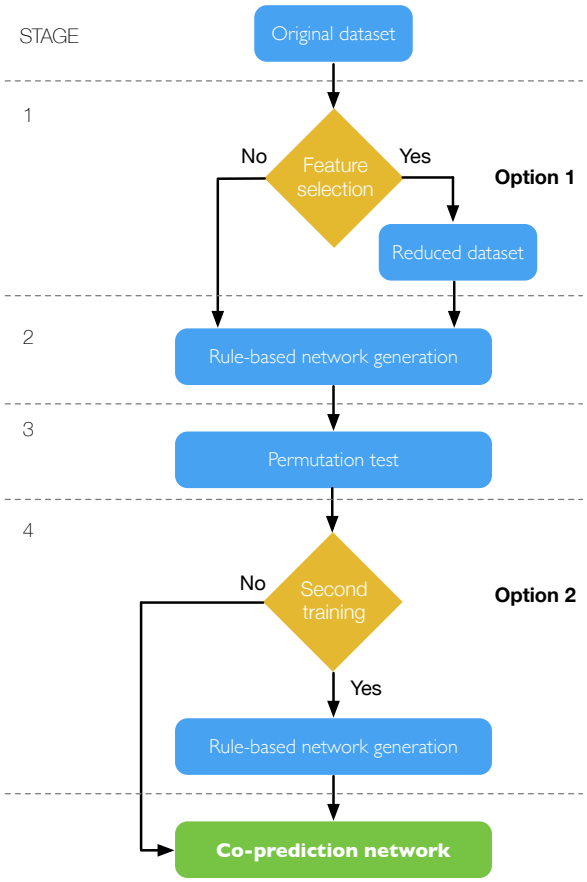


Fig 3.2: The stages of the FuNeL (Functional Network Learning) inference protocol.

In the following sections, firstly FuNeL's ability to correctly identify existing relationships is tested using a set of synthetic datasets. Then, FuNeL is evaluated using eight real-world transcriptomics datasets related to different types of cancer. For each dataset, the four different configurations of the protocol are used to create functional networks. The inferred networks are tested and compared with co-expression networks of equivalent size. To have an extensive evaluation of FuNeL, three different state-of-the-art methods to generate co-expression networks are used. The differences between FuNeL and co-expression networks are assessed from two points of view: (1) the enriched biological terms and (2) the relationships between genes known to be associated with a particular type of cancer. Finally, using a prostate cancer dataset as a case

study, a more detailed biological analysis of the enriched terms and the disease-related genes is performed. The largest hubs and the most central nodes in the prostate cancer co-prediction networks are studied for their involvement in the disease. Literature support is found for the association between these topologically important genes and prostate cancer. This is further confirmed by an independent transcriptomics dataset (not used as a source in the inference process). Overall, the FuNeL inferred networks are shown to (1) capture relevant biological knowledge that is complementary to the knowledge associated with different co-expression networks, and (2) more adequately represent the relationships between genes associated with the disease targeted by each dataset.

3.2 Material and Methods

This section thoroughly describes the proposed FuNeL protocol, the datasets from which the networks were inferred and the experimental design used to evaluate and compare it with co-expression networks created with three different algorithms.

3.2.1 The co-prediction paradigm

Rule-based machine learning models have been successfully applied to extract information from different types of biological data: transcriptomics [157], proteomics [158], lipidomics [159] and protein structure data [160]. In the field of network inference, a new paradigm, based on rule-based machine learning models, was proposed by Bassel et al. [161]. This paradigm, called *co-prediction* (in opposition to the classic co-expression), uses the prediction rules of a classification algorithm to identify relationships between attributes (e.g. genes). Co-prediction is based on the assumption that genes within the same classification rules, due to their co-operation in predicting the sample class, have an increased likelihood of being functionally related to the biomedical process in question (see Figure 3.3). Different than co-expression, the co-prediction approach is employed when solving a classification task in supervised learning. Therefore, it exploits the phenotype information of the data (class labels) to detect functional relations.

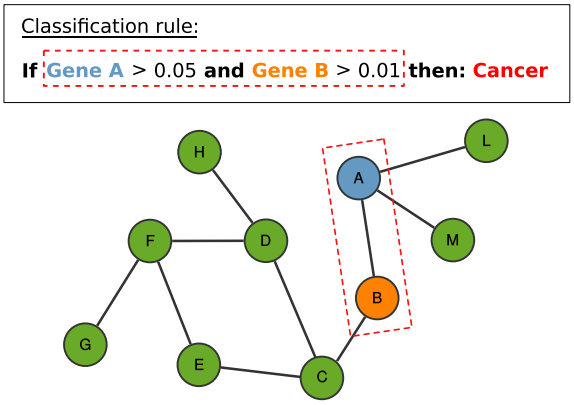


Fig 3.3: The co-prediction paradigm: the association between the genes is inferred from their co-occurrence in classification rules.

In [161], the prediction rules were generated by an evolutionary learning classifier called BioHEL [39]. BioHEL is designed to handle large-scale biological datasets, it generates rules one by one using a mechanism known as separate-and-conquer. BioHEL is enhanced for this type of analysis with a sparse knowledge representation containing a rule-wise embedded feature selection: each rule being generated will only use a very small fraction of attributes. The relevant attributes for a rule are discovered automatically during the learning process and each different rule may use a different subset of attributes due to its stochastic nature. The co-prediction approach was initially tested, in [161], using publicly available gene expression data from Arabidopsis thaliana, for which it was known the seed samples that germinated or not. The functional gene network generated from those data was termed as SCoPNet: Seed Co-Prediction Network. SCoPNet was able to predict functional associations between genes acting in the same developmental and signal transduction pathways irrespective of the similarity in their respective gene expression patterns. Using SCoPNet, four novel regulators of seed germination were identified and experimentally verified. Furthermore, the network was used to predict interactions at the level of transcript abundance between these novel and previously described factors influencing Arabidopsis seed germination.

Lately, other researchers have successfully adopted a similar network inference paradigm by exploiting the classification rules from a Michigan-style learning classifier systems [162, 163]. This approach was applied to SNPs data and was able to identify disease risk factors in a bladder cancer.

3.2.2 The FuNeL protocol

SCoPNet showed that the co-prediction paradigm generates functional networks that incorporate meaningful biological knowledge and can be employed to formulate new research hypothesis. However, from a methodological perspective, many questions remained unanswered about co-prediction:

- Can the co-prediction approach identify known genetic relationships?
- Can the biological significance of the co-prediction networks be quantified?
- What is the impact of data pre-processing on the generated networks?
- Is this methodology able to capture knowledge that escapes other methods?
- Are the discovered functional relationships meaningful in the human disease context?

To address these questions a new network inference protocol called **FuNeL** (Functional Network Learning) is proposed. FuNeL aims to provide a general framework for the inference of functional networks based on the co-prediction paradigm by using rule-based machine learning models. The FuNeL protocol substantially extends the previous work of [161] by incorporating: (1) statistical filtering of inferred functional relationships via permutation tests, (2) a multi-stage network generation to maximise the knowledge extraction, and (3) a configurable feature selection stage to control the size of the generated networks.

Stages of FuNeL

FuNeL is defined by a total of four stages, as illustrated in Figure 3.2. Two of these stages are optional (1 and 4), they lead to a total of four different protocol configurations. If the first optional stage (feature selection) is performed, the original dataset is reduced to the most relevant attributes. In Stage 2, a rule-based machine learning is used to infer a network. This network is statistically refined in Stage 3, in which a permutation test is used to filter out non-significant nodes. The final step, in which

the network generation is repeated for the second time, is again optional. A time complexity analysis of the protocol is available in the Section A.4 of Appendix A.

Feature selection (stage 1) When datasets contain a large number of attributes, a common situation when dealing with biomedical and high-throughput data, some of them might be irrelevant to the prediction of the target. Discard those attributes helps the classification algorithm to focus its learning effort only on the ones that matters. Therefore, the feature selection is the first stage of the inference process. To pick the relevant attributes, FuNeL employs SVM-RFE: a recursive feature elimination method based on Support Vector Machines [111]. The choice fell on SVM-RFE as this algorithm was initially designed to cope with cancer-related transcriptomics data and over the years it showed its potential in identifying relevant features from biomedical data. In FuNeL, SVM-RFE uses an SVM algorithm with a linear kernel as preliminary studies suggested that it can eliminate as much as 90% of the original dataset attributes, without losing much of the classification accuracy. In Figure 3.4 is illustrated the difference in accuracy, calculated using a standard 10-fold cross-validation, when applying SVM-RFE to the datasets (see Section 3.2.3 for a description of the datasets employed) and removing 90% of the original set of attributes. When considering only 10% of the original attributes, the accuracy increased for two datasets, slightly decreased for three datasets and remained unaltered for the remaining datasets.

Rule-based network inference (stage 2) To infer the rule-based classification models, BioHEL was used as in [161]. BioHEL generates sets of classification rules using a genetic algorithm. Figure 3.5 shows an example of a rule set created using BioHEL from a cancer-related dataset. Each rule is sequentially applied to the test set samples, if none of them classifies ("fire") the sample, the default rule is applied. Due to the stochastic nature of BioHEL's learning process, each of its runs generates a different rule set. This fact is leveraged by creating a large number of alternative hypotheses of functional relationships via multiple runs of the algorithm. FuNeL runs BioHEL 10 000 times and infers the network from the *consensus* of all the generated rule sets. To do that, all the pairs of attributes that appear together in the same classification rule are used as the network edges (co-prediction paradigm). Then, each

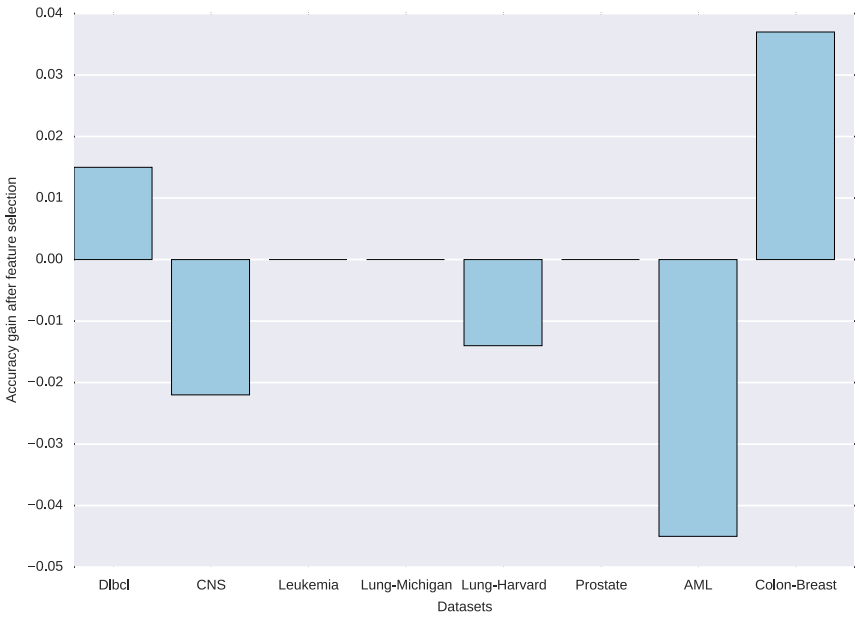


Fig 3.4: Changes in the accuracy (calcualated using a 10-fold cross-validation) when retaining 10% of the original attributes using SVM-RFE.

network node (attribute) receives a score that corresponds to the number of times it has been used in the rules (node score). datasets

Permutation test (stage 3) Given a list of edges (attribute-attribute associations) extracted from the rule sets, FuNeL aims to filter out the non-significant nodes. To determine the node significance, a statistical analysis procedure based on a permutation test, similar to the one described in [162], is followed. 100 permutated datasets are generated by randomly shuffling the class labels. Next, the co-prediction networks are inferred (as in Stage 2) from these permutated datasets. Then, for each node, the distribution of scores across the 100 networks generated from the permutated datasets

If 35742_at **is** > 29.71 **and** 32545_r_at **is** < 1.50 **then** Tumor
If 38322_at **is** > 81.32 **and** 35769_at **is** > 26.24 **then** Tumor
If 34097_at **is** > 25.40 **and** < 88.84 **then** Tumor
If 38602_at **is** < 35.01 **then** Tumor
If 40853_at **is** > 11.11 **then** Tumor
Default rule: Normal

Fig 3.5: Example of a BioHEL classification rule set.

is calculated. Using a one-tailed permutation test, a p-value is assigned to each node, to estimate how likely it is to draw its score from the calculated distribution. With this process it is sure that the nodes with high scores are really tied to the classes present in the data, and that the network truly represents functional relationships. To decide if a node is statistically significant a typical $\alpha = 0.05$ threshold is employed.

Preliminary experiments showed that using significant nodes alone leads to small and dense networks. By having networks that are highly dense, where each node tends to be connected to every other node, the reliability on the meaning of the functional relationships is lost. To counter that, the node pruning is relaxed to keep all direct neighbours of the significant nodes, so that larger and less dense networks can be created.

Network construction (stage 4) There are two ways to interpret the result of the statistical test (option 2 in Figure 3.2). The first is to use the significant nodes as a filter for the inferred relationships (edges) and remove all the edges between two non-significant nodes. The second approach is to use the permutation test as a further feature selection and build a new rule-based machine learning model using only the significant nodes. This second run of the learning algorithm is then focused only on the statistically important genes and creates the final network.

As a result of two independent optional stages in the FuNeL protocol, there are four different configurations that it can run with (see Table 3.1).

Config.	**Description**
C_1	reduced dataset + 1 stage of network generation
C_2	original dataset + 1 stage of network generation
C_3	reduced dataset + 2 stages of network generation
C_4	original dataset + 2 stages of network generation

Table 3.1: Summary of the FuNeL configurations used in the experiments.

3.2.3 Datasets

Synthetic datasets

To verify if FuNeL is able to correctly identify functional relationships, a set of synthetic datasets were used. Although there are several generators that model expression data with genetic relationships, such as GNW used for several DREAM challenges [120], they generate unlabelled samples (without phenotype information, e.g. case vs. control). Unfortunately, the class labels are necessary to perform the supervised learning at the core of FuNeL.

For this reason, the synthetic data were created with the GAMETES instance generator [164], designed to create genetic datasets with multi-locus disease associations, where no fewer than n loci can predict a phenotype (disease status). GAMETES generates genotype data (SNPs array) based on models with specific genetic constraints (e.g. different heritabilities or frequencies of the SNPs).

To generate the synthetic datasets, a set of 2-locus configurations was used. This choice is similar to what employed in the work of Li et al. [70] to evaluate a permutation-based random forest method for the detection of gene interactions . Specifically, the genetic models varied in terms of heritability (0.001-0.4) and number of attributes (5-25), with fixed allele frequency of 0.2 and constant 2 000 samples per dataset equally distributed between the two classes. For each configuration, from 100 000 random models, two models were selected with extreme values of the ease of detection metric (EDM) (the least and the most difficult). Finally, for each selected model, 50 different datasets (GAMETES has a stochastic nature) were generated, obtaining a total of 4 000 datasets.

Real-world datasets

FuNeL was also tested and evaluated using eight publicly available human cancer microarray datasets (see Table 3.2). These datasets represent a broad range of characteristics in terms of biological information (different types of cancers), number of samples (patients or tissues) and attributes (genes). For each dataset, the attributes were defined by the probes used in the microarray experiment. Generally, a gene can

be represented by more than one probe, thus an extra post-processing step is needed to merge the information and generate networks where nodes truly represent genes. The mapping from the probe IDs to the HUGO gene IDs was done using MADGene [165]. All the probes mapped to the same gene were merged into the same node, if a probe was unmapped, it was removed from the network.

Name	Attributes	Samples	Class labels
Dlbcl [166]	2647	77	Dlbcl; Follicular lymphoma
CNS [167]	7129	60	Survivor; Failures
Leukemia [94]	7129	72	AML; ALL
Lung-Michigan [168]	7129	96	Tumor; Normal
Lung-Harvard [169]	12534	181	Mesothelioma; ADCA
Prostate [170]	12600	102	Tumor; Normal
AML [171]	12625	54	Remission; Relapse
Colon-Breast [172]	22283	52	Colon cancer; Breast cancer

Table 3.2: Description of the datasets used to infer networks.

3.2.4 Co-expression networks

One of the main aims of the analysis performed for this part of the dissertation was to compare FuNeL, a machine learning approach, with the co-expression approach, the state-of-the-art in terms of biological network generation. The co-expression paradigm identifies similarity of gene expression pattern under different experimental conditions. Co-expression edges are an abstraction of functional relationships between genes and do not represent physical binding as in protein interaction or gene regulatory networks. Two genes are considered to be functionally related (co-expressed) if their transcript levels are similar across the samples.

Three well-known co-expression network inference methods were compared to FuNeL, each one uses a different metric to assess gene expressions similarity: Pearson Correlation Coefficient, ARACNE [83] and MIC [87]. The methods are extensively described in Section 2.4.

Pearson Correlation Coefficient (PCC) is a well-known measure of linear dependence between two variables. When applied to gene expression profiles, PCC measures

the similarity in the direction of genes' response across samples. Its main disadvantages are the lack of distributional robustness (it assumes data normality) and the sensitivity to outliers. The PCC-based co-expression networks were generated using the *SciPy* Python library [173].

ARACNE The ARACNE method measures the dependence between two gene expression profiles using mutual information [83]. Given two random variables X and Y, the mutual information $MI(X;Y)$ estimates the entropy to quantify the amount of information that Y contains about X. In opposition to correlation, mutual information is able to detect non-linear dependencies. For every pair of genes X and Y, ARACNE calculates $MI(X;Y)$ using their expression profiles and applies a filtering method, called the data processing inequality, to get rid of the majority of indirect dependencies. The ARACNE based co-expression networks were generated with the *minet* R package [174]. The networks were inferred using the following parameter: $estimator = mi.empirical$, $dis = equalwidth$ and $\epsilon = 0$.

Maximal Information Coefficient (MIC) is a recently proposed measure of the strength of association between two variables that is strictly connected to mutual information [87]. Instead of using a single discretisation strategy to bin the compared variables, it chooses individual bins for each variable, such that the value of mutual information $MI(X;Y)$ is maximised. Different than the standard estimation of $MI(X;Y)$ value used in ARACNE, the optimised estimation implemented by MIC can identify a larger set of non-linear associations. MIC based co-expression networks were generated using the *minepy* Python library [88] with the following parameters: $\alpha = 0.6$ (the exponent of the search-grid size that that partitions the data to encapsulate the relationship between the two variables) and $c = 15$ (a number determining the starting point of the X-by-Y search-grid).

Inference of the co-expression networks counterparts To fairly compare the co-prediction (FuNeL) and the co-expression networks generated from the same data, they had to match in size. To accomplish this task, for every co-prediction network C defined by m edges and n nodes, two co-expression counterparts were generated:

- $SE(C)$: co-expression network with m edges
- $SN(C)$: co-expression network with n nodes

PCC and MIC directly compute the pairwise similarity between the genes' expressions. Given that, $SE(C)$ was generated using the top m gene pairs with the highest similarity coefficient. To build $SN(C)$, as many top gene pairs as needed were used, to reach at least n nodes (as all pairs tied on the similarity value were included, sometimes the resulting networks had few nodes more).

ARACNE uses a pruning procedure and directly generates a weighted network, not a list of pairwise similarities. When the resulting network had less than m edges or n nodes, the default tolerance threshold ϵ was increased to obtain a large enough network. This was the case for the *CNS* ($\epsilon = 0.002$) and the *Dlbcl* datasets ($\epsilon = 0.043$). Then, the edge weights were used to select the top gene pairs, as in the case of PCC and MIC.

3.2.5 Enrichment analysis

An enrichment analysis was conducted to understand the biological information captured by the generated networks. The enrichment analysis is a statistical-based procedure for checking whether a set of genes have common characteristics. In this study, the set of genes was defined by the nodes of the generated functional network. PANTHER [175] was the tool employed to perform the enrichment analysis. Because many statistical tests are performed (one for each term) at the same time, PANTHER was set to use a Bonferroni correction for multiple testing with $\alpha = 0.05$. Two categories of biological knowledge were considered: Gene Ontology (GO) terms and PANTHER pathways (176 primarily signalling pathways). From the set of GO terms, only the manually curated annotations, supported by experimental evidence, were selected.

3.2.6 Disease association analysis

To evaluate the predictive power of the generated networks, and to assess their relevance within a cancer-related context, the relationships between known disease-associated genes (G-D) were analysed. Two sources for the disease associations were

employed: Malacards (a meta-database of human maladies consolidated from 64 independent sources) [139] and the union of four manually curated databases (OMIM [143], Orphanet [146], Uniprot [145] and CTD [144]). Two properties have been analysed: (1) the proximity of the disease-associated genes within a network and (2) the number of triangles in a network, containing one or more disease-associated genes.

Higher proximity indicates a stronger functional relationship between genes involved in the disease. Triangles represent groups of attributes used together across different prediction rules, and therefore indicate a strong mutual relationship between the genes (useful in the discovery of potential new G-D associations). Triangles are also the smallest non-trivial motifs that can be found in a complex network and often identify functional units of biological processes in cells [176].

The proximity of disease-associated genes was measured using the average Shortest Path Length (SPL). The proximity was defined as a ratio of two distances: average SPL between all pairs of the non-associated genes and average SPL between all pairs of disease-associated genes A:

$$\frac{1}{n}\sum_{i=1}^{n} w_i \frac{\overline{SPL}(CC_i \setminus A)}{\overline{SPL}(A)}, \text{where } w_i = \frac{|CC_i|}{\sum_{j=1}^{n}|CC_j|}$$

As the generated networks often were disconnected (had more than 1 connected component), a weight w_i that represents the relative size of a connected component CC_i was introduced (bigger components have more impact). Components with less than three nodes or disease-associated genes were not used in the calculation.

3.3 Results

This section presents the analysis performed on the FuNeL functional networks generated from both real-world and synthetic datasets. First, the ability of FuNeL networks to identify predefined relationships between attributes is tested using synthetics datasets. Then, the FuNeL and the co-expression networks are compared from both a topological and a biological point of view. Finally, using a prostate cancer dataset as a case study, the relevance of the biomedical knowledge captured by FuNeL networks is assessed.

3.3.1 Identification of predefined relationships in synthetic datasets

At first, it was assessed if FuNeL networks can capture meaningful associations between biological entities. This was performed using synthetic datasets generated with GAMETES. GAMETES is an instance generator that creates synthetic genomic datasets with multi-locus disease associations that have no fewer than n-loci predicting a phenotype. By using a set of 2-locus configurations, GAMETES created datasets in which the association between two attributes (SNPs) is responsible for the disease status (class label) of the samples. Those synthetic data were used to assess if FuNeL networks capture the existing relationships between the attributes that determine the phenotype of each sample. In total, 80 different model configurations were used, varying in terms of heritability (proportion of variation that can be ascribed to the attributes of the models, low heritability means higher noise), number of SNPs (attributes) and ease of detection, and tested the success rate on 50 datasets per model. Given the small number of attributes available in the synthetic datasets, only the C_2 configuration was used in the tests (no feature selection, single learning phase). The percentage of successfully identified relationships for each model (across the 50 datasets) is reported in Table 3.3. A success was defined by the presence of an edge between the interacting pair of SNPs in the inferred network.

	5 SNP		**10 SNP**		**15 SNP**		**20 SNP**		**25 SNP**	
Her.	**L-EDM**	**H-EDM**	**L-EDM**	**H-EDM**	**L-EDM**	**H-EDM**	**L-EDM**	**H-EDM**	**L-EDM**	**H-EDM**
0.001	6 %	16 %	8 %	18 %	4 %	10 %	4 %	12 %	12 %	16 %
0.005	8 %	82 %	0 %	86 %	6 %	80 %	2 %	82 %	8 %	72 %
0.01	8 %	96 %	8 %	100 %	8 %	100 %	12 %	100 %	14 %	100 %
0.05	14 %	100 %	60 %	100 %	42 %	100 %	34 %	100 %	34 %	100 %
0.1	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
0.2	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
0.3	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
0.4	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %

Table 3.3: FuNeL success rate in the identification of disease-predicting SNPs. The datasets differed with respect to heritability, number of SNPs and detection difficulty (L-EDM models were the hardest, H-EDM the easiest).

As expected, a higher success rate was obtained for models where the relationships were easy to detect (H-EDM). The performance increased with higher values of heritability and 100% success rate was obtained for heritability values above 0.05, regardless of

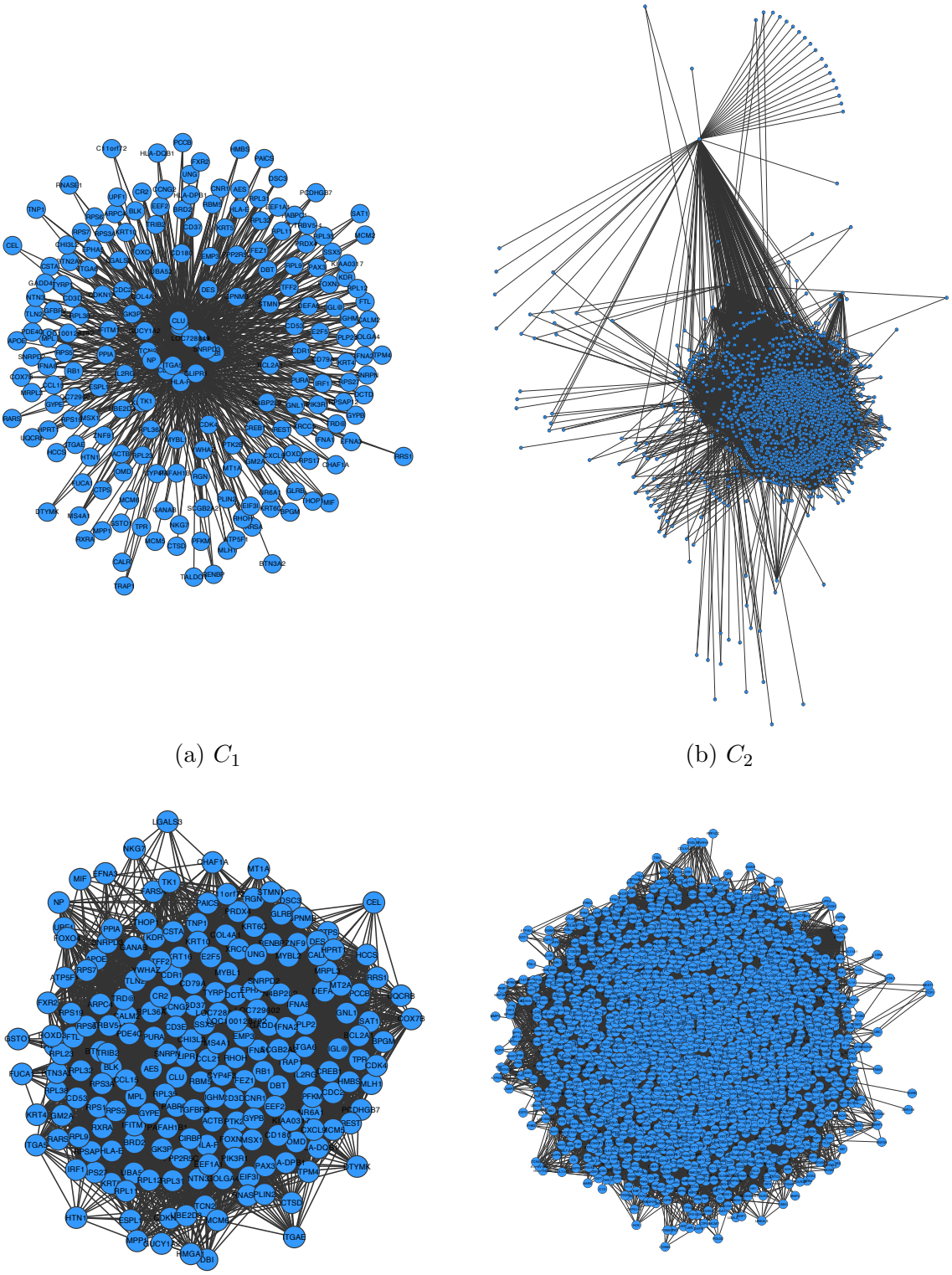
model difficulty. The overall results are similar to those reported by Li el al. [70], or even slightly better, as FuNeL's success rate was unaffected by the increase in the number of SNPs as in the presented permutated random forest approach.

3.3.2 Topological comparison of the inferred networks

Comparison of FuNeL networks

The network topology refers to the spatial arrangements of its elements. The analysis of the topological properties tells us how different nodes are connected to each other and how their communication paths look like. There are many aspects and characteristics that can be evaluated in a network, this analysis was focused on four main measures: number of nodes, number of edges, clustering coefficient and diameter. The *clustering coefficient* is a measure of the degree to which nodes in a network tend to cluster together. It expresses the likelihood that any two nodes with a common neighbour are themselves connected. The *diameter* indicates the maximum distance between two nodes in the network.

Different FuNeL configurations infer networks that are topologically different. In Figure 3.6 are shown the networks generated with different configurations using the *Dlbcl* dataset [166]. All the networks have been visualised with Cytoscape [177] and using the same layout (Organic layout). At first sight, differences can be noticed in how the nodes are connected. For a more detailed evaluation, the main topological measures mentioned earlier were calculated. A summary, dataset by dataset is reported in Table 3.4. In the Dlbcl networks, C_1 and C_3 had the same diameter (5), while C_2 and C_4 resulted more compact and had a smaller diameter equal to 3. Among the four networks, C_1 had by far the highest clustering coefficient, 0.872, almost twice the value of C_4. In general, when analysing all the FuNeL networks generated with the real-world datasets, as expected, the configurations having feature selection (C_1 and C_3) brought to networks with a smaller number of nodes than when the original set of attributes is used (C_2 and C_4). Furthermore, the second phase of machine learning modelling (C_3 and C_4) tends to reduce the number of nodes as it uses a reduced set of attributes as input (only significant nodes and their neighbours from the first training phase) while increasing both clustering coefficient and the number of edges.



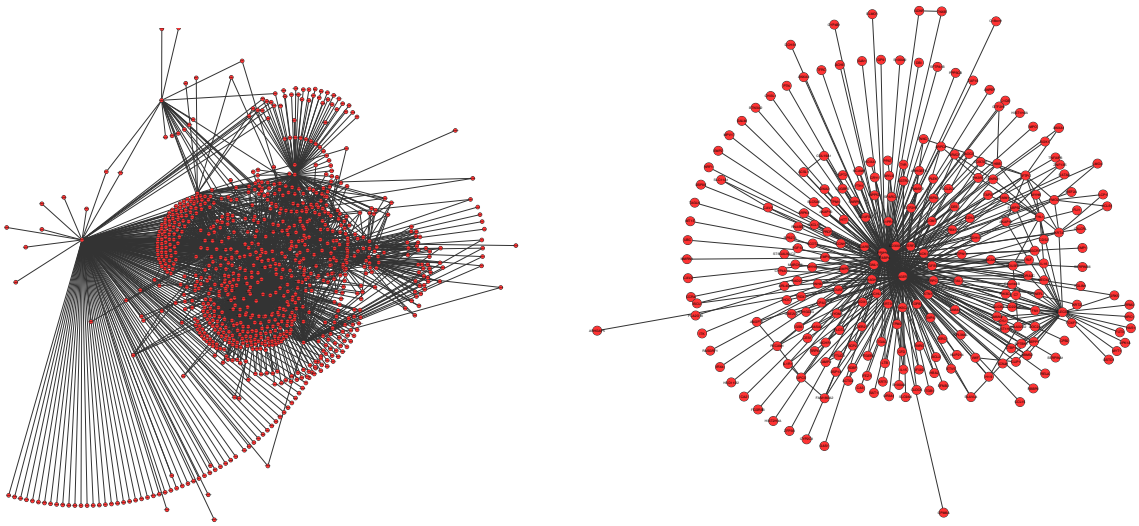
(a) C_1 (b) C_2

(c) C_3 (d) C_4

Fig 3.6: FuNeL networks generated from the Dlbcl dataset.

		FuNeL				Co-expression (SE)				Co-expression (SN)			
Dataset	**Cat.**	C_1	C_2	C_3	C_4	$SE(C_1)$	$SE(C_2)$	$SE(C_3)$	$SE(C_4)$	$SN(C_1)$	$SN(C_2)$	$SN(C_3)$	$SN(C_4)$
Leukemia	**Nodes**	421	1480	294	988	683	873	843	941	422	1482	293	979
	Edges	1529	2294	2154	2646	1529	2294	2154	2646	680	7145	409	2870
	Clust.Coef.	0.712	0.155	0.589	0.33	0.333	0.348	0.354	0.341	0.323	0.388	0.303	0.344
	Diameter	5	6	4	6	24	18	19	22	16	18	9	20
LungH	**Nodes**	429	1419	382	1030	578	930	955	1214	432	1413	384	1027
	Edges	1068	2317	2398	3410	1068	2317	2398	3410	617	4302	476	2650
	Clust.Coef.	0.344	0.298	0.43	0.404	0.356	0.373	0.376	0.372	0.341	0.386	0.296	0.376
	Diameter	5	8	5	7	10	23	23	21	6	22	6	23
LungM	**Nodes**	91	919	48	247	76	280	59	119	90	915	50	248
	Edges	134	1858	78	410	134	1858	78	410	224	13574	64	1510
	Clust.Coef.	0.379	0.262	0.418	0.457	0.465	0.525	0.446	0.514	0.539	0.523	0.493	0.511
	Diameter	3	5	3	3	6	11	6	6	5	14	6	12
CNS	**Nodes**	501	4257	494	3538	945	2152	1616	2607	501	4261	488	3532
	Edges	4302	25069	12769	40840	4302	25069	12769	40840	1553	171052	1502	90395
	Clust.Coef.	0.743	0.255	0.521	0.302	0.354	0.389	0.367	0.400	0.346	0.427	0.35	0.421
	Diameter	4	7	4	6	21	15	23	13	12	13	14	12
Dlbcl	**Nodes**	201	1699	201	1617	207	1411	1238	1790	200	1699	200	1614
	Edges	848	10471	7351	33170	848	10471	7351	33170	832	24280	832	17865
	Clust.Coef.	0.872	0.574	0.642	0.453	0.508	0.438	0.411	0.51	0.501	0.504	0.501	0.481
	Diameter	3	5	3	5	2	16	17	14	2	14	2	15
GSE2191	**Nodes**	890	4802	846	3561	837	1848	1239	1750	897	4799	839	3553
	Edges	3290	13424	6469	12074	3290	13424	6469	12074	3711	90410	3292	47806
	Clust.Coef.	0.488	0.082	0.317	0.291	0.377	0.409	0.394	0.4	0.382	0.415	0.377	0.417
	Diameter	5	9	5	9	25	23	19	21	21	13	25	15
GS3726	**Nodes**	668	2077	524	1170	879	1300	992	1367	759	2300	573	1170
	Edges	1761	3255	2051	3502	1761	3255	2051	3502	1471	9808	1584	3469
	Clust.Coef.	0.134	0.0077	0.307	0.109	0.226	0.23	0.223	0.233	0.213	0.287	0.254	0.346
	Diameter	8	10	7	8	20	26	20	25	26	15	19	15
Prostate	**Nodes**	938	4290	704	2277	356	543	322	448	920	4298	702	2287
	Edges	3796	10175	3090	6546	3796	10175	3090	6546	33250	914829	16934	24427
	Clust.Coef.	0.328	0.245	0.29	0.25	0.565	0.607	0.541	0.584	0.655	0.703	0.641	0.711
	Diameter	7	10	6	8	6	9	5	8	8	11	7	12

Table 3.4: Topological properties of FuNeL and Pearson co-expression networks.



(a) C_2 (b) C_4

Fig 3.7: FuNeL networks generated from the Lung-Michigan dataset.

Comparison of FuNeL and co-expression networks

FuNeL and co-expression networks were first compared from a topological point of view. The topological properties described in the previous section were calculated

		FuNeL				Co-expression (SE)				Co-expression (SN)			
Dataset	**Cat.**	C_1	C_2	C_3	C_4	$SE(C_1)$	$SE(C_2)$	$SE(C_3)$	$SE(C_4)$	$SN(C_1)$	$SN(C_2)$	$SN(C_3)$	$SN(C_4)$
Leukemia	**Nodes**	421	1480	294	988	1024	1426	1356	1577	422	1480	294	989
	Edges	1529	2294	2154	2646	1529	2294	2154	2646	512	2416	327	1479
	Clust.Coef.	0.712	0.155	0.589	0.330	0.002	0.002	0.002	0.002	0.000	0.002	0.000	0.002
	Diameter	5	6	4	6	17	19	22	19	9	17	11	19
LungH	**Nodes**	429	1419	382	1030	907	1614	1653	2066	429	1419	382	1030
	Edges	1068	2317	2398	3410	1068	2317	2398	3410	435	1924	375	1250
	Clust.Coef.	0.344	0.298	0.430	0.404	0.007	0.006	0.006	0.005	0.013	0.006	0.012	0.007
	Diameter	5	8	5	7	23	16	15	13	14	18	10	18
LungM	**Nodes**	91	919	48	247	143	1321	96	370	91	920	48	247
	Edges	134	1858	78	410	134	1858	78	410	72	1127	34	259
	Clust.Coef.	0.379	0.262	0.418	0.475	0.000	0.002	0.000	0.009	0.000	0.005	0.000	0.014
	Diameter	3	5	3	3	13	17	11	18	11	17	5	11
CNS	**Nodes**	501	4257	494	3538	2002	4509	3581	5342	502	4257	494	3538
	Edges	4302	25069	12769	40840	4302	25069	12769	41661	513	20409	505	12358
	Clust.Coef.	0.743	0.255	0.521	0.302	0.004	0.005	0.006	0.026	0.004	0.005	0.004	0.005
	Diameter	4	7	4	6	12	8	12	7	20	9	20	12
Dlbcl	**Nodes**	201	1699	201	1617	380	1452	1191	2236	201	1699	201	1617
	Edges	848	10471	7351	33170	848	10471	7351	33890	269	14149	269	12903
	Clust.Coef.	0.872	0.574	0.642	0.453	0.136	0.126	0.140	0.176	0.113	0.110	0.113	0.115
	Diameter	3	5	3	5	13	9	11	5	12	8	12	9
GSE2191	**Nodes**	890	4802	846	3561	2574	5226	3846	5027	890	4802	846	3561
	Edges	3290	13424	6469	12074	3290	13424	6469	12076	846	10671	794	5564
	Clust.Coef.	0.488	0.082	0.317	0.291	0.002	0.002	0.002	0.002	0.004	0.002	0.004	0.002
	Diameter	5	9	5	9	19	13	16	13	30	15	30	17
GS3726	**Nodes**	668	2077	524	1170	1362	2167	1546	2279	668	2166	524	1170
	Edges	1761	3255	2051	3502	1761	3255	2051	3502	787	3250	597	1455
	Clust.Coef.	0.134	0.077	0.307	0.109	0.024	0.053	0.029	0.050	0.014	0.053	0.016	0.021
	Diameter	8	10	7	8	20	20	19	18	15	20	13	19
Prostate	**Nodes**	938	4290	704	2277	2760	6805	2268	4575	939	4290	704	2277
	Edges	3796	10175	3090	6546	3796	10175	3090	6546	1300	6095	1017	3102
	Clust.Coef.	0.328	0.245	0.290	0.250	0.005	0.003	0.006	0.003	0.001	0.003	0.002	0.005
	Diameter	7	10	6	8	13	13	15	13	12	13	9	15

Table 3.5: Topological properties of FuNeL and ARACNE co-expression networks.

for all the networks. In Figures 3.7 – 3.10 are shown the networks generated using the different inference approaches from the Lung-Michigan dataset [168]. To allow a fair visual comparison, all the networks have been depicted using the same (Organic) layout. The topological measures for each method and dataset by dataset are reported in Table 3.4 (PCC), Table 3.5 (ARACNE) and Table 3.6 (MIC)

When contrasting FuNeL and co-expression networks, the *SE* counterparts have in general more nodes. On the other hand, *SN* networks differ according to the inference method used. In fact, ARACNE generated *SN* counterparts with fewer edges, while this is true only for $SN(C_1)$ and $SN(C_3)$ inferred with MIC and Pearson. The clustering coefficient is constantly lower in ARACNE networks than in FuNeL, this is probably due to the pruning phase operated by the method. A similar trend can be noticed for MIC networks with some exceptions (e.g. Prostate $SN(C2)$ and $SN(C_4)$). When FuNeL is contrasted with the Pearson Correlation Coefficient, the networks generated with feature selection (C_1 and C_3) have a lower coefficient than their co-

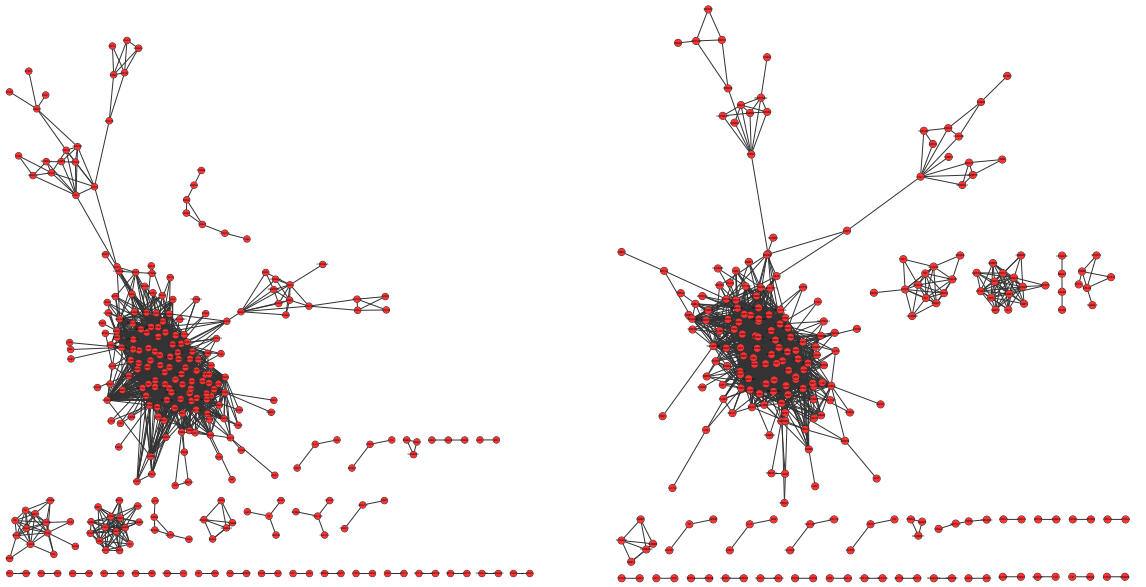
		FuNeL				Co-expression (SE)				Co-expression (SN)			
Dataset	**Cat.**	C_1	C_2	C_3	C_4	$SE(C_1)$	$SE(C_2)$	$SE(C_3)$	$SE(C_4)$	$SN(C_1)$	$SN(C_2)$	$SN(C_3)$	$SN(C_4)$
Leukemia	**Nodes**	421	1480	294	988	640	807	780	896	421	1480	294	989
	Edges	1529	2294	2154	2646	1529	2294	2155	2647	749	6173	432	3096
	Clust.Coef.	0.712	0.155	0.589	0.330	0.162	0.182	0.180	0.179	0.138	0.180	0.127	0.175
	Diameter	5	6	4	6	18	29	27	29	10	17	8	18
LungH	**Nodes**	429	1419	382	1030	384	685	703	944	429	1419	382	1030
	Edges	1068	2317	2398	3410	1068	2317	2399	3410	1264	5867	1045	3841
	Clust.Coef.	0.344	0.298	0.430	0.404	0.349	0.308	0.305	0.302	0.339	0.282	0.343	0.305
	Diameter	5	8	5	7	9	13	13	17	7	18	9	19
LungM	**Nodes**	91	919	48	247	118	626	79	219	91	919	48	247
	Edges	134	1858	78	410	134	1858	78	410	93	3109	38	484
	Clust.Coef.	0.379	0.262	0.418	0.475	0.212	0.272	0.213	0.306	0.208	0.235	0.153	0.302
	Diameter	3	5	3	3	8	18	7	8	6	14	3	7
CNS	**Nodes**	501	4257	494	3538	1424	3104	2357	3725	501	4257	495	3538
	Edges	4302	25069	12769	40840	4305	25131	12771	40850	704	62208	694	36027
	Clust.Coef.	0.743	0.255	0.521	0.302	0.124	0.154	0.144	0.159	0.089	0.162	0.091	0.161
	Diameter	4	7	4	6	17	11	12	10	11	10	11	10
Dlbcl	**Nodes**	201	1699	201	1617	475	1140	1047	1453	203	1699	203	1617
	Edges	848	10471	7351	33170	848	10471	7362	33172	196	74773	196	59307
	Clust.Coef.	0.872	0.574	0.642	0.453	0.111	0.240	0.219	0.319	0.082	0.381	0.082	0.366
	Diameter	3	5	3	5	21	13	16	15	11	11	11	11
GSE2191	**Nodes**	890	4802	846	3561	1700	4129	2617	3883	890	4803	846	3563
	Edges	3290	13424	6469	12074	3299	13433	6469	12207	1380	17797	1271	10540
	Clust.Coef.	0.488	0.082	0.317	0.291	0.109	0.095	0.098	0.098	0.120	0.095	0.118	0.099
	Diameter	5	9	5	9	22	15	18	16	19	15	21	15
GS3726	**Nodes**	668	2077	524	1170	1271	1921	1357	1996	672	2152	526	1172
	Edges	1761	3255	2051	3502	1890	3261	2056	3524	852	3921	538	1705
	Clust.Coef.	0.134	0.077	0.307	0.109	0.110	0.100	0.104	0.100	0.126	0.099	0.121	0.117
	Diameter	8	10	7	8	23	29	23	28	14	24	14	24
Prostate	**Nodes**	938	4290	704	2277	687	839	667	773	964	4290	712	2277
	Edges	3796	10175	3090	6546	3981	10186	3777	8257	15928	1763794	5254	308709
	Clust.Coef.	0.328	0.245	0.290	0.250	0.167	0.278	0.169	0.265	0.313	0.758	0.218	0.661
	Diameter	7	10	6	8	7	8	7	8	7	8	7	9

Table 3.6: Topological properties of FuNeL and MIC co-expression networks.

expression counterparts. Finally, a clear pattern emerges when analysing the diameter of the networks. Co-prediction networks are more compact than co-expression counterparts having up to 3 times lower diameter for MIC and Pearson and up to 7 times lower for ARACNE.

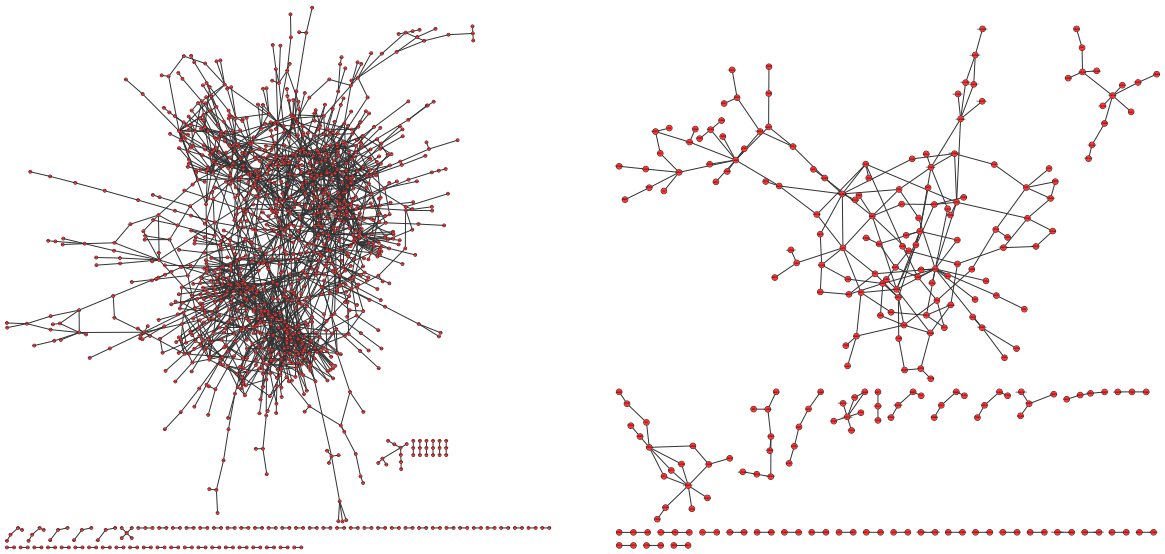
3.3.3 Complementarity of enriched terms

The results described in Section 3.3.2, illustrate that different FuNeL configurations lead to networks with different topological properties. Next, the different configurations were evaluated to assess if they generate networks capturing different biological knowledge. This evaluation was based on the analysis of the enriched terms associated with the nodes of each network. To test how unique the biological terms (GO terms and pathways) over-represented in the inferred FuNeL networks are, the overlap between terms of networks created with different configurations was measured. Given



(a) $SE(C_2)$ (b) $SN(C_4)$

Fig 3.8: Pearson co-expression networks generated from the Lung-Michigan dataset.



(a) $SE(C_2)$ (b) $SN(C_4)$

Fig 3.9: ARACNE co-expression networks generated from the Lung-Michigan dataset.

two networks generated with the configurations C_a and C_b, the pairwise terms overlap was calculated using the Jaccard similarity coefficient:

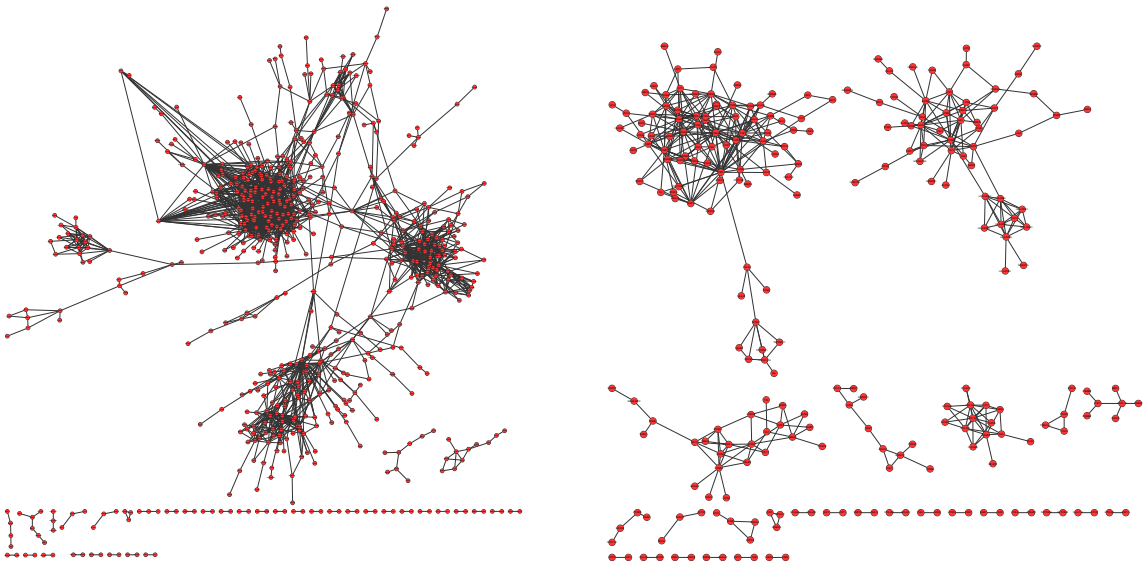
$$O(C_a, C_b) = \frac{c}{u_a + u_b + c}$$

where c is the number common terms, u_a is the number of unique terms for C_a and u_b is the number of unique terms for C_b.

	Gene Ontology				**Pathways**			
	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4
C_1	—	35.3	74.9	40.5	—	18.6	51.3	18.3
C_2		—	32.1	70.1		—	9.5	59.1
C_3			—	36.4			—	10.4
C_4				—				—

Table 3.7: Average overlap of enriched GO terms and pathways between different FuNeL configurations.

Table 3.7 summaries the pairwise overlap between the 4 different FuNeL configurations. For GO terms is reported the average overlap of the: biological process, cellular component and molecular function categories. Although configurations that operate on the same dataset (C_1–C_3 and C_2–C_4) shared the most terms/pathways, in general the overlap is quite far from 100%. Thus, the remaining difference is a result of the second training stage. Configurations used on different datasets (i.e. different set of attributes) resulted in networks sharing less than 40% GO terms and 20% pathways. Overall, the values in Table 3.7 suggest that different steps of the FuNeL protocol results in networks carrying diverse biological knowledge.



(a) $SE(C_2)$ (b) $SN(C_4)$

Fig 3.10: MIC co-expression networks generated from the Lung-Michigan dataset.

Similarly, the term overlap was calculated between co-prediction and co-expression by comparing the C_i networks with their co-expression counterparts $SE(C_i)$ and $SN(C_i)$ (see Table 3.8). The percentages were similar across the different inference methods. The overlap in enriched terms was never higher than 62% (still leading to a difference of terms around 40%) and was the largest for configuration not using feature selection (C_2 and C_4). In general, the values were lower for the biological pathways with a minimum of a mere 2% of shared terms. Low values of terms overlap indicate that the co-prediction and the co-expression approaches can be seen as complementary. Despite starting from the same dataset, they generate networks expressing different biological information.

		Co-expression (SE)				**Co-expression (SN)**			
Method	**Cat.**	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4
Pearson	**GO**	28.0	41.4	29.7	43.2	31.5	57.6	36.7	48.8
	path.	22.3	26.0	25.8	19.0	26.4	40.0	17.5	28.7
ARACNE	**GO**	34.8	62.1	27.2	56.5	33.3	61.2	27.7	53.5
	path.	12.6	46.3	13.9	47.9	8.5	42.3	1.6	35.6
MIC	**GO**	31.6	51.3	28.3	48.7	30.0	61.4	28.9	52.7
	path.	9.7	33.9	14.2	31.5	11.2	46.9	8.0	35.2

Table 3.8: Average overlap of enriched GO terms and pathways between the FuNeL and co-expression networks. Each co-expression network C_i was compared to the corresponding co-expression networks $SE(C_i)$ and $SN(C_i)$.

Random networks as a null for comparison

When looking at the reduced overlap of enriched terms between co-prediction and co-expression networks, a legitimate doubt might arise: could the lack of overlap be due because co-prediction is not capturing useful information? To tackle this question, together with additional analysis presented later on, a set of experiments was performed using randomly generated networks. For each dataset and each FuNeL network, 100 same-size random networks were created by swapping the genes (nodes) of the co-prediction network with randomly selected genes from the original set of genes. Then, an enrichment analysis was conducted on the networks using the same PAN-

Dataset	C_1	C_2	C_3	C_4
Prostate	95.2±85.4	91.2±81.2	77.8±68.4	110.9±101.0
Leukemia	80.9±77.6	15.5±31.2	84.9±83.6	1.6±33.1
Lung-Michigan	59.0±73.3	1.1±5.5	60.0±60.9	10.1±15.8
Dlbcl	74.0±73.5	40.3±39.3	75.6±77.7	38.2±39.1
AML	63.8±71.3	96.6±101.8	100.7±96.3	93.0±101.0
Colon-Breast	67.3±69.9	40.9±39.6	93.9±88.6	43.8±42.1
Lung-Harvard	88.8±95.7	45.8±42.4	148.5±165.3	43.2±44.0
CNS	40.9±48.5	18.0±32.8	3.4±28.2	13.9±31.5

Table 3.9: Average difference in enriched GO terms between FuNeL and random networks. The values have been averaged across three GO categories (biological process, cellular component and molecular function)

THER settings [1]. The goal was to check how the terms overlap between co-prediction and random networks changes in relation to the overlap between co-prediction and co-expression networks. If FuNeL networks capture irrelevant knowledge, their overlap with random networks should appear higher than the overlap with co-expression networks.

First, as shown in Table 3.9 and Table 3.10, the average number of enriched terms in random networks was almost consistently lower than in the co-prediction networks, which suggests that nodes in the networks share more biological information than the nodes of random networks.

Then, to test if the terms overlap between co-prediction and random networks is different than between co-prediction and co-expression networks, a 3-way analysis and measure the $\Delta overlap$ was performed as:

$$\Delta overlap = overlap(\text{co-prediction, co-expression}) - overlap(\text{co-prediction, random})$$

Table 3.11 and Table 3.12 show the $\Delta overlap$ for each network and each co-expression based method. The average overlap percentage of 100 random networks was used and the resulting $\Delta overlap$ was averaged across all eight datasets. Positive differences rep-

[1]Unfortunately, this analysis, performed later in time, could not be replicated exactly as in other parts of the chapter due to changes to the PANTHER web service. The category of experimentally confirmed GO terms previously used has been removed. Although PANTHER has promised to restore this category in the future, the only way forward was to use the set of all GO terms instead. Therefore, for the sake of comparison, the enrichment analysis was repeated for all co-prediction and co-expression networks.

Dataset	C_1	C_2	C_3	C_4
Prostate	22.1±5.2	2.4±2.1	15.7±4.1	7.5±2.4
Leukemia	3.2±4.8	0.4±1.7	1.7±5.4	-1.2±2.7
Lung-Michigan	-0.5±1.8	-0.2±0.6	-2.0±4.2	-0.3±0.7
Dlbcl	3.9±5.3	-1.3±1.8	3.4±5.3	-0.9±1.4
AML	11.5±5.4	13.4±2.7	8.2±4.1	13.8±1.9
Colon-Breast	0.5±0.8	0.8±0.5	4.5±1.7	0.7±0.6
Lung-Harvard	12.8±2.6	1.4±1.0	13.5±3.3	1.3±1.5
CNS	-8.2±5.0	-1.8±2.9	-6.5±3.6	-2.5±3.2

Table 3.10: Average difference in enriched pathways between FuNeL and random networks.

Network	Pearons	ARACNE	MIC
SE(C_1)	-0.5	13.8	9.1
SE(C_2)	-18.6	5.6	-5.4
SE(C_3)	5.5	13.5	10.9
SE(C_4)	-8.8	14.5	3.3
SN(C_1)	1.3	1.9	3.8
SN(C_2)	1.5	2.9	5.0
SN(C_3)	2.4	0.1	3.5
SN(C_4)	2.6	3.5	4.0

Table 3.11: Average $\Delta overlap$ in enriched GO terms overlap between FuNeL and random networks.

Network	Pearson	ARACNE	MIC
SE(C_1)	3.0	8.0	3.0
SE(C_2)	-1.3	9.0	-7.0
SE(C_3)	8.0	11.0	1.0
SE(C_4)	-9.0	22.0	4.0
SN(C_1)	7.0	1.0	4.0
SN(C_2)	4.0	3.0	8.0
SN(C_3)	-2.0	-5.0	2.0
SN(C_4)	1.0	8.0	8.0

Table 3.12: Average $\Delta overlap$ in enriched pathways overlap between FuNeL and random networks.

resent a larger overlap between co-prediction and co-expression networks than between co-prediction and random networks.

The overlap with random networks was rarely greater than the overlap with co-expression networks (mostly for the Pearson method). Negative values for both GO

terms and pathways were observed for the $SE(C_2)$ and $SE(C_4)$ networks. As these are the largest networks (in terms of number of nodes), they contain up to 1/3 of the size of the original dataset. Therefore, the chance that a random network would include the same or extremely similar nodes as in the corresponding co-prediction network is high, leading to a larger overlap of enriched terms. For all the remaining networks positive values were consistently observed. Given the obtained results, it is possible to conclude that the knowledge associated with FuNeL networks is different and more biologically relevant than what can be achieved by chance.

3.3.4 Quantifying the amount of captured biological knowledge

When comparing the co-prediction and co-expression paradigm, the amount of knowledge captured by different networks was contrasted. The amount of biological knowledge (number of enriched terms) captured by a network is related to its size (number of nodes). Therefore, to fairly compare networks of different sizes, the normalised Enrichment Score (ES) was used:

$$ES = \frac{number\ of\ enriched\ terms}{number\ of\ nodes}$$

The score assesses how much a network contains biologically related nodes. The higher it is, the larger is the biological similarity between the nodes of a network.

To have a global view of the performance of each inference method in term of ES, a two-step analysis was performed for each enrichment category. First, using the ES, the networks generated by each method across the datasets were ordered (lower rank indicates higher ES) to identify the best performing one. Then, the best performing networks of each method were compared by calculating their average ES rank across the eight datasets. A rank-based scheme was necessary given the different distributions of ES across the different datasets. This guarantees a more fair and interpretable analysis. The full network ranks for each inference method are available in Section A.1 of Appendix A. The results of this analysis, using the best networks for different inference approaches, are reported in Table 3.13. MIC performed best when ES was

calculated using the GO terms (it was ranked first in each of those categories). When ES was calculated using the biological pathways, C_4 and ARACNE $SE(C_1)$ shared the highest rank.

Category	FuNeL	Pearson	ARACNE	MIC
GO BP	C4 (3)	**SE(C3) (1.5)**	SN(C3) (4)	**SN(C3) (1.5)**
GO MF	C3 (3.5)	SN(C3) (3.5)	SN(C3) (2)	**SN(C3) (1)**
GO CC	C3 (4)	SN(C1) (3)	SN(C3) (2)	**SN(C3) (1)**
Patwhays	**C4 (1.5)**	SN(C2) (3.5)	**SE(C1) (1.5)**	SE(C3) (3.5)
Average	3.00 ± 1.10	2.88 ± 0.90	2.38 ± 1.10	**1.75 ± 1.20**

Table 3.13: Average ranks based on the Enrichment Score for the best performing networks of each inference method. For each category and for each method, is reported the network used in the analysis. The ranks (in brackets) were averaged across all 8 datasets, and the highest ranks are shown with bold font. The last row reports the average ranks across all the biological categories. The following abbreviations were used for GO categories: biological process (BP), molecular function (MF) and cellular component (CC).

Table 3.13 shows that the best performing networks for each method were mostly C_3 co-expression counterparts, in particular $SN(C_3)$. This is consistent with the result of the topological analysis where these networks were found to have the lowest number of nodes and suggests that smaller networks tend to be more enriched. The difference in performance between the FuNeL configurations is mainly due to the application of the second machine learning phase (the best networks were C_3 and C_4).

In addition, another analysis compared each similarity-based inference method against FuNeL. For each dataset and enrichment category, the networks were ranked from 1 to 12 (4 C_i + 4 $SE(C_i)$ + 4 $SN(C_i)$) by decreasing number of enriched terms (lower rank means higher ES). The ranks, averaged across all eight datasets, are reported in Table 3.14 for all the three co-expression inference approaches. In this pairwise analysis there is not a consistent winner across all the enrichment categories, in general, FuNeL networks performed similarly to Pearson and ARACNE. MIC seems to have better results than FuNeL only for GO categories, while FuNeL performed better when considering the biological pathways. Overall, especially among the top performing networks, the difference in ranks is minimal. Consequently, this ES based analysis suggests that co-expression and co-prediction networks tend to capture a sim-

ilar amount of (complementary, according to the results of Section 3.3.3) biological knowledge.

		FuNeL				Co-expression (SE)				Co-expression (SN)			
Method	**Cat.**	C_1	C_2	C_3	C_4	$SE(C_1)$	$SE(C_2)$	$SE(C_3)$	$SE(C_4)$	$SN(C_1)$	$SN(C_2)$	$SN(C_3)$	$SN(C_4)$
Pearson	**GO BP**	7.00 (7.5)	7.00 (7.5)	6.06 (6)	5.88 (3.5)	5.88 (3.5)	5.88 (3.5)	**5.12 (1)**	5.88 (3.5)	7.06 (9)	7.75 (12)	7.12 (10)	7.38 (11)
	GO MF	7.81 (11)	6.19 (5)	5.38 (2)	5.62 (3)	9.12 (12)	6.50 (7.5)	6.50 (7.5)	7.38 (10)	7.19 (9)	6.25 (6)	**4.31 (1)**	5.75 (4)
	GO CC	4.31 (5)	11.00 (12)	4.19 (4)	9.00 (10)	3.88 (2)	6.25 (6.5)	6.25 (6.5)	8.12 (8)	**3.19 (1)**	9.38 (11)	4.06 (3)	8.38 (9)
	Pathways	8.12 (10.5)	4.75 (2)	8.12 (10.5)	**4.38 (1)**	6.94 (8)	6.50 (6)	6.69 (7)	5.88 (5)	7.62 (9)	5.00 (3)	8.50 (12)	5.50 (4)
ARACNE	**GO BP**	6.69 (7)	6.94 (10)	6.25 (5.5)	**4.75 (1)**	6.25 (5.5)	8.12 (11)	6.88 (8.5)	9.25 (12)	5.19 (3)	6.88 (8.5)	5.06 (2)	5.75 (4)
	GO MF	7.44 (10)	6.50 (8)	5.69 (3)	6.19 (6.5)	6.00 (5)	8.75 (12)	5.62 (2)	8.62 (11)	5.81 (4)	7.00 (9)	**4.19 (1)**	6.19 (6.5)
	GO CC	4.31 (4)	10.75 (12)	3.44 (3)	8.25 (8)	5.50 (5)	9.38 (10)	6.75 (7)	10.12 (11)	2.44 (2)	8.88 (9)	**2.31 (1)**	5.88 (6)
	Pathways	7.88 (10.5)	5.38 (3)	7.88 (10.5)	5.25 (2)	**4.88 (1)**	6.25 (6)	5.50 (4)	7.00 (8)	7.56 (9)	6.50 (7)	8.38 (12)	5.56 (5)
MIC	**GO BP**	7.44 (8.5)	7.88 (11)	7.44 (8.5)	6.38 (5.5)	4.12 (2)	7.00 (7)	6.00 (4)	6.38 (5.5)	4.81 (3)	9.12 (12)	**3.94 (1)**	7.50 (10)
	GO MF	8.50 (12)	8.06 (10)	7.19 (8)	7.75 (9)	**3.62 (1)**	6.50 (5.5)	5.12 (3)	6.75 (7)	5.31 (4)	8.12 (11)	4.56 (2)	6.50 (5.5)
	GO CC	5.19 (6)	11.62 (12)	4.44 (4)	10.12 (10)	4.25 (3)	7.12 (7.5)	5.12 (5)	7.12 (7.5)	3.19 (2)	10.38 (11)	**1.69 (1)**	7.75 (9)
	Pathways	8.75 (12)	5.50 (3)	7.50 (10)	**4.75 (1)**	6.00 (5)	6.50 (6)	5.00 (2)	6.62 (7)	7.56 (11)	7.00 (9)	6.94 (8)	5.88 (4)

Table 3.14: Average network ranks based on the Enrichment Score. The ranks were averaged across all 8 datasets. The row-wise rank is given in brackets and the highest ranks are shown with bold font. The following abbreviations were used for GO categories: biological process (BP), molecular function (MF) and cellular component (CC).

3.3.5 Evaluation of the networks in a disease context

To verify if the topology of the inferred networks is biologically meaningful, it was analysed how it defines the relationships between genes that are known to be associated with the disease (cancer) in hand of each dataset. The disease-associated genes were expected to be more closely connected than other genes and to be present in functional units, such as triangle motifs. The proximity of the disease-associated genes (i.e. how closely connected they are compared with non-disease-associated genes) was measured and it was counted the number of triangular relationships present in each network (i.e. the percentage of triangles containing one, two or three disease-associated genes). As presented in Section 3.3.3, a two-step analysis was performed by using the gene-disease association (G-D) metrics for the ranking. The results are reported in Table 3.15, the full ranks for each inference method are available in Section A.2 of Appendix A.

The average ranks, for both sources of G-D associations, show that co-prediction outperforms the other inference paradigms. The proximity of the disease-associated genes was in general higher in C_2 networks. Therefore, the co-prediction paradigm identifies the core elements of the network more accurately. This result highlights the benefits of including functional information, whenever these are available, in the network infer-

Source	Cat.	FuNeL	Pearson	ARACNE	MIC
Curated	**1A**	**C2 (1)**	SN(C2) (4)	SN(C3) (2.5)	SN(C2) (2.5)
	2A	**C3 (1)**	SN(C3) (2)	SE(C2) (3)	SN(C2) (4)
	3A	C1 (2)	SN(C1) (3)	SE(C4) (4)	**SE(C2) (1)**
	Proximity	**C2 (1)**	SN(C3) (2.5)	SE(C4) (2.5)	SE(C2) (4)
	Average	**1.25 ± 0.50**	2.88 ± 0.90	3.00 ± 0.70	2.88 ± 1.40
Malacards	**1A**	**C2 (1)**	SN(C2) (4)	SN(C4) (3)	SE(C4) (2)
	2A	**C2 (1.5)**	SN(C4) (4)	**SE(C4) (1.5)**	SN(C2) (3)
	3A	C3 (2)	SN(C4) (3)	SE(C2) (4)	**SN(C2) (1)**
	Proximity	**C2 (1)**	SE(C4) (4)	SE(C4) (3)	SE(C2) (2)
	Average	**1.38 ± 0.5**	3.75 ± 0.5	2.88 ± 1.00	2.00 ± 0.80

Table 3.15: Average ranks based on the G-D associations for the best performing networks of each inference method. For each category and for each method are reported the network used for the analysis. The ranks (in brackets) were averaged across all 8 datasets, and the highest ranks are shown with bold font. The last row reports the average ranks across all the categories. The number of disease-associated genes participating in a triangle is denoted as 1A, 2A and 3A.

ence process (FuNeL is using the class labels assigned to the samples of the dataset), in contrast to the co-expression approach solely based on gene expression similarity.

There is also a clear difference in the number of disease-associated genes participating in the triangles; FuNeL networks were ranked higher than the co-expression networks. The only category in which MIC had a better rank was $3A$. However, considering the low number of triangles defined by three disease-associated genes, many ties affected the positions in this category. Overall, these results demonstrate the higher ability of the FuNeL networks in identifying relationships between disease driving factors and potentially provides a framework for the discovery of new G-D associations.

Similar to the analysis performed when using the ES (Table 3.14), the networks were also ranked from 1 to 12 ($4\ C_i + 4\ SE(C_i) + 4\ SN(C_i)$) using the G-D associations information. The rankings are reported in Table 3.16 (curated) and Table 3.17 (Malacards). These results further highlight the better performance, in a disease-context, of the FuNeL networks if contrasted with (different) co-expression networks.

		FuNeL				Co-expression (SE)				Co-expression (SN)			
Method	**Cat.**	C_1	C_2	C_3	C_4	$SE(C_1)$	$SE(C_2)$	$SE(C_3)$	$SE(C_4)$	$SN(C_1)$	$SN(C_2)$	$SN(C_3)$	$SN(C_4)$
Pearson	**1A**	4.31(3)	**3.00 (1)**	4.81 (4)	3.25 (2)	8.50 (11)	8.12 (8)	8.38 (9)	6.50 (5)	8.50 (11)	6.50 (5)	7.62 (7)	8.50 (11)
	2A	5.19 (3)	6.94(7)	**3.88 (1)**	5.62 (4)	6.50 (5)	8.00 (11)	7.62 (10)	7.00 (8)	6.62 (6)	7.38 (9)	5.12 (2)	8.12 (12)
	3A	6.38 (4)	6.62 (8)	6.50 (6)	7.12 (12)	6.56 (7)	6.81 (10)	6.44 (5)	6.94 (11)	**5.94 (1)**	6.69 (9)	6.06 (3)	**5.94 (1)**
	Proximity	5.75 (5)	**4.63 (1)**	5.00 (3)	**4.63 (1)**	6.75 (7)	8.75 (12)	7.13 (9)	8.50 (11)	6.44 (6)	6.75 (7)	5.69 (4)	8.00 (10)
ARACNE	**1A**	5.56 (3)	**4.38 (1)**	6.31 (6)	4.75 (2)	6.81 (7)	7.44 (9)	8.00 (12)	7.62 (11)	6.19 (5)	7.50 (10)	6.12 (4)	7.31 (8)
	2A	4.94 (3)	5.19 (4)	**3.94 (1)**	4.44 (2)	7.62 (9)	6.25 (5)	6.88 (7)	6.31 (6)	8.69 (11)	7.69 (10)	8.69 (11)	7.38 (8)
	3A	**5.38 (1)**	6.00 (4)	5.50 (2)	5.62 (3)	6.94 (8)	6.94 (8)	6.94 (8)	6.94 (8)	6.94 (8)	6.94 (8)	6.94 (8)	6.94 (8)
	Proximity	5.50 (4)	**4.13 (1)**	4.88 (3)	4.50 (2)	8.13 (10)	7.38 (8)	6.88 (7)	5.75 (5)	8.44 (11)	7.38 (8)	8.44 (11)	6.63 (6)
MIC	**1A**	4.75 (3)	**3.12 (1)**	5.62 (4)	3.88 (2)	8.12 (10)	6.25 (6)	7.62 (8)	6.50 (7)	9.12 (11)	6.12 (5)	9.12 (11)	7.75 (9)
	2A	4.75 (3)	5.94 (4)	**3.94 (1)**	4.25 (2)	8.06 (9)	6.69 (7)	8.44 (12)	6.56 (6)	8.06 (9)	6.31 (5)	8.31 (11)	6.69 (7)
	3A	6.06 (3)	6.31 (5)	6.19 (4)	6.88 (10)	6.69 (8)	**5.56 (1)**	7.81 (12)	5.81 (2)	6.81 (9)	6.44 (7)	7.06 (11)	6.38 (6)
	Proximity	5.81 (4)	**4.13 (1)**	4.94 (3)	4.38 (2)	7.19 (9)	6.38 (5)	7.06 (7)	7.06 (7)	8.06 (11)	6.88 (6)	7.94 (10)	8.19 (12)

Table 3.16: Average network ranks based on the G-D associations (curated). The ranks were averaged across all 8 datasets. The row-wise rank is given in brackets and the highest ranks are shown with bold font. The number of disease-associated genes participating in a triangle is denoted as 1A, 2A and 3A.

		FuNeL				Co-expression (SE)				Co-expression (SN)			
Method	**Cat.**	C_1	C_2	C_3	C_4	$SE(C_1)$	$SE(C_2)$	$SE(C_3)$	$SE(C_4)$	$SN(C_1)$	$SN(C_2)$	$SN(C_3)$	$SN(C_4)$
Pearson	**1A**	5.00 (3)	**2.71 (1)**	5.93 (4)	3.00 (2)	8.93 (12)	7.86 (9)	8.14 (10)	7.36 (7)	7.64 (8)	6.43 (5)	6.71 (6)	8.29 (11)
	2A	5.36 (4)	**3.86 (1)**	5.21 (3)	4.00 (2)	8.21 (11)	7.14 (8)	7.50 (9)	6.79 (7)	8.57 (12)	6.64 (5.5)	8.07 (10)	6.64 (5.5)
	3A	6.64 (7)	6.64 (7)	**4.93 (1)**	6.64 (7)	6.64 (7)	6.64 (7)	6.64 (7)	6.64 (7)	6.64 (7)	6.64 (7)	6.64 (7)	6.64 (7)
	Proximity	5.71 (3)	**3.86 (1)**	6.68 (7)	4.43 (2)	6.07 (4)	7.14 (8.5)	6.71 (6)	6.29 (5)	7.50 (10)	7.57 (11)	8.71 (12)	7.14 (8.5)
ARACNE	**1A**	5.00 (3)	**3.50 (1)**	5.50 (4)	4.38 (2)	6.38 (9)	6.69 (12)	6.44 (10)	6.06 (7)	6.25 (8)	6.00 (6)	6.44 (10)	5.62 (5)
	2A	5.44 (5)	**4.06 (1)**	5.31 (4)	4.81 (3)	6.38 (10)	6.56 (11)	6.75 (12)	4.62 (2)	6.12 (8)	5.94 (6)	6.00 (7)	6.25 (9)
	3A	5.94 (8)	5.94 (8)	**4.56 (1)**	5.94 (8)	5.94 (8)	5.06 (2)	5.94 (8)	5.19 (3)	5.94 (8)	5.94 (8)	5.94 (8)	5.94 (8)
	Proximity	6.21 (7)	**4.14 (1)**	7.50 (9)	5.29 (2)	8.0 (10)	6.29 (8)	6.14 (5)	5.43 (3)	8.00 (10)	6.14 (5)	8.71 (12)	6.14 (5)
MIC	**1A**	4.75 (3)	**2.88 (1)**	5.62 (6)	3.25 (2)	7.25 (9)	6.12 (7)	7.38 (11)	5.12 (4)	7.25 (9)	5.12 (4)	7.38 (11)	6.12 (7)
	2A	5.19 (6)	**3.56 (1)**	5.06 (5)	4.19 (2)	8.25 (12)	5.69 (8)	6.62 (9)	4.94 (4)	7.81 (11)	4.69 (3)	6.69 (10)	5.56 (7)
	3A	6.62 (11)	6.62 (11)	4.94 (2)	6.62 (11)	6.19 (8)	5.00 (3)	6.31 (9)	5.81 (7)	5.56 (6)	**3.81 (1)**	5.44 (5)	5.31 (4)
	Proximity	5.93 (7)	**3.86 (1)**	7.36 (9)	4.71 (2)	8.14 (10)	5.14 (4)	5.14 (4)	6.28 (8)	10.07 (11)	5.14 (4)	10.79 (12)	5.43 (6)

Table 3.17: Average network ranks based on the G-D associations (Malacards). The ranks were averaged across all 8 datasets. The row-wise rank is given in brackets and the highest ranks are shown with bold font. The number of disease-associated genes participating in a triangle is denoted as 1A, 2A and 3A.

3.3.6 Prostate cancer case study: enriched terms

To compare in detail the difference in the biological knowledge captured by the co-prediction and co-expression networks, the global analysis presented earlier was followed by a case study focused on a dataset characterised by a single disease – prostate cancer [170]. Particular focus was put on the specific knowledge captured by one paradigm but not the other.

In Figures 3.11 and 3.12 are compared the co-prediction and the Pearson co-expression networks inferred from the prostate cancer dataset. The attention was set on GO terms and pathways enriched uniquely in one type of network. For the sake of readability, the generic GO terms (with depth < 9 in the GO hierarchical structure) were filtered out. C_2 was the network with the largest number of unique terms, followed by C_4 and $SN(C_2)$. A total of 16 GO terms and 21 pathways were unique to co-prediction

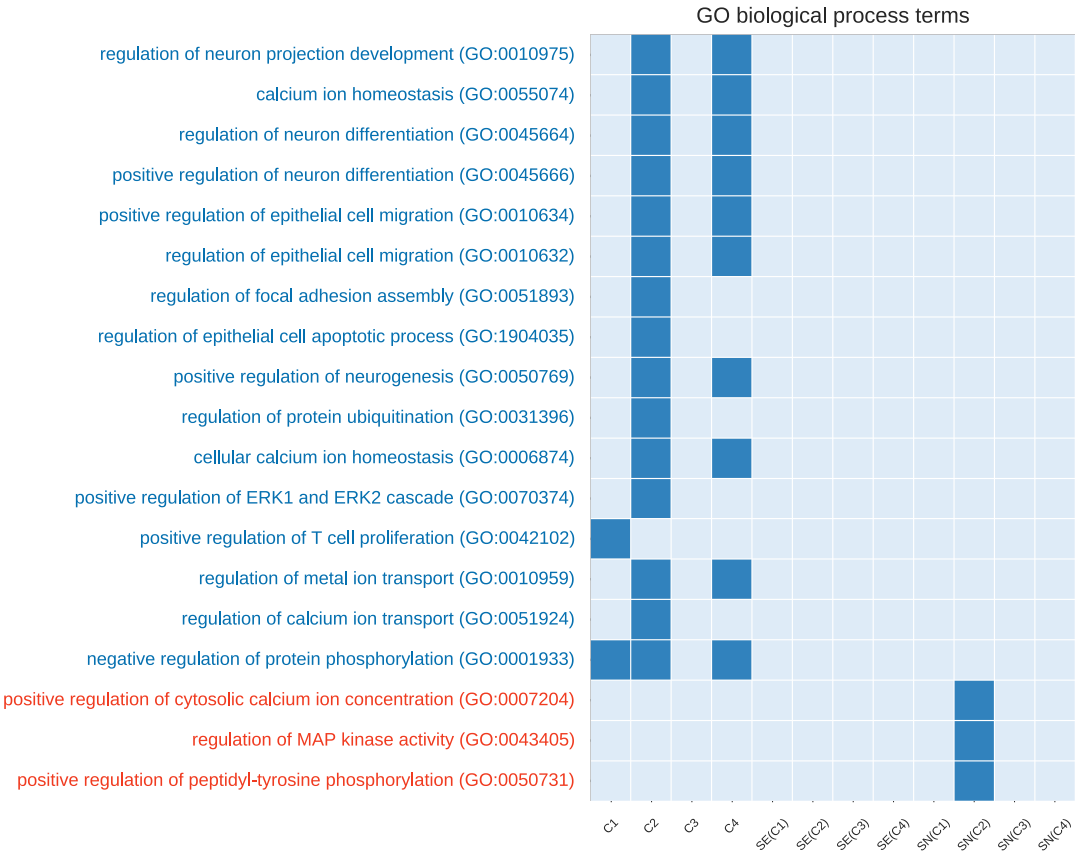


Fig 3.11: Unique enriched GO terms (biological process) for each network configuration. The x-axis shows the 12 investigated networks. The y-axis shows the names of enriched terms unique to co-prediction or Pearson co-expression networks. Red terms are associated with co-expression networks, blue with co-prediction. Empty columns indicate networks with no unique terms.

networks while only 3 GO terms and 4 pathways were specific to co-expression networks. A similar disproportion in favour of the co-prediction networks was found when comparing with MIC and ARACNE networks (see Section A.3.1 of Appendix A for the complete analysis).

Several of the unique GO terms enriched in the co-prediction networks are related to prostate cancer, according to the specialised literature. The role of the *Protein ubiquination* in prostate cancer was recently analysed and showed an impact for its treatments [178]. The *ERK* pathway is involved in the motility of prostate cancer cells [179]. Prostate cancer cells seem to alter the nature of their *calcium* influx to promote growth and acquire *apoptotic* resistance [180]. Furthermore, the role of *calcium home-*

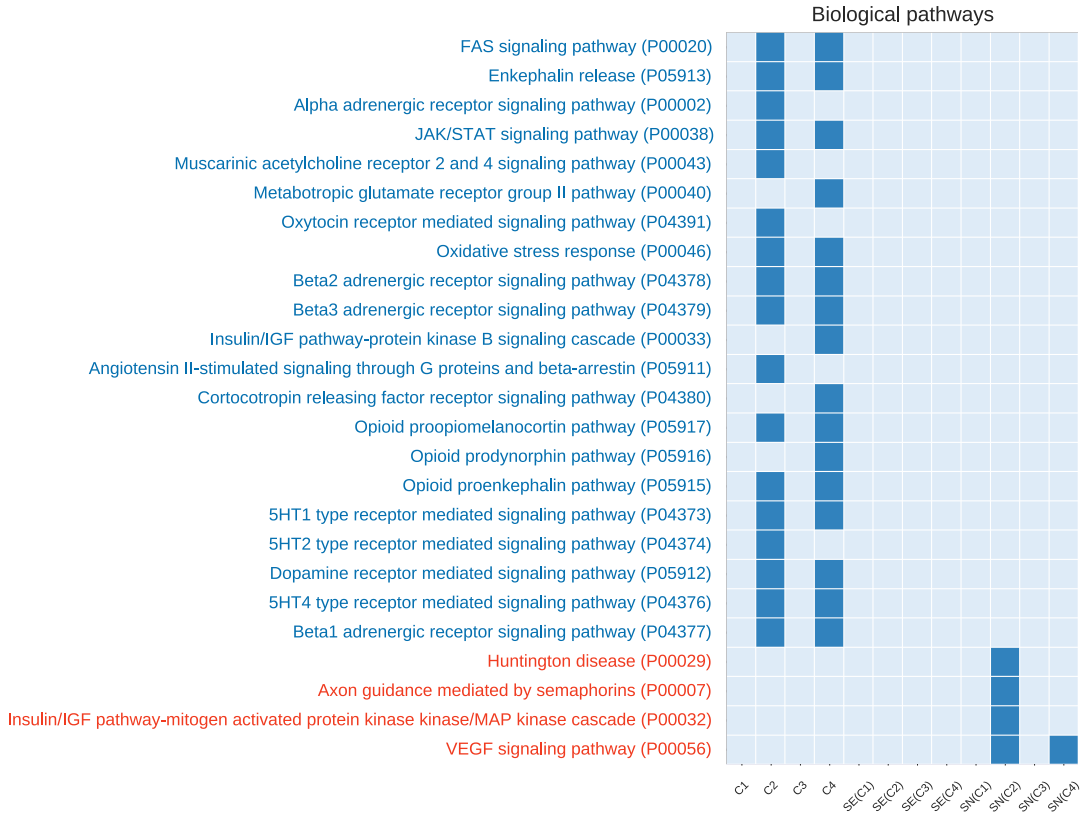


Fig 3.12: Unique enriched biological pathways for each network configuration. The x-axis shows the 12 investigated networks. The y-axis shows the names of enriched terms unique to co-prediction or Pearson co-expression networks. Red terms are associated with co-expression networks, blue with co-prediction. Empty columns indicate networks with no unique terms.

ostasis in the majority of the cell-signalling pathways involved in carcinogenesis has been well established, prostate cancer included [181].

Some enriched pathways specific to co-prediction networks are also highly relevant to prostate cancer. Several studies showed the involvement of the *JAK/STAT pathway* in the prostate cancer development [182, 183]. There is multiple evidence suggesting that one of the major ageing-associated influences on prostate carcinogenesis is *oxidative stress* and its cumulative impact on DNA damage [184, 185]. Finally, *FAS* (also called Apo1 or CD95) plays a central role in the physiological regulation of programmed cell death and has been implicated in the pathogenesis of various malignancies and diseases of the immune system including prostate cancer [186].

An additional analysis was performed on the biological terms related to the hubs (highly connected nodes) of the inferred networks. A node v was considered to be a

hub if its degree was at least one standard deviation above the average network degree, that is if:

$$d(v) > \mu_d + \sigma_d$$

where $d(v)$ is the degree of the node v (number of direct neighbours), and μ_d and σ_d are the mean and standard deviation of the network node degree distribution. To compare the networks, the top ten most frequent Gene Ontology terms, shared between each network's hubs, were used. To make this analysis more specific, the most generic/common terms (which could be associated with many genes) were discarded, only the GO terms situated at level 10 or higher in the GO hierarchy, were considered. Figure 3.13 provides the top ten most frequent GO terms associated to the hubs of co-prediction and Pearson co-expression networks. Blue terms were found only in co-prediction networks, red terms were found only in co-expression networks and green terms were in common.

Terms	**Pearson**	**ARACNE**	**MIC**
Co-prediction	16	18	16
Co-expression	19	20	19
Common	11	9	11

Table 3.18: Unique and common terms from networks' hubs

In total, 16 unique terms for co-prediction networks were found, 19 unique terms for co-expression networks and 11 common terms. Table 3.18 summarises the number of unique and common terms shared between networks created with different approaches. The plots associated to the comparison of FuNeL with ARACNE and MIC are available in the Figure 3.14 and Figure 3.15. The results further highlight biological terms exclusively associated either with co-prediction and co-expression networks.

An analysis of term overlap was conducted using only the best performing networks in the curated G-D association analysis (namely C_2 for FuNeL, $SN(C_3)$ for Pearson, $SE(C_4)$ for ARACNE and $SE(C_2)$ for MIC, see Section A.2 in Appendix A for details). The aim was to further show how the knowledge captured by networks inferred with different approaches is partially shared and often highly network-specific. Figure 3.16

Fig 3.13: Top 10 most frequent biological processes from Gene Ontology found in the network hubs when comparing FuNeL and Pearson co-expression networks. Blue terms were found only in co-prediction networks, red terms were found only in co-expression networks, and green terms were found in both.

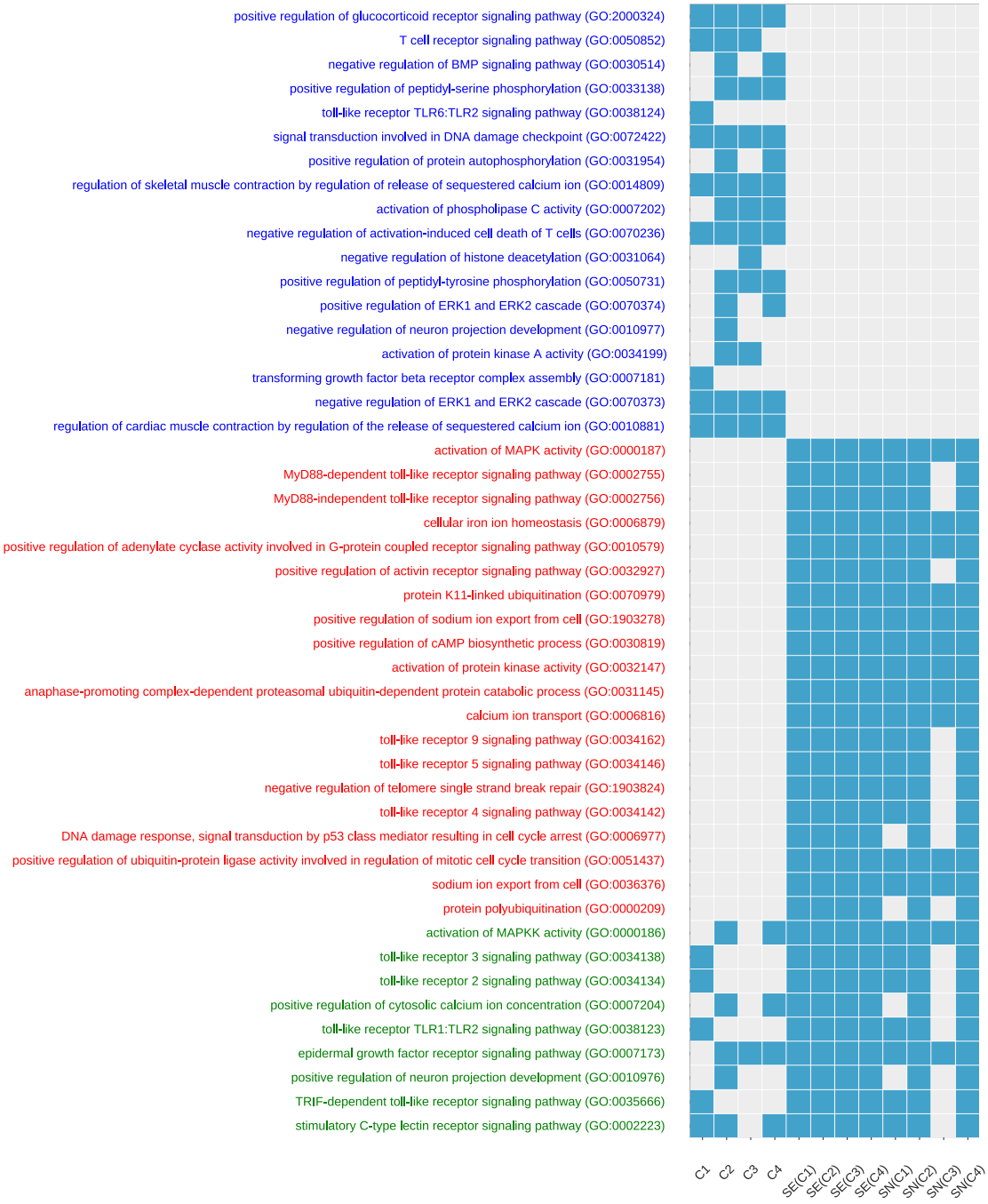


Fig 3.14: Top 10 most frequent biological processes from Gene Ontology found in the network hubs when comparing FuNeL and ARACNE co-expression networks. Blue terms were found only in co-prediction networks, red terms were found only in co-expression networks, and green terms were found in both.

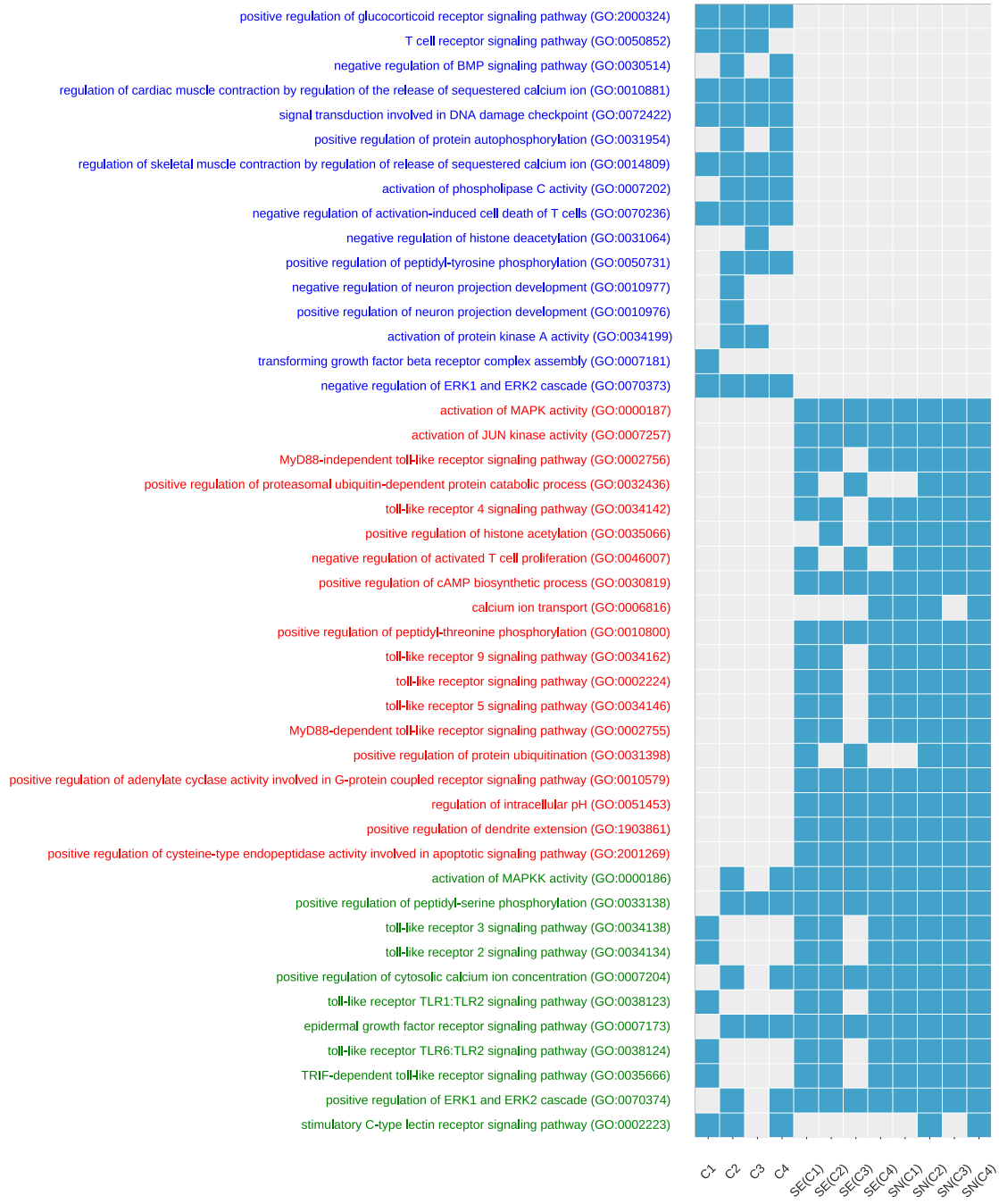


Fig 3.15: Top 10 most frequent biological processes from Gene Ontology found in the network hubs when comparing FuNeL and MIC co-expression networks. Blue terms were found only in co-prediction networks, red terms were found only in co-expression networks, and green terms were found in both.

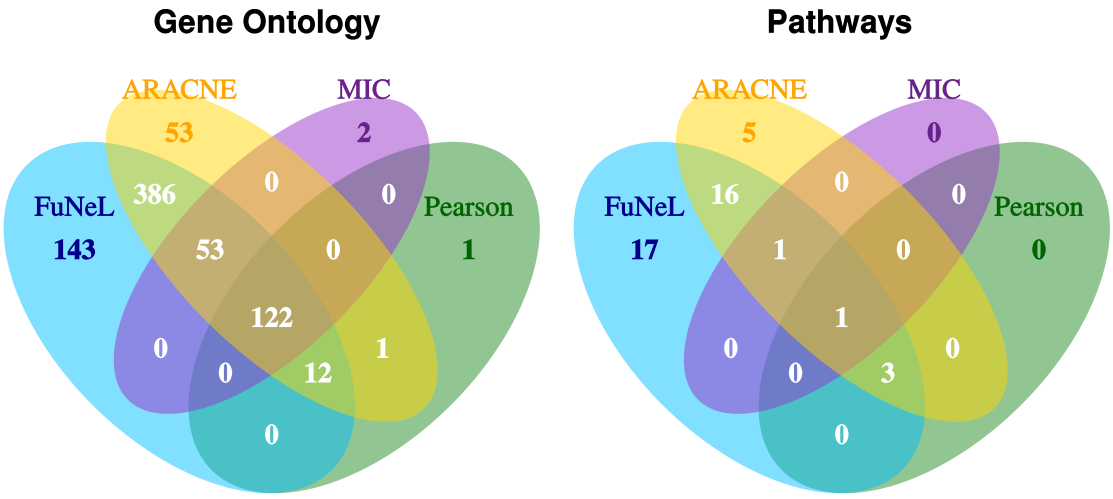


Fig 3.16: Overlap of PANTHER enriched terms between the best performing networks identified with the G-D association analysis (curated databases). The values represent the number of enriched terms that are unique or shared between different networks. On the left the overlap of GO terms (including all 3 categories: BP, CC and MF), on the right the overlap of pathways.

shows the number of shared/unique enriched GO terms (including all three GO categories) and pathways across different networks. In total 122 GO terms were found to be captured by all the networks, while only one pathway was in common to all. The FuNeL network is similar to the ARACNE one, a total of 386 GO terms and 16 pathways were associated to both. Few terms were purely specific of MIC and Pearson networks, on the contrary a large number of unique terms was related to C_2 (143 GO terms and 17 pathways). Overall, Figure 3.16 emphasises even more the complementarity between the co-prediction and co-expression approaches regarding the captured biological knowledge.

3.3.7 Prostate cancer case study: disease associations

The literature and the public cancer databases (not used in the inference process) were searched to verify if key nodes in the generated networks are associated with prostate cancer. The node degree (number of connections) and the betweenness centrality (number of shortest paths between all the pair of nodes that pass through a given node) were used as a measure of the node importance.

Literature analysis The top three most connected nodes (hubs) were picked for each of the four FuNeL networks (configurations). The set contains six genes: *GSTM2*, *NELL2*, *CFD*, *PTGDS*, *PAGE4* and *LMO3*. All the genes from this set, except *LMO3*, were also found to be the most central nodes (with highest betweenness centrality).

Almost all these genes are related to prostate cancer, according to the specialised literature:

- *NELL2* contributes to alterations in epithelial-stromal homeostasis in benign prostatic hyperplasia and codes for a novel prostatic growth factor [187], and is also an indicator of expression changes in cancer samples [188].
- *CFD* (adipsin gene) is over expressed in PP periprostatic adipose tissue of prostate cancer patients [189].
- *PTGDS* (and two other genes) are expressed at consistently lower levels in clinical prostate cancer tissues and form a signature that predicts biochemical relapse [190].
- *PAGE4* modulates androgen receptor signalling, promoting the progression to advanced lethal prostate cancer [191], and has a significantly lower expression level in patients with prostate recurrent disease [192].
- *LMO3* interacts with *p53*, a well known gene tumour suppressor in prostate cancer [193].

The only gene without literature support is *GSTM2*. It might represent a good target for further experimental verification.

Validation on independent data To further validate the biological relevance of the inferred networks, an independent prostate cancer dataset [194] was used from the collection of data available in the cBioPortal for Cancer Genomics [195]. The independent data was used to check the genomic alteration of the key topological genes. The top ten hubs (nodes with the highest degree) and the top ten central nodes (with highest betweenness centrality) were analysed in the co-prediction network that

better performed in the gene-disease association analysis using the curated databases: C_2 (see Section A.2 in Appendix A). The genes with highest degree were: *PTGDS*, *PAGE4*, *NELL2*, *GSTM2*, *PARM1*, *MAF*, *LMO3*, *COL4A6*, *RBP1* and *ABL1*. For the betweenness centrality, the set was almost identical; only *RBP1* was replaced by *MYH11*. On average the expression in samples was altered in 31.8% cases for hubs and in 35.6% cases for central nodes. The most altered genes were found to be downregulated at the mRNA level: *COL4A6* (65%), *MYH11* (58%), *PARM1* (53%) and *GSTM2* (52%). In addition, genomic alterations in several key genes were found to be strongly co-occurrent (e.g. *PTGDS* – *GSTM2*, *PAGE4* – *COL4A6*, *PAGE4* – *RBP1*, etc.).

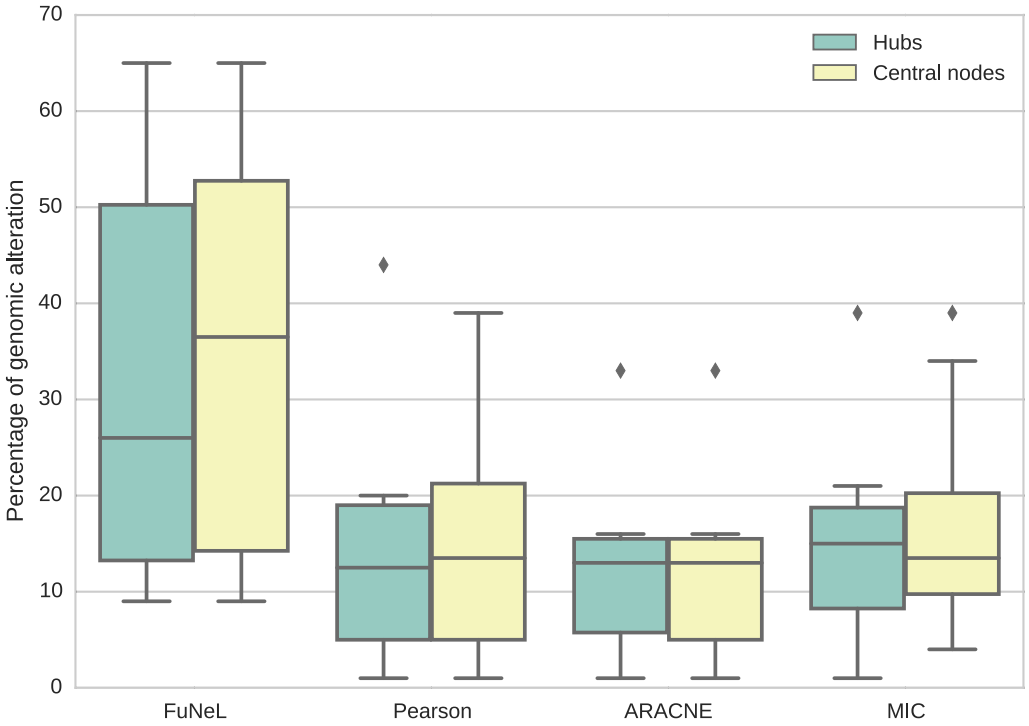


Fig 3.17: Distribution of the percentage of genomic alteration in the samples of an independent dataset for top 10 hubs and central nodes. The topologically important genes were selected from the best performing networks in the G-D association analysis on curated datasets.

When repeating the analysis with the best ranked co-expression networks, (namely $SN(C_3)$ for Pearson, $SE(C_4)$ for ARACNE and $SE(C_2)$ for MIC), on average the alteration level was consistently lower, at most half of the FuNeL key genes. The percentages of alterations are represented as boxplots in Figure 3.17, while the average

alterations are reported, for each method, in Table 3.19. As Figure 3.17 shows, FuNeL can identify many more genes with a higher percentage of alteration than other methods. Therefore, the topologically relevant nodes in the best co-prediction network represent genes more strongly related to the prostate cancer that can possibly be considered as biomarkers, with over two times more frequent genomic alterations.

Genes	**FuNeL**	**Pearson**	**ARACNE**	**MIC**
Hubs	31.8 %	14.2 %	12.3 %	15.2 %
Central nodes	35.6 %	14.7 %	12.2 %	17.1 %

Table 3.19: Average percentage of genomic alteration for top hubs and central nodes in the independent dataset.

The detailed list of genomic alterations for the top hubs and the central nodes for each analysed network is available in Section A.3.2 of Appendix A.

3.4 Discussion

This chapter introduced FuNeL: a protocol to infer functional networks based on the *co-prediction* paradigm where the structure of a rule-based machine learning model is used to identify functional relationships between genes. FuNeL generates functional networks using a different approach than the state-of-the-art methods, commonly based on a similarity paradigm. Machine learning is at the core of the FuNeL protocol, the networks are generated via the mining of machine learning models (rule-based models in this instance) inferred to solve a classification task. Different options in the FuNeL protocol provide a total of 4 different configurations, each one generating diverse networks. These have been contrasted with networks generated using the co-expression paradigm, the most widely adopted similarity-based approach.

Before the comparison with co-expression methods, synthetic data were used to evaluate the ability of FuNeL to retrieve known biological associations. It was fundamental to assess if the information obtained from the mining of machine learning was indeed **relevant**. That is, if the relationships inferred with FuNeL were found to be meaningful. When applied to synthetic datasets, FuNeL was able to identify existing pairwise relationships between genetic attributes (SNPs). The obtained results were in line,

and in some cases superior, to a recently proposed approach based on permutated random forest [70]. These findings mean that the assumption on which the co-prediction approach, and in general FuNeL, are based on, was proven to be correct. Attributes (SNPs) that (statistically) appear together more frequently than by chance in the BioHEL's classification rules are also associated within the tested synthetic data.

Encouraged by those findings, it was checked if a rule-based machine learning model, with its complex knowledge representation, might be used to identify biologically meaningful relationships that escape the standard inference methods. This analysis was performed using eight real-world cancer-related transcriptomics datasets. FuNeL was compared with three co-expression inference methods (ARACNE, Pearson and MIC) by using networks of matching size and generated from the same data. The differences between co-prediction and co-expression were observed from three points of view: basic topological properties, enriched biological terms and relationships between known disease-associated genes.

The comparison of the topological properties revealed the influence of the protocol options. Not surprisingly, both the feature selection and the second training phase reduce the size of the networks, but at the same time, increase the clustering coefficient and the number of connections. The clustering coefficient was found to be lower in almost all the ARACNE networks, probably due to the pruning procedure. It was also lower in many MIC networks. Moreover, when feature selection was applied, the resulting networks had higher clustering coefficient than Pearson co-expression networks with the same number of edges. Interestingly, all of the co-expression networks were less compact (lower diameter). This is probably because many attributes appear together in the same classification rules, reducing the distance from each other in the FuNeL network.

The differences in networks topology translated into differences in the contained biological information. The overlap between enriched GO terms and pathways across protocol configurations was generally low, indicating that different configurations infer networks that capture different biological knowledge. The term overlaps between the co-prediction networks and their equivalent co-expression counterparts were even lower. This can be interpreted as evidence that the biological knowledge captured by the two

paradigms is not entirely redundant, but in large part complementary. Associations defined by FuNeL are not limited to attributes that show similar expression patterns but are extended to pairs of attributes that participate in the same classification rule.

Differences between the networks were also observed in the analysis of the connections between genes known to be related to a specific disease. The disease-associated genes were more closely connected (higher proximity) in the co-prediction networks, which means that the disease-related nodes of the network were closer to its core. In addition, the number of functional units (triangle motifs), that can identify new gene-disease associations, was found to be higher in the co-prediction networks. Therefore, it can be concluded that the co-prediction approach better captures the abstract concept of functional relationship. The superior performance of FuNeL networks in identifying the disease-associated genes is likely a result of effective use of the class labels of the samples, which the similarity-based methods ignore. Although it would be tempting to attribute this performance difference entirely to the use of supervised learning in FuNeL, it would be an overstatement, as the knowledge of explicit links between genes and diseases is not available to it in training. The hypothesis is that this is rather a result of differences in expression values of the disease-associated genes, which taken together are able to discriminate between sample phenotypes.

To further analyse and compare the two paradigms, a case study on the prostate cancer dataset [170] was performed. FuNeL generated networks that were enriched with knowledge totally missed by all the co-expression networks. Topologically important genes (nodes) in the co-prediction networks were found: (1) to be altered in a high percentage of tumour samples in an independent cancer transcriptomic study, and (2) to be already associated with prostate cancer according to the specialised literature. Therefore, the co-prediction networks not only capture biological knowledge complementary to the co-expression networks but also better highlight the important genes involved in the disease process. The key nodes (hubs and central nodes in this instance) from FuNeL networks could be considered as candidate biomarkers for the disease of the data. This is directly linked to their relevance in the BioHEL's rules. Attributes frequently appear in the classification rules and become hubs if their expression can be

used to discriminate the samples of the data. Thus, they are likely to have an major role in the condition/disease.

3.5 Future work

One of the main limitations of FuNeL is the expensive computational effort required by the network generation phase that is based on the classification rules created with BioHEL. A time complexity analysis of the protocol is available in the Section A.4 of Appendix A. BioHEL belongs to the class of evolutionary machine learning algorithms, notoriously famous for being relatively slow. Nevertheless, each run of BioHEL (10 000 in the presented analysis) is totally independent, thus the generation of the rule sets can be trivially parallelised without any extra overhead. Another possible solution to this problem might be the substitution of BioHEL with other "faster" machine learning algorithms, such as random forest or decision trees. The analysis of FuNeL with other supervised approaches would be interesting from two points of view: (1) to check how those networks would be different from the BioHEL-based networks and (2) to assess if those networks would also capture complementary knowledge to that of similarity-based approaches.

The machine learning step proposed in FuNeL involves the employment of rule-based models generated with BioHEL. However, this does not represent the only possible solution. Other machine learning algorithms could be adopted within the learning phase to replace BioHEL. Among many examples there are unsupervised methods, such as the Apriori algorithm for association rule learning, or other supervised methods, such as decision tree algorithms (e.g. C4.5 or random forest). Some adjustment would be necessary to extract the knowledge from a different model representation, but the rest of the protocol could remain unchanged. For example, in the case of the decision trees, relationships could be inferred between attributes that share the same path from the root to the leaves of the tree. This potential flexibility in the choice of a learning algorithm, together with the ability to apply the protocol to different types of data, makes FuNeL a powerful tool for network inference.

The complementarity between co-prediction and co-expression networks has been extensively shown in this chapter. Given the characteristics of the two approaches, it seems like a wasted opportunity not try to integrate them. A network resulting from the union of both approaches could capture a wider amount of knowledge if compared to its single components. In addition, another possibility would be to exploit the similarity information within the learning process of the classification rules (BioHEL in this instance). More specifically, check whether co-expressed attributes can define meaningful (i.e. highly predictive) classification rules that can be used for the inference of a co-prediction network

The presented version of FuNeL generates functional networks from single biomedical datasets. However, it could be extended to handle multiple sources. For example, it could integrate multiple rule sets produced from a similar biomedical condition (e.g. prostate cancer or leukaemia) and infer a single functional network. Considering that some biomedical studies focus on a subset of the population (e.g. patients of a particular age or from a specific geographic location), the integrated network could be more representative of the condition than the one inferred from a single dataset. This integration strategy of combining multiple data as already been proven successful in providing more relevant and robust solution [196–198].

Summary

This chapter presented FuNeL, a protocol for the inference of functional networks. The analysis of the networks (prostate cancer in particular) has shown that an effect of the inference process is the identification of candidate biomarkers. Topologically important nodes (hubs and central nodes) were found to be highly altered in independent data, suggesting a possible role in the studied condition/disease. In this part of the dissertation, biomarkers played a minor role and emerged from the analysis of the inferred networks. The following chapter will directly assess the problem of biomarkers discovery presenting the RGIFE heuristic. While FuNeL analyses machine learning models to generates functional networks, in RGIFE the models are exploited to identify the best performing subsets of biomarkers.

4

RGIFE: a ranked guided iterative feature elimination heuristic for biomarkers identification

Contents

Abstract

This chapter studies RGIFE, a heuristic for the identification of small sets of highly predictive biomarkers. RGIFE uses the information extracted from the structure of machine learning models to discover irrelevant features that can be discarded without compromising the predictive performance. The RGIFE heuristic represents an important contribution to the field of biomedicine, as specifically designed to tackle the problem of biomarkers discovery and provide small relevant solutions.

4.1 Introduction

In recent years the rapid progress in bio-technologies, together with their cost decrease, has led to an explosive rise in the availability of good quality biomedical data. For example, in the last decade, the growth of transcriptomics (ArrayExpress), proteomics (PRIDE) and metabolomics (Metabolights) data stored at the European Bioinformatics Institute (EMBL-EBI) has been exponential, as illustrated in Figure 4.1. In 2015, the data stored at EMBL-EBI were more than 70 petabytes (1 petabyte = 1000 terabytes) [199]. Because of costs that are no more prohibitive, nowadays many laboratories, produce large-scale data from biological samples as a routine task. High-throughput experiments allow the analysis of the relationships and the properties of many biological entities (e.g. gene, proteins, etc.) at once. As a result, the observations are defined in a high dimensional space. However, while the cost of bio-technologies is reasonably low, the costs required for clinical studies are still high. Thus, biomedical data often contain only a small set of data points. Those two characteristics, when combined, lead to a problem known as the *curse of dimensionality*. This phenomenon was introduced by Bellman [200] to describe a problem generated by the exponential increase in volume associated with extra dimensions. In a machine learning context, it implies that a small addition in the data dimensionality requires a substantial increase in the number of samples to maintain the same quality of classification, regression, clustering, etc. [201]. Given the difficulties in obtaining large sets of samples (e.g. patients), coupled with the high dimensionality, the analysis of biomedical data with

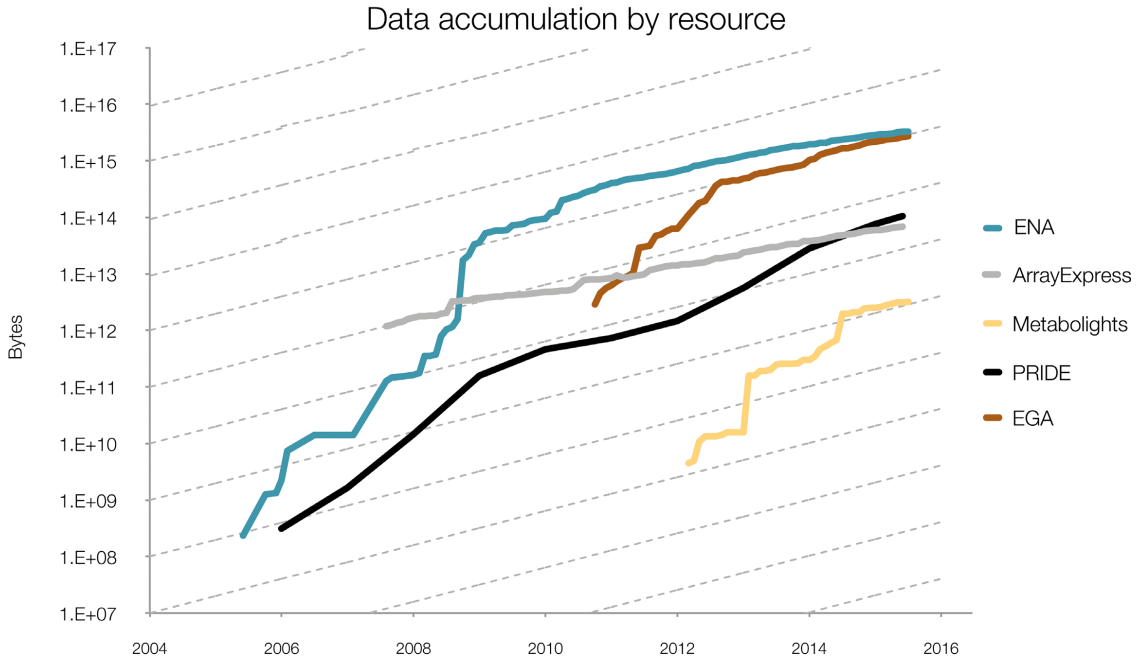


Fig 4.1: Data accumulation at EMBL-EBI [199].

machine learning techniques is a challenging task. In addition, biomedical data are often noisy. The noise can arise from the equipment used to perform the experiments, from human errors or from the stochasticity that might affect the biological phenomena being analysed. As a result, there is a need for relevant analytic techniques that deal with these noisy and high-dimensional data.

One of the major research fields in bioinformatics and biomedicine involves the discovery of driving factors from disease-related datasets. They are also known as *biomarkers* and have been described 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"* [95]. The challenging part, when looking for biomarkers, is the ability to identify and discard the irrelevant information (e.g. genes whose role is not important for the progress of a disease) so that important factors can emerge. Statistical methods have been traditionally used in biomedical studies to select the variables responsible for the presence of a disease, using both univariate [96] and multivariate approaches [99] (see Section 2.6 for more details).

Along with the traditional statistical methods, machine learning has been extensively and successfully employed, in many different forms, to solve the problem of biomarkers

discovery over the years [10]. Machine learning has been used in the form of ensemble methods [202], particle swarm optimisation [113], deep neural networks [203], etc. Feature selection, in machine learning, is defined as the process of selecting relevant variables to be used in the model construction. By removing useless features, the machine learning algorithm can generate better models at a lower computational cost. It is easy to think about the discovery of biomarkers, from biomedical data, as a feature selection problem. From this perspective, the goal is to identify a subset of features (biomarkers) that can build models that correctly predict the class of the samples. Although this is not their primary focus, when applied in biomedicine, feature selection methods can provide new scientific insights by recognising the factors that determine the presence of a condition in the data.

Over the years different feature selection methods have been designed. Some have been created explicitly to tackle biological problems, others were instead more generic and could be applied to a broad variety of problems. A popular approach for feature selection is to rank the attributes based on some importance criteria and then select only the top ones [103]. However, one of the main drawbacks is that the number of features to be selected needs to be set up-front and decide its exact value is a not-trivial problem. Other methods such as CFS [104] or mRMR (minimum Redundancy Maximum Relevance) [204] are designed to evaluate the goodness of a given subset of variables in relation to the class/output variable. When coupled with a feature subset search mechanism (e.g. BestFirst), they can automatically identify the optimal number of attributes to be selected.

A large class of feature selection methods is based on an iterative reduction process. The basic concept of these methods is to iteratively remove the useless feature(s) until a stopping condition is reached. The most known and used iterative reduction algorithm is the SVM Recursive Feature Elimination (SVM-RFE) [111]. It iteratively repeats three steps: (1) trains an SVM classifier, (2) ranks the attributes based on the weights of the classifier and (3) removes the bottom ranked attribute(s). SVM-RFE was initially designed to work with transcriptomics data, but nowadays it is commonly used in many different contexts. Several approaches have been presented after this method [205–207].

This chapter focuses on **RGIFE** (**R**anked **G**uided **I**terative **F**eature **E**limination): a heuristic for the identification of reduced biomarker signatures. RGIFE employs an iterative process for the biomarkers identification similar to SVM-RFE as the features are ranked based on their importance (contribution) within the machine learning model. However, two main characteristics differentiate the methods: (a) in RGIFE the removed features not always are the ones at the bottom of the importance rank and (b) in RGIFE the use of the *soft-fail* allows, under certain circumstances, to consider an iteration successful when it suffered a drop in performance, as long as it is within a tolerance level. The work presented here is a substantial extension of the work by Swan et al. [208] where the heuristic was first introduced. Every aspect of the original work has been thoroughly revisited by: (1) using a different machine learning algorithm to rank the features and evaluate the feature subsets, (2) introducing strategies to reduce the probability to stuck at a local optimum, (3) limiting the stochastic nature of the heuristic, (4) comparing the method with some well-known approaches commonly used in bioinformatics, (5) evaluating the performance using synthetic datasets and (6) validating the biological relevance of the signatures using a prostate cancer dataset as a case study.

In the next sections, first, the new version of RGIFE is compared with the original method proposed in [208], then is contrasted with five other algorithms both from a computational (using synthetic and real-world datasets) and a biomedical point of view. Afterwards, using a prostate cancer dataset as a case study, the knowledge associated with the signature (term used to identify a set of biomarkers) generated by RGIFE is thoroughly evaluated. The analysis performed showed that the new version of the heuristic outperforms its original version both in terms of prediction accuracy and number of selected attribute while being less computationally expensive. When compared with other feature reduction approaches, RGIFE obtained similar prediction performance while constantly selecting fewer features. The analysis completed on synthetic data demonstrated the ability of RGIFE to identify relevant attributes while discarding irrelevant and redundant information. Finally, the case study evaluation showed a higher biomedical relevance of the genes selected by RGIFE when compared with other methods.

4.2 Material and Methods

This section describes in detail the RGIFE heuristic and all the changes implemented in comparison with the original version. Then, the five different benchmarking methods are presented followed by a description of the datasets employed for the comparison. Finally, the experimental design and the approaches used for analysis of the predictive performance and the biomedical validation of the signatures are introduced.

4.2.1 The RGIFE heuristic

A detailed pseudo-code that describes the RGIFE heuristic is depicted in Algorithm 1, while in Figure 4.2 is illustrated its generic iterative nature.

RGIFE is able to analyse and extract biomarker signatures from any type of biomedical dataset. The only requirement is that the samples need to be associated to a finite set of categories or classes (e.g. control vs. case), that is they can be used for a classification problem. As briefly mentioned in the introduction, RGIFE removes attributes if their role in the predictive model is irrelevant. Therefore, the first step of the heuristic is to estimate the performance of the classifier using the original set of attributes and assess their importance (line 29). Any classifier that ranks the attributes, based on their relevance in the classification task, can be used in the heuristic. The original version of the heuristic employed BioHEL [39] as base classifier to generate the predictive models and the attribute rankings. In this new version of RGIFE, BioHEL has been replaced with a random forest classifier [50]. This choice was primarily due to reduce the overall computational cost, as will be shown later (Figure 4.5). The function RUN_ITERATION( ) splits the dataset into training and test data by implementing a k-fold cross-validation (by default $k = 10$) process to assess the performance of the current set of attributes. A k-fold cross-validation scheme was preferred, rather than the leave-one-out used in the previous RGIFE version, because of its better results when it comes to model selection [30]. In here, to describe the RGIFE heuristic, the generic term *performance* will be used to refer to how well the model can predict the class of the test samples. In reality, within RGIFE many different measures can be employed to estimate the model performance (accuracy, F-measure, AUC, etc.).

Algorithm 1 RGIFE: Rank Guided Iterative Feature Elimination

Input: dataset *data*, cross-validation repetitions N
Output: selected attributes

```
1:
2: function REDUCE_DATA(data)
3:     numberOfAttributes ← current number of attributes from data
4:     ▷ If blockSize is larger than the attributes reduce it (and check for soft-fail)
5:     if (startingIndex + blockSize) > numberOfAttributes then
6:         blockRatio = blockRatio × 0.25
7:         blockSize = blockRatio × numberOfAttributes
8:     end if
9:     attributesToRemove ← attributesRanking[startingIndex : (startingIndex +
   blockSize)]
10:    reducedData ← remove attributesToRemove from data
11:    startingIndex = startingIndex + blockSize
12:    return reducedData
13: end function
14:
15: function RUN_ITERATION(data)
16:    for N times do
17:        ▷ generate training and test set folds from data
18:        performances ← cross-validation over data
19:        attributesRank ← get the attributes ranking from the models
20:    end for
21:    performance = average(performances)
22:    attributesRank = average(attributesRank)
23:    return performance, attributesRank
24: end function
25:
26: blockRatio = 0.25
27: blockSize = blockRatio× (attributes in data)
28: startingIndex = 0
29: performance, attributesRank = RUN_ITERATION(data)
30: referencePerformance = performance
31:
32: while blockSize ≥ 1 do
33:    data = REDUCE_DATA(data)
34:    numberOfAttributes ← current number of attributes from data
35:    performance, attributesRank = RUN_ITERATION(data)
36:    if performance < referencePerformance then
37:        failures = failures + 1
38:        if (failures = 5) OR (all attributes have been test) then
39:            if there exist a soft-fail then
40:                referencePerformance = softFailPerformance
41:                numberOfAttributes, selectedAttributes ← attributes of the
   dataset at the softFail iteration
42:                blockSize = blockRatio × numberOfAttributes
```

```
43:             else
44:                 blockRatio = blockRatio × 0.25
45:                 blockSize = blockRatio × numberOfAttributes
46:             end if
47:             failures = 0; startingIndex = 0
48:         end if
49:     else
50:         referencePerformance = performance
51:         selectedAttributes ← current attributes from data
52:         blockSize = blockRatio × numberOfAttributes
53:         failures = 0; startingIndex = 0
54:     end if
55: end while
56: return selectedAttributes
```

The N parameter indicates how many times the cross-validation process is repeated with different training/test partitions, to minimise the potential bias introduced by the randomness of the data partition. The generated model (classifier) is then exploited to rank the attributes based on their importance within the classification task. Afterwards, the block of attributes at the bottom of the rank is removed and a new model is trained over the remaining data (lines 33-35). The number of attributes to be removed in each iteration is defined by two variables: *blockRatio* and *blockSize*. The former represents the percentage of attributes to remove (that decreases under certain conditions), the latter indicates the absolute number of attributes to remove and is based on the current size of the dataset. Then, if the new performance is equal or better than the reference (line 49), the removed attributes are permanently eliminated. Otherwise, the attributes just removed are placed back in the dataset. In this case, the value of *startingIndex*, a variable used to keep track of the attributes been tested for removal, is increased. As a consequence, RGIFE evaluates the removal of the next *blockSize* attributes, ranked (in the reference iteration) just after those placed back. The *startingIndex* is iteratively increased, in increments of *blockSize*, if the lack of the successive blocks of attributes keeps decreasing the predictive performance of the model. With this iterative process, RGIFE evaluates the importance of different ranked subsets of attributes. Whenever either all the attributes of the current dataset have been tested (i.e. have been eliminated and the performance did not increase), or there has been more than 5 consecutive unsuccessful iterations (i.e. performance was

degraded), *blockRatio* is reduced by a fourth (line 44). The overall RGIFE process is repeated while *blockSize* (number of attributes to remove) is ≥ 1.

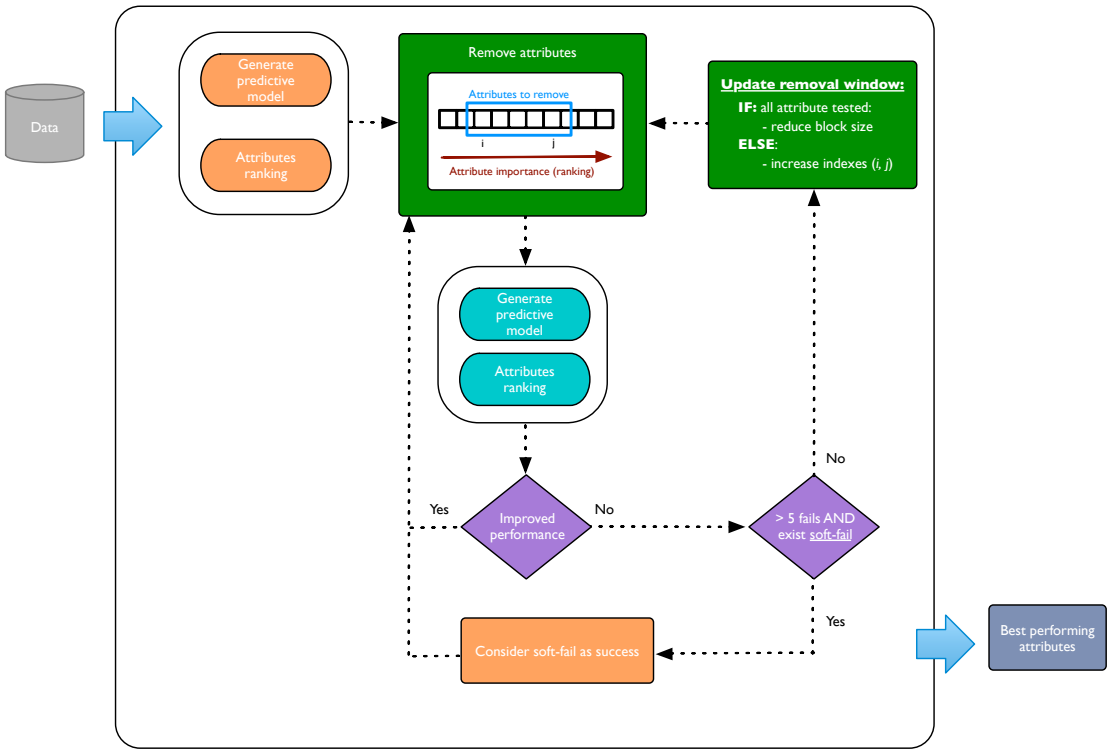


Fig 4.2: The iterative nature of the RGIFE heuristic and its overall behaviour.

An important characteristic of RGIFE is the concept of the*soft-fail*. After five unsuccessful iterations, if some past trial failed and suffered a "small" drop in performance (one misclassified sample more than the reference iteration) it is still considered successful (line 40). The reason behind this approach is that by accepting a temporary small degrade in performance, the probability of incurring in a local optimum is reduced. Thus, the likelihood of obtaining better solutions is increased. Given the importance of the soft-fail, as illustrated later in Section 4.3.4, in this new RGIFE implementation, the searching for the soft-fail is not only performed when five consecutive unsuccessful trials occur, as in the original version, but it occurs before every reduction of the block size. Furthermore, the iterations that are tested for the presence of a soft-fail are extended. While before only the last five iterations were analysed, now the searching window is expanded up to the most recent between the reference iteration and the iteration in which the last soft-fail was found.

4.2.1.1 Relative block size

One of the main changes introduced in this new version of the heuristic is the adoption of a *relative block size*. The term block size defines the number of attributes that are removed in each iteration. In [208], the 25% of the attributes was initially removed, then whenever having: all the attributes been tested, or five consecutive unsuccessful iterations, the block size was reduced by a fourth. However, the analysis suggested that this approach was prone to get stalled early in the iterative process and prematurely reduce the block size to a very small number. This scenario either slows down the iterative reduction process because successful trials will only remove few attributes (small block size), or it prematurely stops the whole feature reduction process if the size of the dataset being analysed becomes too small (few attributes) due to large chunks of attributes being removed (line 32 in Algorithm 1). To address this problem, the new implementation of the heuristic introduces the concept of the relative block size. By using a new variable called *blockRatio*, the number of attributes to be removed is now proportional to the size of the current attribute set being processed, rather than to the original attribute set. While before the values of *blockSize* were predefined (given the original attribute set), now they vary based on the size of the data in hand. Preliminary tests (not reported) showed that this block size policy is much more reliable.

4.2.1.2 Parameters of the classifier

RGIFE can be used with any classifier that can provide an attribute ranking after the training process. The presented version of RGIFE uses a random forest classifier that is known for its robustness to noise and its efficiency, so it is ideally suited to tackle biomedical data. Furthermore, as suggested in [209], random forest tends not to overfit, incorporates interactions among predictor variables, can be easily used when the number of features is extremely larger than the observations and can tackle both binary and multi-class problems. The current version of the heuristic is implemented in Python and uses the random forest classifier available in the *scikit-learn* library [210]. In this package the attributes, by default, are ranked based on the *gini impurity*. The gini impurity represents the expected error rate at a node M if the category label is selected randomly from the class distribution present at M. The feature importance

is calculated as the sum over the number of splits (across every tree) that include the feature, proportionally to the number of samples it splits. Default values for all the parameters of the classifier are used within the heuristic, except for the number of trees (set to 3000 because it provided the best results in preliminary tests not reported here). The attribute importance based on entropy was also tested, but considering that it did not produce any improvement in performance, the default criteria was chosen.

4.2.1.3 RGIFE policies

The current version of the heuristic uses a random forest as core classifier, rather than BioHEL as originally proposed [208]. The random forest is a stochastic ensemble classifier as each decision tree is built by using a random subset of features. As a consequence, RGIFE inherits this stochastic nature, that is each run of the algorithm results in a potentially different optimal subset of features. The presence of multiple optimal solutions is a common scenario when dealing with high dimensional -omics data [211]. Therefore, this situation is addressed by running RGIFE multiple times and using different policies to select the final model (signature):

- *RGIFE-Min*: the final model is the one with the smallest number of attributes
- *RGIFE-Max*: the final model is the one with the largest number of attributes
- *RGIFE-Union*: the final model is the union of the models generated across different executions

In the presented analysis the signatures were identified from 3 RGIFE runs.

4.2.2 Benchmarking algorithms

RGIFE has been compared with five well known feature selection algorithms: CFS [104], SVM-RFE [111], ReliefF [212], Chi-Square [213] and L1-based feature selection [214]. These algorithms were chosen in order to cover the different approaches that can be used to tackle the feature selection problem, each of them employs a different strategy to identify the best subset of features.

CFS is a multivariate correlation-based feature selection method. By exploiting a best-first search, it assigns high scores to subsets of features highly correlated to the class attribute but with low correlation between each other. Similarly to RGIFE, CFS automatically identifies the best size of the signature.

SVM-RFE is a well known iterative feature selection method that employs a backwards elimination procedure. The method ranks the features by training an SVM classifier (linear kernel) and discarding the least important (last ranked). SVM-RFE has been successfully applied in classification problems with -omics and in general biomedical datasets.

ReliefF is an extension of the Relief algorithm proposed by Kira and Rendell [106]. ReliefF is a supervised learning algorithm that considers global and local feature weighting by computing the nearest neighbours the samples. The feature importance is calculated by checking which features differs (in terms of values) between samples of different classes. This method is well employed due to its fast nature as well with its simplicity.

Chi-Square is a univariate feature selection approach that computes chi-squared (χ^2) stats between each feature and class. The score can be used to select the K attributes with the highest values for the chi-squared statistic calculated from the classes.

L1-based feature selection uses a linear model penalised with the L1 norm to identify relevant attributes [214]. The L1 norm tends to generate sparse solutions (models) where many of the estimated coefficients are zero. The features with a non-zero coefficient are selected because of their importance when predicting the outcome (class label) of the samples. A linear support vector classifier (SVC) was used to generate the linear penalised model.

The L1-based feature selection was evaluated using the *scikit-learn* implementation of the SVC [210], the other benchmarking algorithms were tested with their implementation available in WEKA [215]. Default parameters were used for all the methods.

4.2.3 Datasets

Synthetic datasets

The ability of RGIFE to identify relevant features was tested using a large set of synthetic datasets. The main characteristics of the data are available in Table 4.1. Different possible scenarios (correlation, noise, redundancy, non-linearity, etc.) were covered using the datasets employed in [103] as a reference (the LED data were not used as they consist of a 10-class dataset that does not reflect typical biological problem).

CorrAL is a dataset with 6 binary features (i.e. $f_1, f_2, f_3, f_4, f_5, f_6$) where the class value is determined as $(f_1 \wedge f_2) \vee (f_3 \wedge f_4)$. The feature f_5 is irrelevant while f_6 is correlated to the class label by 75%. In addition, the data contains 93 irrelevant features randomly added [216].

XOR-100 includes 2 relevant and 97 irrelevant (randomly generated) features. The class label consists of the XOR operation between two features: $(f_1 \oplus f_2)$ [216].

Parity3+3 describes the problem where the output is $f(x_1, ...x_n) = 1$ if the number of $x_i = 1$ is odd. The *Parity3+3* extends this concept to the parity of three bits and uses a total of 12 attributes [103].

Monk3 is a typical problem of the artificial robot domain. The class label is defined as $(f_5 = 3 \wedge f_4 =) \vee (f_5 \neq 4 \wedge f_2 \neq 3)$ [217].

SD1, SD2 and SD3 are 3-class synthetic datasets where the number of features (around 4 000) is higher than the number of samples (75 equally split into three classes) [218]. These characteristic try to reflect the problematic of microarray data. They contain both full class relevant (FCR) and partial class relevant (PCR) features.

FCR attributes serve as candidate biomarkers to distinguish all the cancer types, while PCR discriminates subsets of cancer types. SD1 includes 20 FCR and 4 000 irrelevant features. The FCR attributes are divided into two groups of ten, genes in the same group are redundant. The optimal solution consists of any two relevant feature coming from different groups. SD2 includes 10 FCR, 30PCR and 4 000 irrelevant attributes. The relevant genes are split in groups of ten; the optimal subset should combine one gene from the set of FCRs and three genes from the PCRs, each one from a different group. Finally, SD3 contains only 60 PCRs and 4 000 irrelevant features. The 60 PCRs are grouped by ten, the optimal solution consists of six genes, one from each group. Collectively, SD1, SD2 and SD3 will be referred as the SD datasets.

Madelon is a dataset used in the NIPS'2003 feature selection challenge [219]. The relevant features represent the vertices of a 5-dimensional hypercube. 495 irrelevant features are added either from a random gaussian distribution or multiplying the relevant features by a random matrix. In addition, the samples are distorted by flipping labels, shifting, rescaling and adding noise. The characteristic of Madelon is the presence of many more samples (2400) than attributes (500).

Name	Attributes	Samples	Characteristics
CorrAL [216]	99	32	Corr.; F $\gg$ S
XOR-100 [216]	99	50	N.L; F $\gg$ S
Parity3+3 [103]	12	64	NL
Monk3 [217]	6	122	No.
SD1 [218]	4020	75	F $\gg$ S
SD2 [218]	4040	75	F $\gg$ S
SD3 [218]	4060	75	F $\gg$ S
Madelon [219]	500	2400	N.L; No.

Table 4.1: Description of the synthetic datasets used in the experiments. *Corr.* stands for correlation, *N.L* indicates nonlinearity, $F \gg S$ is used for datasets where the number of features is higher than the number of samples and *No.* represents noisy data.

Furthermore, two biological conditions (control and case) synthetic microarray datasets were generated using the *madsim* R package [220]. Madsim is a flexible microarray data simulation model that creates synthetic data similar to those observed with common platforms. Twelve datasets were created using default parameters but varying in

terms of number of attributes (5 000, 10 000, 20 000 and 40 000) and percentage of up/down regulated genes (1%, 2% and 5%). Each dataset contained 100 samples equally distributed in controls and cases.

Real-world datasets

RGIFE was evaluated using ten different cancer-related transcriptomics datasets, see Table 4.2). These datasets represent a broad range of characteristics in terms of biological information (different types of cancers), number of samples and number of attributes (genes).

Name	Attributes	Samples
Prostate-Sboner [221]	6144	281
Dlbcl [166]	7129	77
CNS [167]	7129	60
Leukemia [94]	7129	72
Prostate-Singh [170]	12600	102
AML [171]	12625	54
Colon-Breast [172]	22283	52
Bladder [222]	43148	166
Breast [223]	47293	128
Pancreas [224]	54675	78

Table 4.2: Description of the real-world datasets used in the experiments.

4.2.4 Experimental design

4.2.4.1 Relevant features identification

The scoring measure proposed by Bolon et. al [103] was used to compute the efficacy of the different feature selection methods in identifying important features from synthetic data. The *Success index* aims to reward the identification of relevant features and penalise the selection of irrelevant ones:

$$Success\,Index = 100 \times \left(\frac{R_s}{R_t} - \alpha\frac{I_s}{I_t}\right)\ ; \alpha = min\left\{\frac{1}{2}, \frac{R_t}{I_t}\right\}$$

where R_s and R_t are the number of relevant features selected and the total number of relevant features. Similarly, I_s and I_t represent the number of selected and the total number of irrelevant features.

4.2.4.2 Predictive performance validation

While CFS and the L1-based feature selection automatically identify the optimal set of features, the other algorithms require to specify the number of feature to retain. To obtain a fair comparison, this parameter was set to be equal to the number of attributes selected by the RGIFE's Union policy (as it generates by definition the largest signature among the policies).

The most common metric to assess the performance of a feature selection method is by calculating the accuracy when predicting the class of the samples. A typical n-fold cross-validation scheme randomly divides the dataset D in n equally-sized disjoint subsets $D_1, D_2, ..., D_n$. In turn, each fold is used as test set while the remaining $n-1$ are used as training set. If the folds are forced to maintain the original distribution of the classes, the cross-validation is named as *stratified*. However, the stratified cross-validation does not take into account the presence of similar samples (clusters) within each class. This might lead to a distorted measure of the performances [30]. Dealing with transcriptomics datasets that have a small number of observations (e.g. CNS has only 60 samples), this distortion might also be be amplified. To avoid this problem, the DB-SCV (Distributed-balanced stratified cross-validation) scheme was adopted [31]. The original DB-SCV scheme was modified so that the residual samples are randomly assigned to the folds. A dataset with m samples, when using a n-fold cross-validation scheme has in total ($m \bmod n$) residual samples. By randomly assigning the residual samples to the folds, rather than sequentially as in the proposed DB-SCV, the validation schema can better estimate the predictive performance of unseen observations. A 10-fold DB-SCV was employed within RGIFE (line 17-18) with $N = 10$, the model performance was estimated using the accuracy metric).

All the feature selection methods were tested using a 10-fold DB-SCV scheme. The methods were applied to the training sets and the results (selected attributed) were mirrored to the test sets. The predictive performance were assessed using four classifiers: Random Forest (RF), Gaussian Naive Bayes (GNB), Support Vector Machine (SVM) (with a linear kernel) and K-nearest neighbour (KNN). Each classifier uses different approaches and criteria to predict the label of the samples, therefore

the predictive performance of each method was tested in various classification scenarios. The random forest uses an ensemble of decision trees to perform the prediction, the SVM tries to define an hyperplane that maximises the distance between objects of different classes, GNB is a naive bayes classifier that assume that the continuous values associated with each class are distributed according to a Gaussian distribution. Finally, the KNN classifies the instances by looking at the K closest neighbours. All the classifiers were employed using the *scikit-learn* implementation with default parameters, except for the depth of the random forest trees, which was set to 5 to avoid overfitting (given the relatively small number of attributes in each signature). The stochastic nature of RF was addressed by generating ten different models for each training set and defining the predicted class via a majority vote.

4.2.4.3 Biomedical relevance analysis of the signatures

The biomedical importance of the signatures, generated by different methods, was validated using the *Prostate-Singh* dataset [170] as a case study. The biomedical relevance was assessed examining the role of the signatures' genes: in a cancer-related context, in a set of independent prostate-related datasets and within a Protein Protein Interaction (PPI).

Gene-disease associations To assess the relevance of the signatures within a cancer-related context, it was checked whether their genes were already known to be associated with a specific disease (G-D association). From the literature, it was retrieved the list of genes known to be associated with prostate cancer. Two sources of information were used: *Malacards* (a meta-database of human maladies consolidated from 64 independent sources) [139] and the union of 4 manually curated databases (OMIM [143], Orphanet [146], Uniprot [145] and CTD [144]). Using the number of disease-associated genes included in the signatures, *precision*, *recall* and *F-measure* were calculated. The *precision* is the fraction of genes that are associated to the disease, while the *recall* is the fraction of disease-associated genes (from the original set of attributes) included in the signature. Finally the *F-measure* is calculated as the harmonic mean of *precision* and *recall*.

Gene relevance in independent datasets The public prostate cancer databases were searched to verify if the genes selected by the different methods are relevant also in data that were not used for the inference of the signatures. Eight prostate cancer related datasets were selected from the cBioPortal for Cancer Genomics [195]: *SUC2, MICH, TCGA, TGCA 2015, Broad/Cornell 2013, MSKCC 2010, Broad/Cornell 2012* and *MSKCC 2014*. The aim was to check if the selected genes were genomically altered in the samples of the independent data. For each method and each independent dataset, was calculated the average fraction of samples with genomic alterations for the identified biomarkers. To consider the different size of each signature, the values have been normalised across methods (i.e. divided by the number of selected genes).

Signature induced network A part of the validation of the signatures was based on its analysis in a network context. Possible interactions between the genes selected by RGIFE were assessed. To verify it, a signature induced network was generated from a PPI network by aggregating all the shortest paths between all the genes in the signature. If multiple paths existed between two genes, the path that overall (across all the pairs of genes) was the most used was included. The paths were extracted from the PPI network employed in [115] that was assembled from 20 public protein interaction repositories (BioGrid, IntAct, I2D, TopFind, MolCon, Reactome-FIs, UniProt, Reactome, MINT, InnateDB, iRefIndex, MatrixDB, DIP , APID, HPRD, SPIKE, I2D-IMEx, BIND, HIPPIE, CCSB), removing non-human interactions, self-interactions and interactions without direct experimental evidence for a physical association.

4.3 Results

This section presents the analysis performed to evaluate the RGIFE heuristic. First, RGIFE is compared with its original version proposed by Swan et al. Then, five well-established approaches are used to benchmark RGIFE from both a computational and a biomedical point of view. The comparison is performed analysing synthetic and real-world (cancer-related transcriptomics) datasets. Finally, using a prostate cancer dataset as a case study, the relevance of the biomedical knowledge associated with the genes selected by RGIFE is assessed.

4.3.1 Comparison with the original heuristic

The presented RGIFE heuristic is an extension of the work proposed in [208]. The main substantial changes involved: the use of a different base classifier (from Bio-HEL [39] to a random forest) and therefore a different attributes ranking criteria, the adoption of different block size policy (from absolute to relative) and the employment of a more robust validation scheme within the reduction process (from LOOCV to DB-SCV). Therefore, the first natural step for the validation of the new RGIFE (RGIFE-RF) was to compare it to its original version, in here named RGIFE-BH. The predictive performance of the signatures identified by the two versions were compared using a 10-fold cross-validation. That is, calculating the accuracy of the attributes selected from the training set when trying to predict the class of the test set samples. In Figure 4.3 is shown the distribution of accuracies obtained using the 10 datasets presented in Table 3.2. The accuracy of RGIFE-BH is calculated as the average of the values obtained over three runs of RGIFE-BH (same number of executions employed to identify the final models with RGIFE-RF). The predictive performance was assessed with four different classifiers. Across different datasets and classifiers, RGIFE-BH performed similarly or worse than the new proposed policies based on a random forest. To establish whether the difference in performance was statistically significant, the Friedman rank based test was used followed by a Nemenyi post-hoc correction. This is a well-known approach in the machine learning community when it comes to the comparison of multiple algorithms over multiple datasets [225]. The Friedman test is a non-parametric equivalent of the repeated-measures ANOVA to evaluate the significance of differences between multiple means (i.e. multiple classifiers). The ranks, for all the tested classifiers, are provided in Table 4.3. The attributes selected by RGIFE-BH performed quite well when using a random forest, while for the remaining tested classifiers the performance were low. In particular, RGIFE-BH obtained statistically significant worse results (confidence level of 0.05), compared with RGIFE-Union, when analysed with the KNN classifier. Overall, the best new RGIFE policy appears to be RGIFE-Union being ranked as first with all the tested classifiers except with SVM-Linear, second best in that instance.

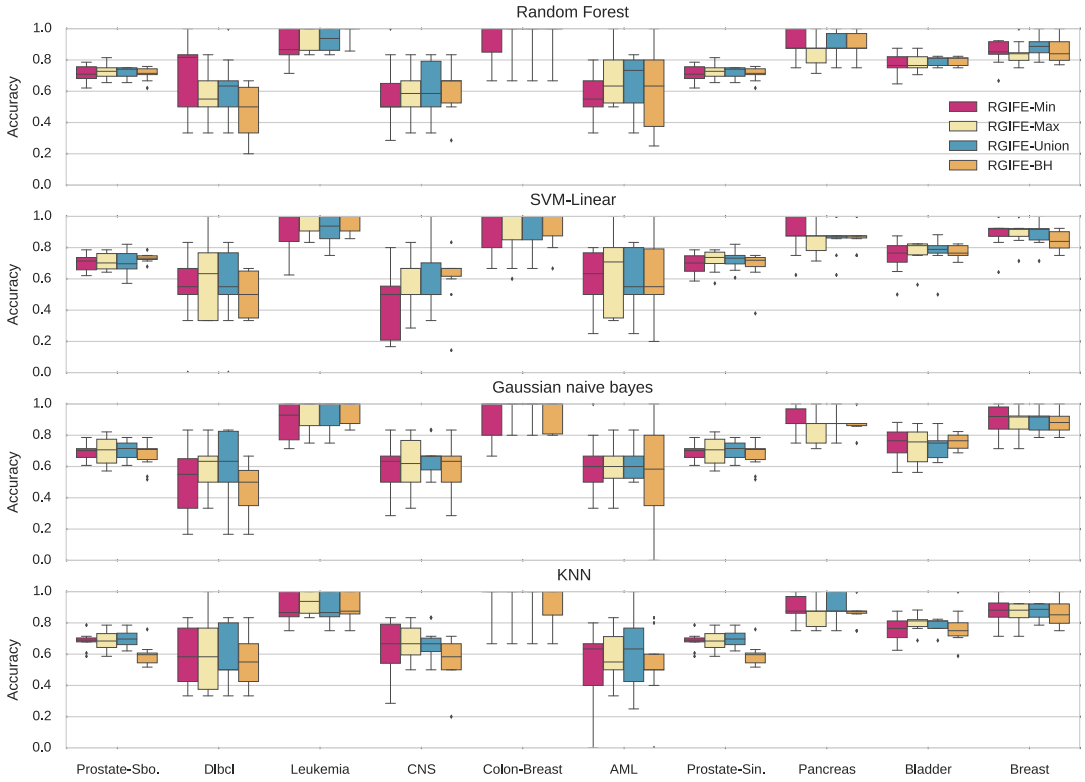


Fig 4.3: Distribution of the accuracies, calculated using a 10-fold cross-validation, for different RGIFE policies. Each subplot represents the performance, obtained with ten different datasets, assessed with four classifiers.

It might be tempting to associate the better performance of the new heuristic with the usage of a better base classifier. However, this is not the case as, when tested with a standard 10-fold cross-validation (using the presented 10 transcriptomics datasets with the original set of attributes), random forest and BioHEL obtained statistically equivalent accuracies when using the Wilcoxon rank-sum statistic (a non-parametric test which ranks the differences in performances of two classifiers for each dataset and compares the ranks). In fact, on average, the accuracy associated with the random

Classifier	RGIFE-Min	RGIFE-Max	RGIFE-Union	RGIFE-BH
Random Forest	3.15 (4)	2.60 (3)	**1.85 (1)**	2.40 (2)
SVM-Linear	3.10 (4)	**1.60 (1)**	2.40 (2)	2.90 (3)
Gaussian naive bayes	2.70 (3)	2.65 (3)	**1.75 (1)**	2.90 (4)
KNN	2.70 (3)	2.20 (2)	**1.80 (1)**	3.30 (4)*

Table 4.3: The average performance ranks obtained by each RGIFE policy across the ten datasets using four classifiers. The highest ranks are shown in bold. * indicates statistically worse performance.

forest was only higher by 1.62 when compared to the performance of BioHEL. The accuracies obtained by the methods are presented in Table 4.4.

Dataset	Random Forest	BioHEL
Prostate-Sbo.	74.0 ± 02.3	74.9 ± 03.1
Dlbcl	59.7 ± 19.1	55.3 ± 17.1
Leukemia	98.6 ± 04.3	94.6 ± 06.9
CNS	63.7 ± 11.7	64.5 ± 11.6
AML	68.7 ± 15.8	62.5 ± 11.7
Prostate-Singh	91.3 ± 10.3	91.4 ± 08.3
Pancreas	89.8 ± 05.1	87.3 ± 06.2
Bladder	80.6 ± 02.1	80.0 ± 02.8
Colon-Breast	94.7 ± 11.1	92.7 ± 12.4
Breast	86.0 ± 07.1	87.7 ± 07.9

Table 4.4: BioHEL and random forest classification accuracy for each dataset calculated using a 10-fold cross-validation with the original set of attributes.

Afterwards, the number of attributes selected by different RGIFE policies were contrasted when using different datasets. Figure 4.4 provides the average number of attributes selected, across the folds of the cross-validation, by the original and the new proposed version of RGIFE. The attributes associated with RGIFE-BH are averaged across its three different executions. In each of the analysed dataset, the new version of the heuristic was able to obtain a smaller subset of predictive attributes while providing higher accuracies. The better performance of the new heuristic is likely the result of the less aggressive reduction policy introduced by the relative block size. By removing chunks of attributes whose sizes are proportional to the volume of the dataset being analysed, the heuristic is more prone to improve the predictive performance across iterations. Moreover, by guaranteeing more successful iterations, a smaller set of relevant attributes can be identified. The difference is particularly evident when analysing the largest datasets (in Figure 4.4 the datasets are sorted by increasing size).

Random forest is a faster classifier than BioHEL that is based on evolutionary learning. Thus, when comparing the execution time required by the new version of RGIFE and its original form, a huge improvement could be noticed. In Figure 4.5 are reported the average number of seconds, across the three executions of RGIFE for the 10-fold cross-

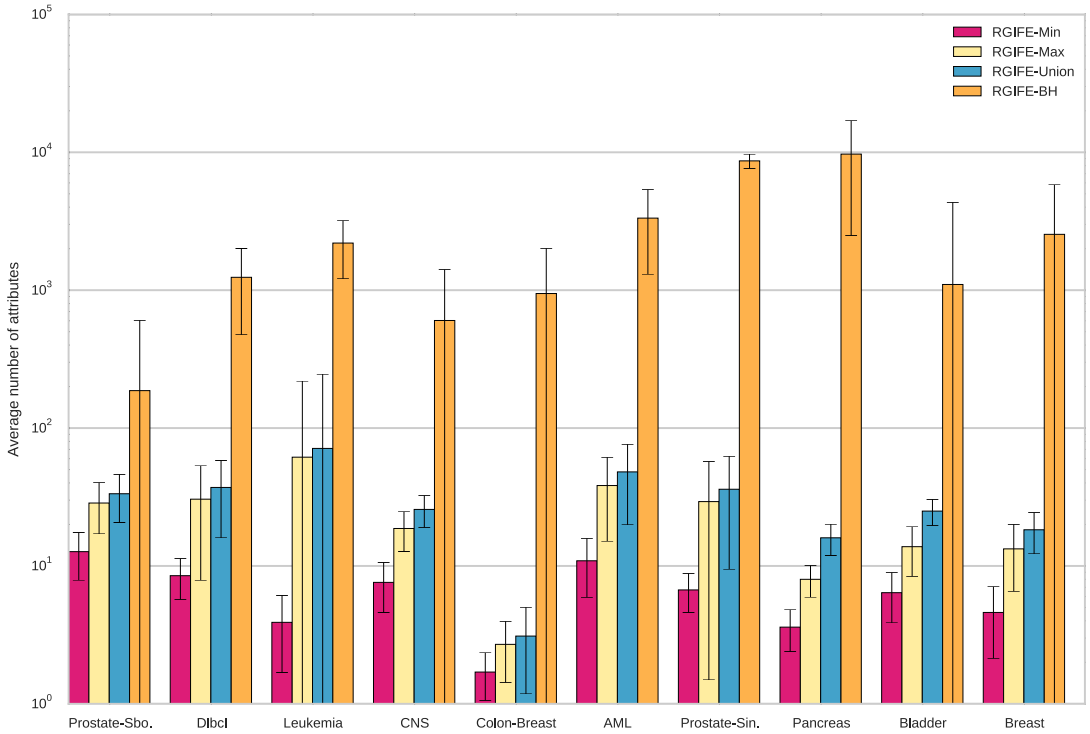


Fig 4.4: Comparison of the number of selected attributes by different RGIFE policies. For each dataset is reported the average number of attributes obtained from the 10-fold cross-validation together with the standard deviation.

validation experiments, required to select the optimal biomarker set. Across datasets of different size, in terms of number of attributes and samples, the new implementation of RGIFE is always at least 100 times faster than the heuristic based on BioHEL. The largest boost in performance was seen when analysing the Colon-Breast dataset. The differences, in execution time, seems to become slightly milder as the number of attributes in the datasets increase.

Overall, the changes introduced in the proposed version of RGIFE greatly improved its performance in terms of: computational time, number of selected attributes, predictive ability of the generated signatures.

4.3.2 Analysis of the RGIFE iterative reduction process

RGIFE is an iterative reduction heuristic where each iteration ends with two possible outputs: the predictive performance is either better/equal, compared to the reference iteration, or worse. Because of this, it is possible to visualise the behaviour of the whole

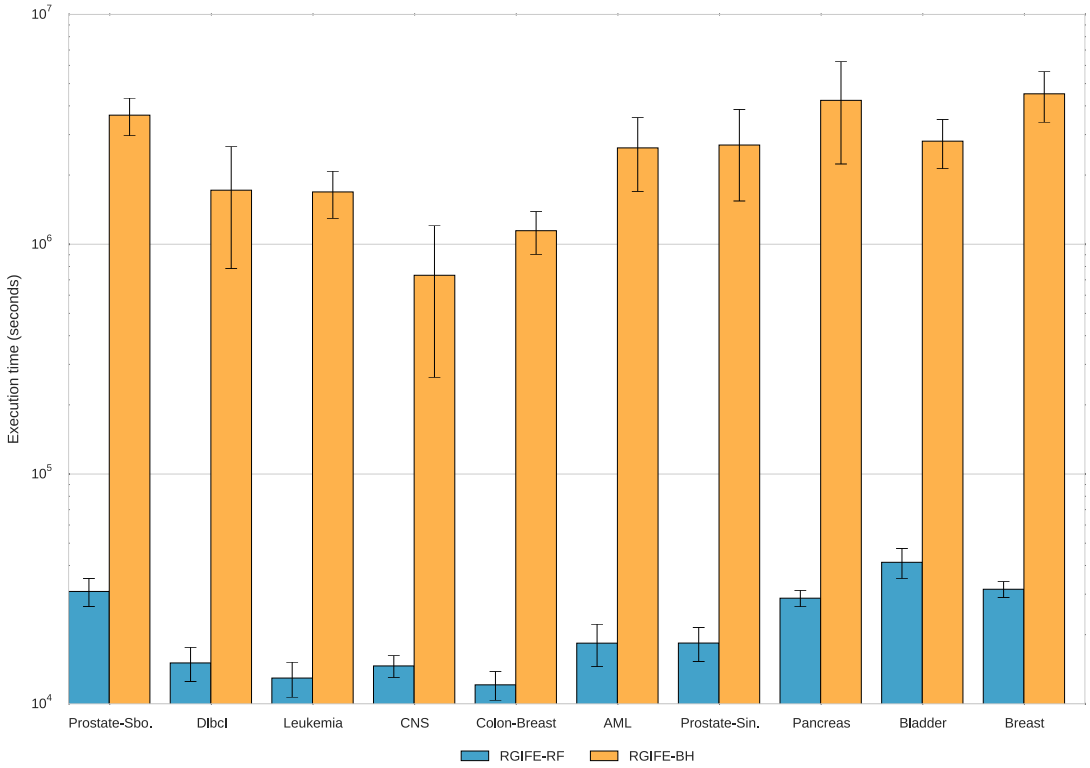


Fig 4.5: Comparison of the execution time (measured in seconds) between the original (RGIFE-BH) and the improved (RGIFE-RF) version of the heuristic. For each dataset is reported the number of seconds, from the 10-fold cross-validation tests, together with the standard deviation, required to identify the best performing attributes. The timing associated to each fold was calculated as the average across three different executions of RGIFE.

reduction process graphically. In Figure 4.6 it is illustrated the application of RGIFE to two datasets: *Breast* [223] and *AML* [171]. The plot shows the reduction process generated from three different runs of RGIFE to obtain the biomarker signature. A different colour represents a different output for the iteration: green and blue show an improvement (or equality) of the performance, blue is used when the removed attributes had not the lowest attribute importance (were not the bottom ranked). Red indicates a decrease in predictive performance, while a yellow square marks the identification of a soft-fail (a past iteration that suffered a small drop in performance). RGIFE looks for soft-fails if all the attributes have been tested (e.g. iteration 9 on Run 1 for *AML*) or there are five consecutive decreases in performance (e.g. iteration 41 on Run 3 for *AML*).

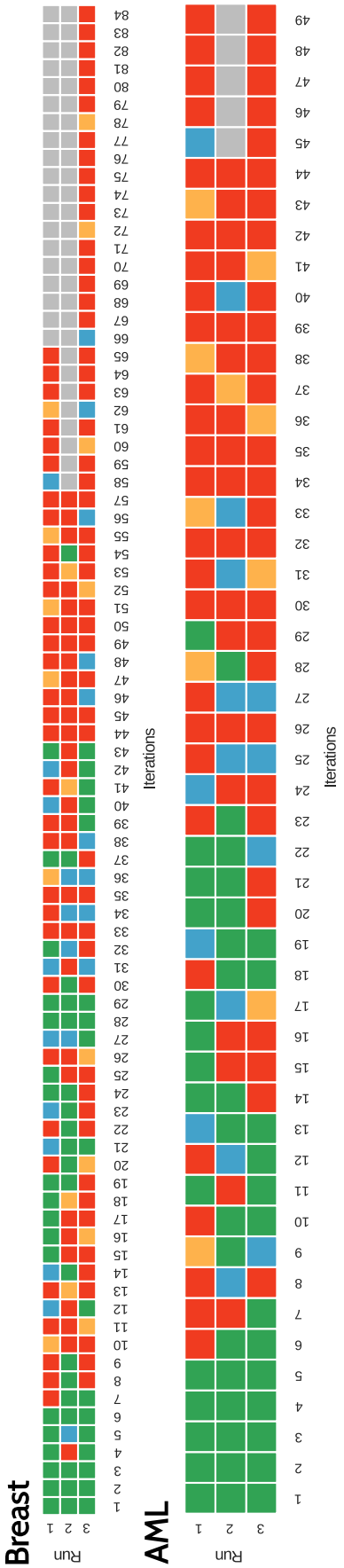


Fig 4.6: Result of each iteration during the iterative feature elimination process when applied to two datasets (*Breast* and *AML*) for three different runs of RGIFE. Green and blue indicate equal or better performance than the reference iteration. Green is used when the removed attributes were the lowest ranked, otherwise blue is employed. Red represents worse performance, yellow shows the identification of a soft-fail. The last non-grey square indicates the last iteration of the RGIFE run.

The Figure 4.6 shows the presence of several yellow marks, this is likely the result of the new strategies introduced to let the heuristic working with smaller data. In fact, differently than the original version, RGIFE now additionally performs a search for soft-fails before the block size is reduced. Furthermore, the iterations evaluated for the presence of a soft-fails are not anymore limited to the past five trials (as in the original version), but are extended up to the reference one (or the last trial in which a soft-fail was found). In many cases, after a soft-fail, RGIFE was able to produce smaller models with higher predictive capacity (e.g. iteration 37 on Run 1 and iteration 19 on Run 2 for *Breast*). Figure 4.6 also helps in highlighting the importance of restoring blocks of attributes back after an unsuccessful trial, which is an integral and novel part of RGIFE but not used in similar methods such as SVM-RFE. Across iterations, in both datasets, many blue squares are visible and indicate an increase of performance when the removed attributes were not the last ranked (lower attribute importance). Most of the methods based on an iterative reduction paradigm only remove the bottom placed features, however, as shown in Figure 4.6, discarding higher ranked features might lead to a better predictive model (e.g. iteration 14 on Run 1 for *Breast*). The examples just provided have emphasised the key role played by two of the main novel features introduced by the RGIFE heuristic: (a) the relevance of placing back features whose removal negatively affects the overall performance and (b) the importance of the soft-fail and its ability to drive the reduction process towards an easier and simpler solution.

4.3.3 Identification of relevant attributes in synthetic datasets

A large set of synthetic data was used to assess the ability of each method to identify relevant features in synthetic datasets. The *Success Index* was used to determine the success of discarding irrelevant features while focusing only on the important ones. Table 4.5 reports a summary of this analysis, the values correspond to the average Success Index obtained when using a 10-fold cross-validation. The last rows report the average Success Index and rank across the different datasets. The higher the Success Index, the better the method, 100 is its maximum value. In Section B.1 of

Appendix B are reported the accuracies of each method using four different classifiers.

Dataset	RGIFE-Min	RGIFE-Max	RGIFE-Union	CFS	ReliefF	SVM-RFE	Chi-Square	L1
CorrAL	59.93 (8)	77.11 (5)	**87.07 (1)**	84.57 (2)	72.04 (6)	64.53 (7)	82.06 (4)	84.11 (3)
XOR-100	**89.99 (1)**	79.88 (3)	89.88 (2)	24.72 (6)	49.86 (5)	14.84 (7)	9.84 (8)	79.30 (4)
Parity3+3	44.44 (3.5)	44.44 (3.5)	**76.67 (1.5)**	-5.93 (6)	**76.67 (1.5)**	-15.93 (8)	-8.52 (7)	5.56 (5)
Monk3	**84.17 (2.5)**	**84.17 (2.5)**	**84.17 (2.5)**	62.50 (6)	**84.17 (2.5)**	N/A	59.17 (7)	73.33 (5)
Madelon	59.98 (5)	77.98 (4)	87.97 (3)	17.99 (8)	89.97 (2)	23.97 (7)	39.97 (6)	**99.01 (1)**
Average	67.70	72.72	**85.15**	36.77	74.54	21.85	36.50	68.26
Average Rank	4.0±2.7 (5)	3.6±1.0 (3.5)	**2.0±0.8 (1)**	5.6±2.2 (6)	3.0±2.0 (2)	7.3±0.5 (8)	6.4±1.5 (7)	3.6±1.7 (3.5)

Table 4.5: Average Success Index calculated using a 10-fold cross-validation. In brackets are reported the rank of each method. The last row reports the average Success Index and rank across the five datasets. The highest indexes are shown in bold. N/A is used for SVM-RFE when analysing the *Monk3* dataset as the method cannot deal with categorical attributes.

RGIFE-Union is the method with the highest average Success Index, followed by RGIFE-Max and ReliefF. The Union policy clearly outperforms the other methods when analysing the *Parity3+3* and the *XOR-100* datasets. Overall, SVM-RFE seemed unable to discriminate between relevant and irrelevant features. Low success was also observed for CFS and Chi-Square. For the analysis of the SD datasets [218] are reported measures that are more specific for the problem. The SD datasets are characterised by the presence of relevant, redundant and irrelevant features. For each dataset, Table 4.6 includes the average number of: selected features, features within the optimal subset, irrelevant and redundant features.

Dataset	Metrics	RGIFE-Min	RGIFE-Max	RGIFE-Union	CFS	ReliefF	SVM-RFE	Chi-Square	L1
SD1	**Selected**	113.3	253.6	289.5	24.3	289.5	289.5	289.5	144.2
	OPT(2)	0.2	0.8	0.9	1.5	2.0	2.0	1.7	2.0
	Redundant	0.0	2.7	2.7	0.3	9.0	8.7	5.4	5.3
	Irrelevant	114.1	248.4	284.2	23.1	270.5	271	278.5	132.6
SD2	**Selected**	103.4	279.7	319.4	23.1	319.4	319.4	319.4	137.1
	OPT(4)	0.6	1.1	1.2	2.7	3.9	4.0	2.8	4.0
	Redundant	0.6	2.4	2.6	0.1	9	8.9	3.6	3.5
	Irrelevant	102.6	271.4	310.4	20.7	281.4	281.4	301.8	117.9
SD3	**Selected**	114.6	284.3	337.3	24.4	337.3	337.3	337.3	143.4
	OPT(6)	1.0	2.6	3.8	3.4	4.8	4.2	3.5	6.0
	Redundant	1.0	4.1	6.1	0.1	9.0	4.0	7.4	3.5
	Irrelevant	113.2	267.8	312.9	21.1	292.0	306.6	309.2	119.2

Table 4.6: Summary of the analysis on the SD datasets. The values are averaged from a 10-fold cross-validation. OPT(x) indicates the average number of selected features within the optimal subset.

The L1-based feature selection was the only method always able to select the minimum number of optimal features, however it also picked a large number of irrelevant

features. On the other hand, CFS was capable of avoiding redundant and irrelevant features while selecting a high number of optimal attributes. ReliefF, SVM-RFE and Chi-Square performed quite well for SD1 and SD2, but not all the optimal features were identified in SD3. The RGIFE policies performed generally poorly on the SD datasets. Among the three policies, RGIFE-Union selected the highest number of optimal features (together with a large amount of irrelevant information). Despite that, the number of redundant features was often lower than methods which selected more optimal attributes. Interesting, when analysing the accuracy obtained by each method (reported in Section B.1 Appendix B), it can be observed that the attributes selected by RGIFE-Union, although not completely covering the optimal subsets, provide the best performance for SD2 and SD3 (with random forest and GNB classifier). Finally, Table 4.7 shows the results from the analysis of the data generated with *madsim* [220]. The values have been averaged from the results of the data containing 1%, 2% and 5% of up/down regulated genes. Different from the SD datasets, there is not an optimal subset of attributes, therefore only the average number of relevant and irrelevant (not up/down-regulated genes) features are reported. The accuracies of each method (available in Section B.1 of Appendix B) were constantly equal to 1 for most of the methods regardless the classifier used to calculate them. Exceptions are represented by RGFE-Max, RGIFE-Min and Chi-Square. All the RGIFE policies performed better than CFS and L1 in terms of relevant selected attributes. Few up/down regulated attributes, compared with the dozens of the other two methods, were enough to obtain a perfect classification. In addition, RGIFE never used irrelevant genes in the proposed solutions. The other methods, whose number of selected attributes was set equal to that used by RGIFE-Union, performed equally well.

Overall, the analysis completed using synthetic datasets highlighted the ability of RGIFE, in particular of RGIFE-Union, to identify important attributes from data with different characteristics (presence of noise, nonlinearity, correlation, etc.). Good performance was also reached from data similar to microarray datasets (madsim). On the other hand, the SD datasets led to unsatisfactory RGIFE results. This can be attributed to the low number of samples (only 25) available for each class that

can generate an unstable internal performance evaluation (based on a 10-fold cross-validation) of the RGIFE heuristic.

Attributes	Metric	RGIFE-Min	RGIFE-Max	RGIFE-Union	CFS	ReliefF	SVM-RFE	Chi-Square	L1
5 000	**Rel.**	1.0	1.2	2.7	54.9	2.7	2.7	2.7	18.4
	Irr.	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0
10 000	**Rel.**	1.0	1.5	3.1	68.4	3.1	3.1	3.1	22.1
	Irr.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
20 000	**Rel.**	1.0	1.3	3.2	96.4	3.2	3.2	3.2	29.2
	Irr.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
40 000	**Rel.**	1.0	1.5	3.5	133.8	3.5	3.5	3.5	28.2
	Irr.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 4.7: Summary of the analysis on the *madsim* datasets. The values represent the average from the analysis of data containing 1%, 2% and 5% of up/down-regulated genes. For each set of data are reported the average number of relevant (up/down-regulated) and irrelevant attributes (from a 10-fold cross-fold validation).

4.3.4 Comparison of the predictive performance with other feature selection methods

Having established the better performance provided by the presented heuristic compared with its original version, and encouraged by the results obtained using synthetic datasets, RGIFE was evaluated with real-world biomedical data. For each dataset and base classifier, the accuracy of the biomarker signatures generated by each method was calculated. Then, the methods were ranked in ascending order (the higher the rank, the higher the accuracy). Table 4.8 reports all the resulting accuracies and the ranks (in brackets), the last column shows the average rank across the datasets. With 3 out of 4 classifiers, the presented heuristic was the first ranked (1 RGIFE-Max, 2 RGIFE-Union). ReliefF was the first ranked when evaluated with random forest (RF), while it performed quite poorly when using SVM. Similarly, RGIFE-Max was first and second ranked respectively with SVM and KNN, while it was the second and the third-worse with RF and gaussian naive bayes (GNB). In general, as already deducted from Table 4.3, RGIFE-Union is the best performing policy in terms of predictive capacity being ranked as first when tested with KNN and GNB. Conversely, RGIFE-Min constantly performed badly across different classifiers and datasets. To statistically compare the performances of the methods, the Friedman test was used again. In all the four scenarios there was no statistical difference in the performances of the tested

Class.	Method	Pro-Sbo.	Dlbcl	Leukemia	CNS	Colon-Breast	AML	Pro-Sin.	Pancreas	Bladder	Breast	Rank
RF	**RGIFE-Un.**	0.723 (6)	0.643 (3)	0.927 (7)	0.617 (5)	**0.947 (3)**	0.667 (3)	0.923 (3)	0.898 (3.5)	0.794 (4)	0.884 (2.5)	4.00 (4)
	RGIFE-Max	0.727 (4)	0.573 (7)	0.940 (5.5)	0.600 (6)	**0.947 (3)**	**0.680 (1)**	0.913 (6.5)	0.859 (8)	0.782 (6)	0.844 (8)	5.50 (6)
	RGIFE-Min	0.712 (8)	**0.680 (1)**	0.886 (8)	0.589 (7)	0.927 (7)	0.577 (8)	0.884 (8)	**0.900 (1.5)**	0.770 (8)	0.851 (7)	6.35 (8)
	CFS	**0.741 (1)**	0.627 (4)	0.957 (2.5)	0.622 (4)	**0.947 (3)**	0.597 (7)	0.922 (5)	0.886 (6)	0.800 (3)	0.869 (6)	4.15 (5)
	SVM-RFE	0.733 (2)	0.623 (5)	0.944 (4)	0.668 (3)	0.887 (8)	0.675 (2)	0.923 (3)	0.898 (3.5)	**0.819 (1)**	0.877 (4)	3.55 (2)
	ReliefF	0.726 (5)	0.577 (6)	**0.961 (1)**	0.681 (2)	0.930 (6)	0.633 (5)	**0.932 (1)**	**0.900 (1.5)**	0.800 (2)	**0.892 (1)**	**3.50 (1)**
	Chi-Square	0.716 (7)	0.660 (2)	0.940 (5.5)	0.520 (8)	**0.947 (3)**	0.622 (6)	0.913 (6.5)	0.886 (6)	0.776 (7)	0.870 (5)	5.60 (7)
	L1	0.730 (3)	0.520 (8)	0.957 (2.5)	**0.684 (1)**	**0.947 (3)**	0.650 (4)	0.923 (3)	0.886 (6)	0.788 (5)	0.884 (2.5)	3.80 (3)
SVM	**RGIFE-Un.**	0.709 (4)	0.523 (4.5)	0.917 (6)	0.565 (4)	0.927 (4)	0.633 (4)	0.895 (5)	0.861 (6)	0.757 (5)	0.892 (3)	4.55 (4)
	RGIFE-Max	**0.716 (1)**	0.573 (2)	0.957 (3.5)	0.572 (3)	**0.947 (1.5)**	0.617 (5)	0.915 (2)	0.873 (5)	0.775 (2)	0.876 (5)	**3.00 (1)**
	RGIFE-Min	0.694 (6)	0.567 (3)	0.908 (7)	0.421 (8)	0.907 (5.5)	0.585 (7)	0.852 (8)	0.875 (4)	0.744 (6)	**0.894 (1.5)**	5.60 (8)
	CFS	0.712 (3)	0.523 (4.5)	0.961 (2)	0.546 (5)	0.943 (3)	0.638 (3)	**0.952 (1)**	**0.911 (1)**	0.770 (3.5)	0.832 (8)	3.40 (2)
	SVM-RFE	0.644 (8)	0.500 (6)	0.942 (5)	0.699 (2)	0.850 (7)	0.500 (8)	0.894 (6)	0.848 (8)	0.770 (3.5)	**0.894 (1.5)**	5.50 (7)
	ReliefF	0.690 (7)	0.407 (8)	0.886 (8)	0.535 (6)	0.747 (8)	0.590 (6)	0.904 (4)	0.900 (2)	**0.795 (1)**	0.886 (4)	5.40 (6)
	Chi-Square	0.705 (5)	**0.603 (1)**	0.957 (3.5)	0.462 (7)	**0.947 (1.5)**	**0.652 (1)**	0.893 (7)	0.857 (7)	0.739 (8)	0.869 (6)	4.70 (5)
	L1	0.716 (2)	0.490 (7)	**0.988 (1)**	**0.746 (1)**	0.907 (5.5)	0.650 (2)	0.913 (3)	0.896 (3)	0.740 (7)	0.851 (7)	3.85 (3)
GNB	**RGIFE-Un.**	0.701 (2)	**0.623 (1)**	0.932 (5)	0.650 (3)	**0.963 (2.5)**	0.627 (4)	0.922 (4)	0.887 (5)	0.733 (6)	0.884 (4)	**3.65 (1)**
	RGIFE-Max	0.698 (3)	0.590 (3)	0.932 (5)	0.620 (5)	**0.963 (2.5)**	0.610 (6.5)	0.922 (4)	0.846 (8)	0.727 (7)	0.876 (6)	5.00 (7)
	RGIFE-Min	0.691 (5)	0.503 (5)	0.890 (8)	0.589 (7)	0.907 (6.5)	0.617 (5)	0.895 (7)	0.900 (2)	0.751 (4)	0.900 (3)	5.25 (8)
	CFS	0.690 (6)	0.520 (4)	**0.973 (1)**	0.665 (2)	0.927 (5)	0.650 (2)	**0.932 (1.5)**	0.871 (7)	0.740 (5)	0.870 (7)	4.05 (3)
	SVM-RFE	0.655 (7)	0.450 (8)	0.958 (3)	0.626 (4)	0.873 (8)	0.593 (8)	0.912 (6)	0.898 (3.5)	**0.807 (1)**	**0.907 (1)**	4.95 (6)
	ReliefF	0.616 (8)	0.473 (7)	0.919 (7)	0.570 (8)	0.907 (6.5)	0.633 (3)	**0.932 (1.5)**	0.898 (3.5)	0.764 (2)	0.901 (2)	4.85 (5)
	Chi-Square	0.694 (4)	0.593 (2)	0.932 (5)	0.620 (6)	**0.963 (2.5)**	0.610 (6.5)	0.922 (4)	**0.925 (1)**	0.727 (8)	0.878 (5)	4.40 (4)
	L1	**0.719 (1)**	0.500 (6)	0.971 (2)	**0.746 (1)**	**0.963 (2.5)**	**0.655 (1)**	0.885 (8)	0.886 (6)	0.753 (3)	0.855 (8)	3.85 (2)
KNN	**RGIFE-Un.**	**0.698 (1)**	0.593 (2)	0.901 (7.5)	0.665 (4)	**0.947 (2.5)**	0.602 (3)	0.894 (2.5)	**0.911 (1.5)**	0.788 (3)	0.876 (6)	**3.30 (1)**
	RGIFE-Max	0.684 (3)	**0.597 (1)**	0.927 (3)	0.670 (3)	**0.947 (2.5)**	0.582 (5)	0.893 (4.5)	0.861 (8)	**0.806 (1)**	0.862 (8)	3.90 (2)
	RGIFE-Min	0.684 (4)	0.587 (3)	0.901 (7.5)	0.635 (5)	**0.947 (2.5)**	0.528 (8)	**0.903 (1)**	0.886 (5.5)	0.758 (8)	0.876 (6)	5.05 (7)
	CFS	0.669 (6)	0.407 (8)	0.917 (5.5)	0.587 (8)	0.910 (6)	0.580 (6)	0.884 (6)	**0.911 (1.5)**	0.776 (5)	0.876 (6)	5.80 (8)
	SVM-RFE	0.662 (7)	0.523 (7)	0.946 (2)	**0.771 (1)**	0.847 (8)	0.562 (7)	0.875 (7)	0.898 (3.5)	0.801 (2)	**0.892 (1)**	4.55 (4.5)
	ReliefF	0.698 (2)	0.553 (5)	0.917 (5.5)	0.615 (7)	0.907 (7)	0.630 (2)	0.894 (2.5)	0.898 (3.5)	0.783 (4)	0.884 (2)	4.05 (3)
	Chi-Square	0.680 (5)	0.537 (6)	0.919 (4)	0.618 (6)	**0.947 (2.5)**	0.595 (4)	0.893 (4.5)	0.873 (7)	0.770 (6)	0.877 (3)	4.80 (6)
	L1	0.645 (8)	0.570 (4)	**0.973 (1)**	0.698 (2)	0.927 (5)	**0.713 (1)**	0.825 (8)	0.886 (5.5)	0.769 (7)	0.876 (4)	4.55 (4.5)

Table 4.8: Accuracies and ranks (in brackets) obtained by each method across the 10 datasets using 4 classifiers. The highest accuracies and ranks are shown in bold. The last column reports the average ranks across the datasets for each method, in brackets is shown the absolute ranks. The accuracies are rounded to the third decimal but the ranks are based on higher precision. RF: random Forest, KNN: K-nearest neighbour, GNB: Gaussian Naive Bayes.

methods. The only exception was ReliefF (first ranked) that statistically outperformed RGIFE-Min when using random forest (confidence level of 0.05).

4.3.5 Analysis of the signatures size

The size (number of the selected attributes) of the signatures generated by the RGIFE policies, CFS and the L1-based feature selection were compared. With methods such as ReliefF or SVM-RFE, the comparison is meaningless because the number of the selected features is a parameter that needs to be set up-front. The results, dataset by dataset, are shown in Figure 4.7. Each bar represents the average number of chosen features across the ten training sets of the cross-validation process described in Section 4.2.4.2. There is a clear and remarkable difference in the number of selected attributes by RGIFE and CFS, this is extreme in datasets such as *Colon-Breast* and *Pancreas*. The L1-based feature selection performed quite badly when applied to the smallest dataset (*Prostate-Sboner*). A large standard deviation can be noticed in the *Leukemia* dataset. This is due to a large signature (around 500 attributes) identified by RGIFE. This large number of attributes is associated with an early stopping condition reached by RGIFE (the block size was reduced too soon due to the impossibility to improve the performance of the reference iteration). As expected, the best performing policy is RGIFE-Min. When applying the Friedman test to the average signature size of the methods, RGIFE-Min and RGIFE-Max were statistically better than CFS and the L1-based approach with a confidence level of 0.01. Moreover, RGIFE-Min also statistically outperformed RGIFE-Union. Although the Union policy did not statistically outperform the other two methods, the results in Figure 4.7 show how it consistently selected fewer features.

4.3.6 Biomedical relevance of the signatures

When feature selection is applied to biomedical data, with the aim of discovering new biomarkers, the signature not only has to be small and highly predictive, but it also needs to contain relevant features. In this analysis, dealing with cancer transcriptomics datasets, it was necessary to assess if the selected genes are relevant in a disease/biological context. This goal was achieved using the *Prostate-Singh* dataset

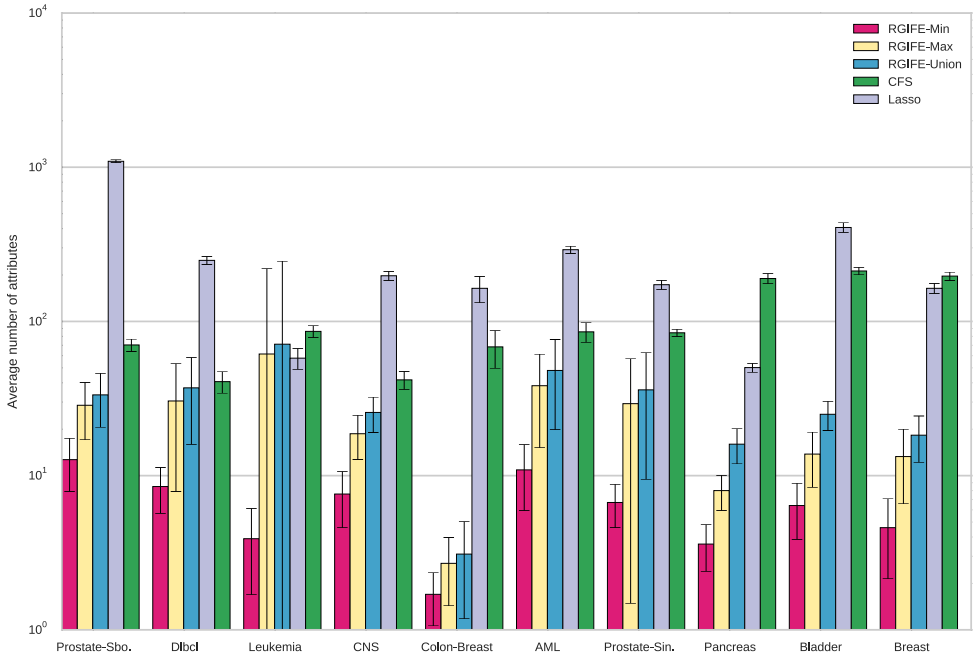


Fig 4.7: Comparison of the number of selected attributes by the RGIFE policies, CFS and L1-based feature selection. For each dataset is reported the average number of attributes obtained from the 10-fold cross-validation together with the standard deviation.

[170] as a case study. In the first part of this section, RGIFE is compared with the other methods, while later on, the focus is set on the signature generated by RGIFE-Union (the best performing policy). The signatures identified by each method are available in the Section B.2 of Appendix B.

Gene-disease association analysis

A similar approach to the analysis performed for the evaluation of FuNeL (Section 3.2.6) was used to assess the relevance of the signatures in a disease context. While in Chapter 3 the disease-associated genes were employed to analyse their relationship within FuNeL networks, in this section their presence within the signatures is evaluated. More specifically, it was assessed how many of the signature genes were already associated with prostate cancer. Precision, recall and F-score were calculated for each signature (see Section 4.2.4.3) using the information from two different sources (Malacards and manually curated data). The higher those metrics are, the better a feature selection algorithm performs as it can identify the relevant disease-associated

genes from the large set of original attributes. Figure 4.8 shows the performances for all the signatures generated in the case study.

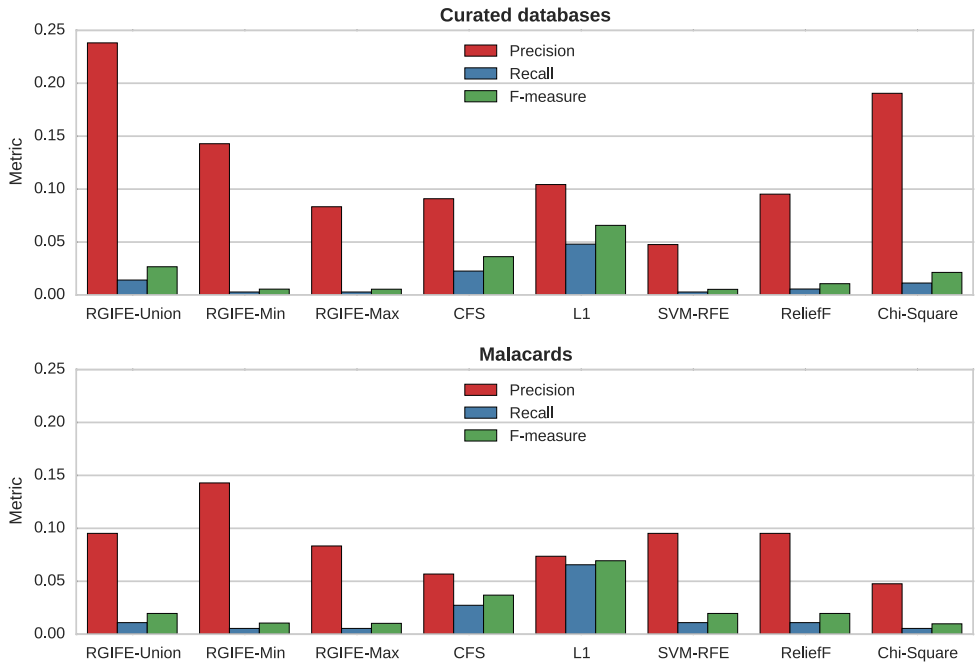


Fig 4.8: Analysis in a disease-context of the signatures selected by the methods. The G-D associations were retrieved from two different sources. Each metric is referred to the number of disease-associated genes available in the signatures.

When using the curated sources for the associations, RGIFE-Union had the higher precision followed by Chi-Square and RGIFE-Min. The other methods performed similarly except for SVM-RFE. High precision means that several genes selected by RGIFE-Union are already known to be relevant in prostate cancer. The recall was in general low for every method and was the highest for the L1-based feature selection (which also generates the largest signatures). Likewise, L1-selected attributes provided the largest F-measure, while similar values emerged for CFS and RGIFE-Union. The remaining approaches all scored low values. Using Malacards, and contrasting the performance with the curated analysis, SVM-RFE had higher precision, while similar (RGIFE-Min and ReliefF) or worse performances were obtained by the other methods. An important decrease was noticed for the L1-based feature selection and Chi-Square. Recall and F-measure did not vary a lot. In general, RGIFE policies tended to have higher or similar precision than the compared methods. RGIFE-Union provided overall the best results outperforming SVM-RFE, ReliefF and Chi-Square, its signature had

higher precision than CFS and L1 that, on the other hand, obtained higher recall (helped by the large number of selected attributes) and F-measure. RGIFE-Max and RGIFE-Min offered reasonable precision values while having a low recall; this is likely the result of the small set of attributes identified by the two policies.

Genomic alteration of the signatures in independent datasets

The genes selected by each method were checked if relevant in prostate cancer-related data that were not used during the learning process. The genomic alterations (mutation, deletion and amplification) of each signature were analysed in eight independent data, available from the cBioPortal [195]. The alterations averaged across the genes of the signatures are reported, dataset by dataset, in Figure 4.9. The L1-based feature selection method was excluded from this analysis as it generated a signature larger than the limit of 100 genes allowed for the queries in cBioPortal. To take into account the different size (number of attributes) of each signature, the percentages of alterations have been normalised. The methods are ranked by the increasing percentage of alteration (the same colours are used in different datasets). The bottom-right plot shows the average rank of each method across all the datasets (higher rank means higher alteration). The two methods selecting genes that are highly altered in independent data are SVM-RFE and RGIFE-Union, with the last one clearly outperforming the others in *SUC2*, *TGCA 2015* and *Broad/Cornell 2013*. Among the other algorithms, RGIFE-Max and CFS perform quite badly, overall they are the bottom ranked, while the remaining methods obtained similar performances. This analysis shows that RGIFE-Union selects genes that are not only highly predictive in the analysed dataset but also are largely altered in datasets, containing samples that are affected by the same disease, not used during the learning process.

In the next sections, the focus will be on the analysis of the signature generated by the RGIFE-Union policy (the best performing). The selected signature consists of 21 genes: *ANXA2P3, TGFB3, CRYAB, NELL2, MFN2, TNN, KIAA1109, PEX3, ATP6V1E1, HPN, HSPD1, LMO3, PTGDS, SLC9A7, SERPINF1, KCNN4, EPB41L3, CELSR1, GSTM2, EPCAM, ERG*.

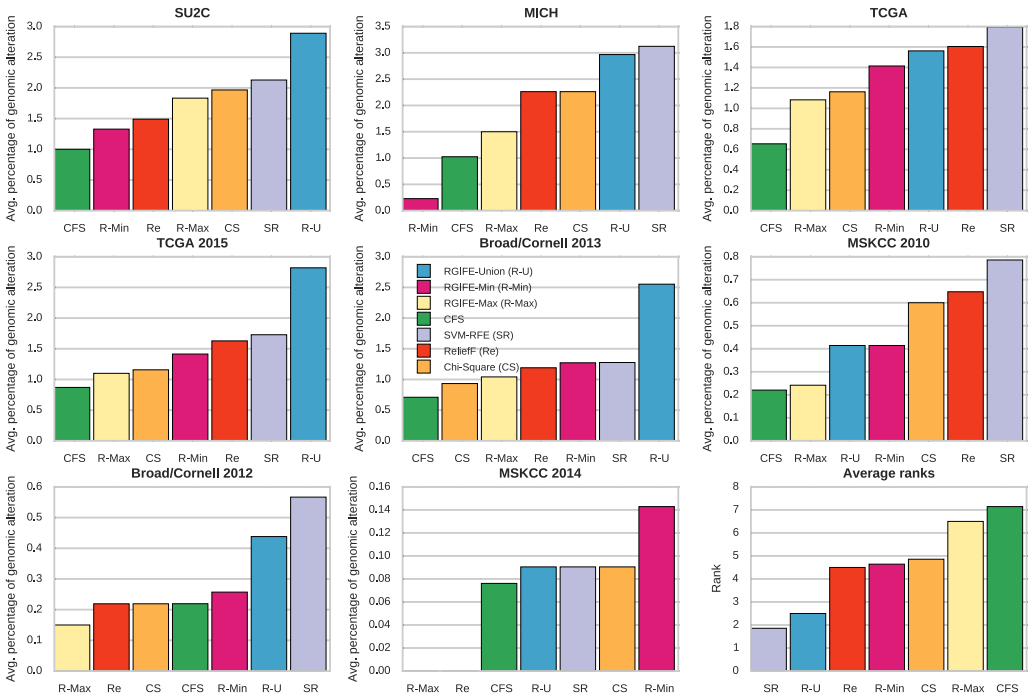


Fig 4.9: Normalised genomic alteration percentages of the signatures inferred for the case study. The alterations refer to the samples available from eight prostate cancer related datasets. The bottom-right plot shows the average ranks across the datasets. Higher rank indicates higher average alterations. The abbreviations and the colors for the plots are defined in the legend of the central subplot.

Gene-disease association from the specialised literature

The specialised literature was used to verify if the genes of the RGIFE signature are already associated with prostate cancer. Interestingly, many of them seem related to prostate cancer. Just to cite few examples:

- *NELL2* is an indicator of expression changes in prostate cancer samples [188], it also contributes to alterations in epithelial-stromal homeostasis in benign prostatic hyperplasia and codes for a novel prostatic growth factor [187].
- *ANXA2P3* (annexin II) was differentially expressed in prostate carcinoma samples from USA and India [226].
- *TGFB3* is expressed in metastatic prostate cancer cell lines and induces the invasive behaviour in these cell [227].

- *CRYAB* expression values can be used to discriminate between cancerous and non-cancerous prostatic tissues [228]
- *HSPD1* was part of a four gene expression signature to detect Gleason grade 3 and grade 4 cancer cells in prostate tissue [229].
- *EPB41L3* has a potential role as a target for treatment of advanced prostate cancer [230].

Afterwards, an enrichment analysis was performed on the RGIFE-Union signature. The enrichment analysis is a statistical-based method to assess if a set of genes share common biological characteristics. The analysis was conducted with the PANTHER classification system [175]; the knowledge base consisted of the PANTHER pathways: a set of 176 primarily signalling pathways. Four pathways resulted statistically (confidence value of 0.05) overrepresented in the signature:

- *Heterotrimeric G-protein signalling pathway-rod outer segment phototransduction* (P00028)
- *B cell activation* (P00010)
- *T cell activation* (P00053)
- *Heterotrimeric G-protein signalling pathway-Gi alpha and Gs alpha mediated pathway* (P00026)

Their role in prostate cancer appear to be relevant from the specialised literature. In particular, the family of heterotrimeric proteins is involved in prostate cancer invasion [231] and the (G protein)-coupled receptors (GPCRs) may contribute to tumour growth [232]. B-cells are increased in prostate cancer tissues according to a research by Woo et al. [233] and lymphotoxin derived by those cells can promote castration-resistant prostate cancer [234]. Finally, chimeric antigen receptor-engineered T cells have the potential to be used for the treatment of metastatic prostate cancer [235].

Signature induced network

As a further validation for the RGIFE heuristic, a signature induced biological network (see the section 4.2.4.3 for details) was studied. The aim was to check the relationships among the signature genes in a PPI network context. The network generated using the RGIFE-Union genes resulted in 93 nodes and 190 edges, see Figure 4.10. The network was tested for biological enrichment, that is if the nodes share some biological characteristics, using two different tools: ClueGO [236] and EnrichNet [128].

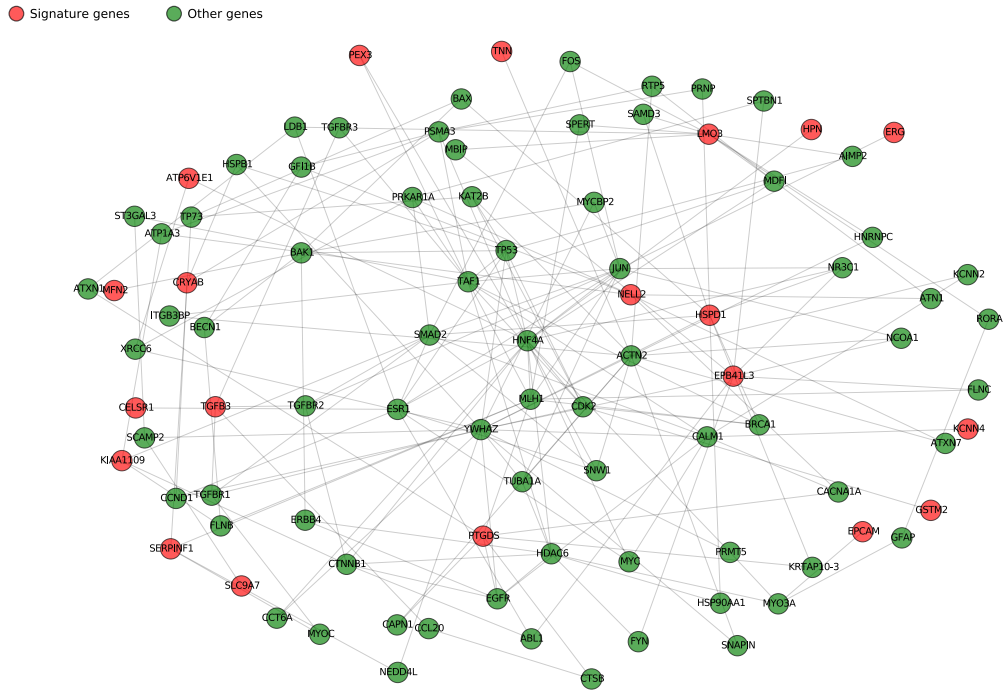


Fig 4.10: The network generated by aggregating all the shortest paths, between the genes of the RGIFE-Union signature, selected from the PPI networks employed in [115].

ClueGO is a Cytoscape plug-in that visualises the non-redundant biological terms for groups of genes in a functionally grouped network. KEGG pathways were used as the biological knowledge base. The result of the enrichment analysis for the nodes of the signature induced network is shown in Figure 4.11, only pathways that are statistically enriched (p-value < 0.05) are reported. The edges between nodes represent the relationship between terms based on their shared genes. The size of the node reflects the enrichment significance of the node, while the colour gradient shows the

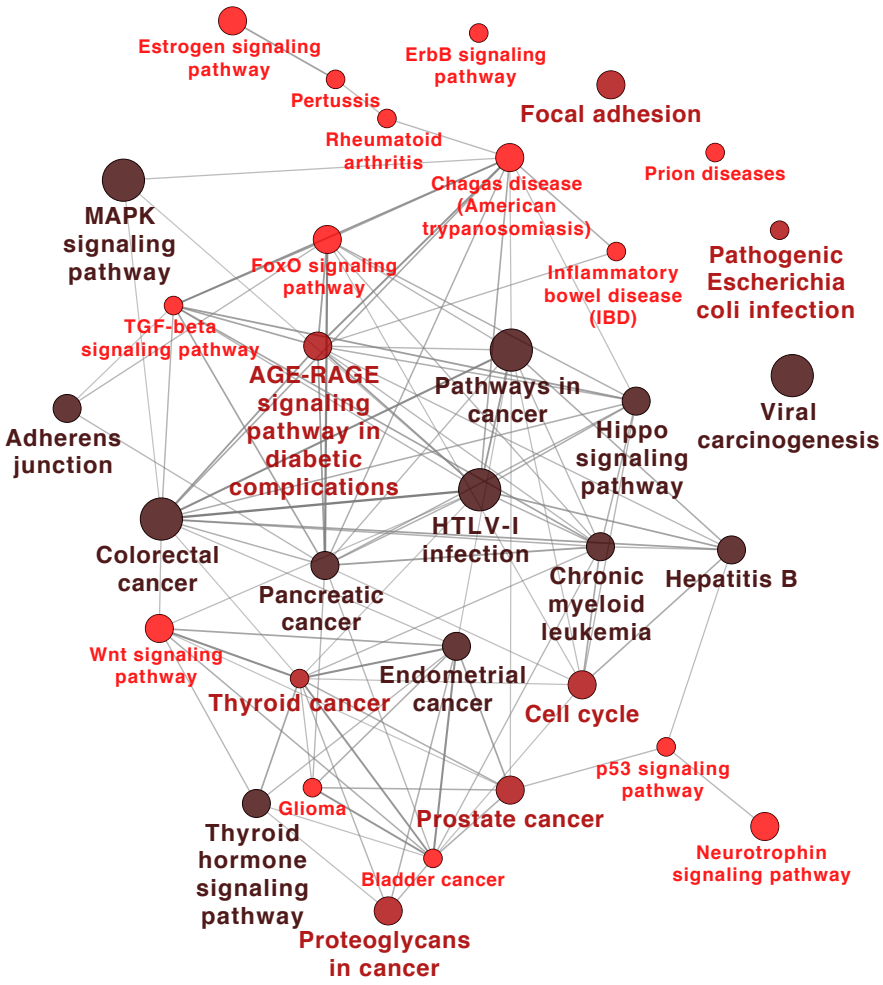


Fig 4.11: Graphical visualisation of the enrichment terms (KEGG pathways found by *ClueGO*) associated to the signature induced network nodes. Edges represent the relationship between terms based on their shared genes. The size of the node indicates the enrichment significance, the colour gradient is proportional to the genes associated with the term. Only terms enriched with p-value < 0.05 are shown.

gene proportion of each cluster associated with the term. One of the highest enriched terms is *pathways in cancer*, this further confirms the role of the selected genes in a cancer context. Among many cancer-related terms, the presence of the *prostate cancer* pathway is particularly relevant as the signature was inferred from prostate cancer data. Finally, *MAPK* is a further pathway already associated, in the literature, with prostate cancer [237].

EnrichNet The signature induced network was also validated with an enrichment tool that uses a different approach than ClueGO. EnrichNet is a gene set analysis

Pathway/process
hsa05210:Colorectal cancer
hsa05216:Thyroid cancer
hsa05213:Endometrial cancer
hsa05212:Pancreatic cancer
hsa05219:Bladder cancer
hsa05220:Chronic myeloid leukemia
hsa04520:Adherens junction
hsa05130:Pathogenic Escherichia coli infection
hsa05020:Prion diseases

Table 4.9: Enriched KEGG pathways (with statistically significant XD-score) identified by *EnrichNet*.

tool that combines network and pathway analysis methods. It maps gene sets onto an interaction network and, using a random walk, scores distances between genes and pathways (taken from a reference database). The XD-score is a network-based association score relative to the average distance to all pathways. The list of pathways with a statistical significant XD-score (using the STRING network as background PPI) is reported in Table 4.9. Several types of cancer are associated with the induced network, among them it appears *colorectal cancer* that, according to *Malacards*, is linked with prostate cancer. Within the list of terms with a significant overlap with the pathways (not reported here) also emerges the *prostate cancer* pathway.

4.4 Discussion

This chapter presented and thoroughly analysed RGIFE: a ranked guided iterative feature elimination heuristic that aims to select small sets of highly predictive features from biomedical datasets. The main differences between RGIFE and other iterative reduction algorithms are: (1) the presence of a back-tracking step that allows to "place back" the features when their removal causes a decrease in the classification performance, (2) the optimal number of selected features is automatically identified rather than being specified up-front and (3) the presence of soft-fails (iterations accepted if the classification performance drops within a tolerance level). To cope with the stochastic nature of RGIFE, three different policies have been proposed (RGIFE-Union, RGIFE-Min and RGIFE-Max) to select the final signature. The evaluation of RGIFE has been

performed using ten cancer-related transcriptomics datasets together with a large set of synthetic data. In addition, RGIFE was contrasted with five other methods commonly used to find new biomarkers.

The presented heuristic is an improvement of the original method proposed by Swan et al. [208]. The heuristic has gone under substantial changes to address the problems of the original version, mainly its large computational time and the high probability of incurring in local optimum solutions. The major changes implemented in this new version include: the use of a different base classifier (from BioHEL to a random forest), a dynamic selection of the number of attributes to remove at each iteration, an extended search and usage of the soft-fails and a more robust validation scheme to be used within the heuristic. The updated RGIFE version outperformed the original method in terms of: prediction accuracy, number of selected attributes and computational time. Better prediction accuracy and a smaller number of selected attributes are likely the results of both the new relative block size criteria and the extended usage of soft-fails. With these new features, the heuristic performs a less greedy attribute filtering, analyses smaller set of data and identifies simpler solutions in an easier way.

The relative block size leads to a less greedy search in the feature space, it requires more iterations to determine the optimal signature, but avoid getting stuck at a local optimum. The improved behaviour can be explained looking at cases when large chunks of attributes are removed. When this happens, it is likely that the reference performance does not improve as a lot of information (attributes) is not considered. Consequently, due to many consecutive unsuccessful iterations, the shrinking of the block size might occur too fast and result in a large final set of predictive attributes, as shown in Figure 4.4. This can be avoided by iteratively reducing the number of attributes to be removed (relative block size).

The ability of RGIFE to identify relevant attributes was tested using many different synthetic datasets. RGIFE-Union clearly outperformed the other five feature selection methods in the analysis of synthetic dataset with different characteristics: noise, nonlinearity, correlation, etc. The heuristic was proven good in selecting relevant features while discarding irrelevant information. The other two policies performed slightly worse but in line with the other methods. When analysing datasets that aim to reflect

the problematics of microarray data, opposite results were obtained. Compared with CFS and L1-based feature selection, the RGIFE policies constantly selected fewer relevant attributes (up/down-regulated genes), while producing a perfect classification, when applied to the *madsim* data [220]. On the other hand, poorer performance were obtained by RGIFE in the analysis of the SD datasets [218]. The bad results of RGIFE are likely to be associated with the low number of samples (25 for each of the three classes) available in the SDs datasets. When dealing with only a few samples, both the predictive performance and the attribute importance, used by RGIFE to evaluate feature subsets, become unstable. Noisy attributes are misinterpreted as relevant and eventually are chosen in the final solution. In addition, the problem of the small number of instances was also amplified by the double cross-validations: external to assess the performance of each method and internal in RGIFE to estimate the goodness of the feature sets. However, quite interesting, the accuracy provided by the RGIFE selected attributes, when determined with a random forest and a Gaussian Naive Bayes classifier, was the best for SD2 and SD3.

When RGIFE was compared with the other approaches, using the real-world datasets, no statistical difference was found in the predictive performance. While the Union and Max policies had similar results, RGIFE-Min clearly had worse accuracy performance. This poor behaviour probably is caused by the extremely small number of selected attributes (up to 15 and 18 times lower if compared with RGIFE-Max and RGIFE-Union respectively). In the context of biomarker discovery, a suitable method needs to minimise the number of proposed attributes while maintaining good predictive performance. RGIFE-Min seems to over-reduce the number of attributes and lead to low predictive performance. On the other hand, when contrasting the number of selected attributes by the RGIFE policies, CFS and the L1-based feature selection, a clear difference emerged. All the three policies selected fewer features than the other two approaches, for RGIFE-Max and RGIFE-Min this difference was statistically significant. RGIFE-Union did not statistically outperform CFS and L1, however, for most of the tested dataset, it selected far fewer features, in particular for datasets with larger numbers of attributes. This is likely a consequence of the RGIFE iterative process that is totally different from the other approaches. The importance within

the classification task is used to identify features not helpful in the discrimination of the samples that could be removed. This leads to a final solution that provides high performance while using fewer attributes. Furthermore, in this context the soft-fail also plays an important role because by accepting a small degrade in performance during the learning phase, more attributes are discarded while a good predictive power is maintained.

Good predictive performance is not enough when proposing a biomarker discovery method, a biological and clinical validation is fundamental. For this purpose, a prostate cancer dataset was used. The specialised literature and the enrichment analysis performed with PANTHER supported the evidence that the genes chosen by the best performing policy (RGIFE-Union) are relevant in prostate cancer. The relevance in a disease context was further confirmed when using the gene-disease associations retrieved from different sources.

With RGIFE, attributes appear in the final model because their (combined) removal causes a drop in predictive performance. Therefore, the biomedical validation suggests that sets of attributes that together are computationally important (because with their values correct predictions can be made) are also key factors from a biomedical point of view. This might be obvious for an univariate approach where a gene, whose regulation is affected by a disease, will be selected because its expression values will greatly differ between controls and cases. On the other hand, when using a multivariate approach, as within RGIFE, such findings were not obvious at the beginning. Overall, the combination of multiple factors, which lead to the correct prediction, seems to reflect the complex role they play in the biomedical condition.

The analysis of the genomic alterations in independent tumour samples (prostate cancer datasets from cBioPortal [195]) showed that RGIFE-Union and SVM-RFE pick genes that are highly altered. A great average alteration in samples not used in the learning process demonstrates that RGIFE selects genes that are not exclusively tied to the training dataset, but that are potentially informative for the disease. CFS, by definition, selects attributes that are highly correlated with the class label but low correlated with each other. The performed analysis suggests that uncorrelated attributes in one dataset tend to not be altered in independent data. This may hint at the ability

of RGIFE in select genes that are less "tied" to the training data and shows a more general importance for the disease. However, further analysis using other datasets (and other types of cancer) need to be performed to confirm this suggestion.

Among the five tested feature selection methods, Chi-Square is the only univariate one, the remaining methods (RGIFE included) are multivariate. RGIFE-Union regularly outperformed Chi-Square in terms of prediction accuracy, disease-relevance and independent data analysis. These suggest that a simple univariate approach is not enough when dealing with high-dimensional data. As already mentioned, biological and biomedical conditions are often the product of multiple factors that interact/cooperate together. Therefore, by looking at single entities (one at the time) complex interactions are missed, is then difficult to obtain good results both in prediction and in biomedical terms and provide reliable biomarkers. Driven by these findings, a multivariate approach is recommended when studying biomedical data that reflect complex conditions.

From a computational point of view, RGIFE is a heuristic for the identification of biomarkers that: (1) performs as good as the compared feature selection methods and (2) consistently selects fewer attributes. From a biomedical perspective, RGIFE extracts genes that: (1) share common biological characteristics and are enriched for pathways important from a clinical point of view, (2) are relevant in a disease context according to both the specialised literature and two different sources of G-D associations and (3) are highly altered in related independent datasets not used for the signature identification process. Altogether, the presented heuristic is a useful and powerful tool for the identification of small predictive sets of biomarkers. RGIFE-Union was found to be the best RGIFE policy leading to good predictive accuracy and relevant biomarkers. However, RGIFE-Min could result useful when lower predictive performance can be tolerated in favour of an extremely small number of biomarkers (more suitable to test with in vitro / in vivo experiments).

In the next section some future ideas, on how to improve the discovery process, will be presented.

4.5 Future work

One of the main advantages of RGIFE is its extreme flexibility in terms of attribute ranking and estimation of the predictive power of the biomarker signatures. In the presented analysis, RGIFE uses a random forest classifier coupled with the gini impurity as metric of the feature importance. However, many other classification algorithms can be employed. In fact, from a purely technical point of view, it is extremely easy to switch from a random forest to a single decision tree or a gradient boosting classifier (or any other classifier implemented in *scikit-learn* that provides an attribute importance score). This requires the modification of a single line of code. Therefore, the performances and the complementarity of different classifiers and feature importance metrics could be tested. Furthermore, given that a regression classifier can be employed as a part of the RGIFE heuristic, it would be possible to apply it to regression problems such as time-series data, common in biomedicine.

In Section 4.3.5, and more specifically in Figure 4.7, RGIFE is shown to be really excellent at identifying small panels of biomarkers. However, the standard deviation of the number of selected features remains large for some datasets. This is particularly evident in the case of the *Leukemia* [94] dataset, where, one of the three runs of RGIFE selected around 500 predictive attributes. This is the result of an early stopping condition reached by RGIFE. It happens when the reference performance cannot be matched for many consecutive iterations and no soft-fails are present. A large number of unsuccessful trials leads to a fast decrease of the block size and in consequence to an early stop. A possible solution to this problem could be a dynamic configuration of the base classier tailored to the dataset being analysed (e.g. the number of attributes). In the performed experiments, the random forest (base classifier) had a fixed configuration (i.e. 3000 trees, unlimited depth of the decision trees, etc.). However, for different dataset sizes, different random forest configurations might perform better. An implementation of RGIFE that automatically adjusts the parameters of its base classifier could solve the problems of the early stop. Nonetheless, this would require a large number of samples, with which a parameter tuning could be performed. Unfortunately, this is a rare circumstance when dealing with biomedical data.

The replacement of BioHEL with a faster base classifier (random forest) have led to a substantial decrease in the computational cost. The improved version of RGIFE requires up to 100 times less time to identify the best biomarkers. Compared to CFS, RGIFE demanded similar execution times for large datasets (such as *Pancreas* [224], *Bladder* [222] or *Breast* [223]), but was far slower for smaller datasets. The detailed results of the complexity analysis are available in the Appendix B.3. The other feature selection methods were instead faster for all the datasets. Therefore, RGIFE could still benefit from a reduction of its overall computational time. At the current stage, the performance of each iteration is assessed via a 10-fold cross-validation repeated N times (default = 10). Given that the training of each fold can be executed separately and, given that the repetitions of the cross-validation are independent, the estimation of the performance on each fold could be done in parallel. In addition, an even faster classifier could be employed in place of the random forest, for example, decision trees. However, this solution needs to be thoroughly tested to guarantee that similar performance can be reached.

Although three different policies were proposed to identify the best predictive signature, the stochastic nature of RGIFE might still lead to different optimal solutions, especially if the analysed data contain a large number of features or are particularly noisy. Various reasons can lead to noisy data: errors introduced by the technology, wrong assignment of the class labels, intrinsic noise of the studied process/condition, etc. In general, the stochastic behaviour of the heuristic not necessarily represents a drawback. In fact, in the presence of multiple optimal solutions [211], RGIFE has higher chances to identify equally well performing diversified sets of biomarkers. Nonetheless, in order to reduce the variability of the inferred signatures, a fixed attributes ranking could be used across the iterations. That is, the attributes to be removed in each iteration could be selected according to the importance ranking from the first iteration (calculated using the original dataset). This procedure has been suggested in [238], where it is stated that a re-calculation of the feature importance, as currently done for RGIFE, is much greedier and provides worse performance.

Finally, as mentioned in Chapter 2, an emerging research path involves the integration of prior biological knowledge in the model inference process. In the case of RGIFE, it

could mean the integration of biological knowledge in the reduction process. That is, the selection of features to be removed could be (partially) based on external knowledge. Practically, there could be a version of RGIFE where the attributes associated with the disease that characterises the data (extracted from the specialised literature or portals such as Malacards [139]), become hard to discard. Another option would be to adjust the rank of the attributes to take in account their role in a molecular context (e.g. node degree within a PPI network). Ultimately, the block of attributes to be removed could be selected based on the shared biological characteristic, that is looking at the number of common GO terms or biological pathways in which they are involved.

Summary

This chapter has presented RGIFE, a machine learning based heuristic for the discovery of biomarkers. RGIFE has been extensively tested and evaluated using cancer-related transcriptomics data. In the next chapter, RGIFE will be used to analyse datasets that integrate information from multiple biomedical sources. The analysis will focus on evaluating RGIFE with different types of data, in particular having: (a) missing values, (b) categorical attributes, (c) information from heterogeneous sources and (d) imbalanced distribution of the samples (i.e. negative instances substantially outnumbering the positive instances).

5

Identification of biomarkers for knee osteoarthritis

Contents

Abstract

This chapter describes the application of machine learning techniques, mainly RGIFE, for the identification of biomarkers for knee osteoarthritis in overweight women. The RGIFE heuristic shows its flexibility when applied to biomedical data affected by missing values and imbalanced class distributions. Pursuing one of the aims of the thesis, both RGIFE and FuNeL are shown to be effective in extracting relevant knowledge from biomedical data of different forms (not limited to transcriptomics data as illustrated in Chapter 3 and 4). In addition, the inferred predictive models are extensively exploited and analysed to better contextualise the role and the importance of the proposed biomarkers.

5.1 Introduction

Osteoarthritis (OA) is a progressive disorder of the joints that features a gradual loss of cartilage, low-grade synovial inflammation, and the development of cysts and bony spurs at the borders of the joints. In a 2010 study [239], hip and knee OA was ranked as the 11th highest contributor to global disability and its prevalence is expected to increase due to the rise of the worldwide obesity along with the ageing of the population. Despite the high prevalence of OA, currently there is no cure for this disease [240], the available treatments only diminish symptoms such as pain and disability. Nowadays, knee OA is mainly diagnosed using clinical and radiographic changes generated by structural damages that occur late in the disease progression. Unfortunately, these techniques have a relatively large precision error and low sensitivity [241]. Given the limitations of these imaging-based biomarkers (also known as "dry"), there is an increased need for identifying new and sensitive biochemical biomarkers (also called "wet"), other dry biomarkers (such as coming from MRI), or a combination of both that can early detect OA before structural damages and established clinical OA develop.

D-BOARD is a European partnership, funded with the EU 7th framework programme of research, to bring together leading academic institutions and European Small and Medium Enterprises (SMEs) with the goal of finding reliable biomarkers and diagnostic

tests that can facilitate earlier diagnosis of OA, and inform the prognosis, monitoring and therapeutic strategies for chronic and disabling forms of this disease. A part of the work for this PhD project involved the collaboration with the Work Package 5 - WP5 (bioinformatics and data analysis) of the D-BOARD consortium. The goal of WP5 is the application of machine learning and bioinformatics techniques to identify new relevant knee-OA biomarkers from the analysis of multiple and different biomedical data. This chapter presents the analysis performed during the collaboration with the D-BOARD consortium trying to tackle the problem of biomarkers discovery in a knee OA context.

Recently, several different approaches have been presented trying to solve the lack of early knee OA biomarkers. For example, the levels of serum COMP (Cartilage Oligomeric Matrix Protein) have been correlated with the development the condition [242], the incidence of clinical knee OA among middle-aged overweight and obese women has been linked with the baseline fibulin-3 concentrations [243] while it was shown to be negatively associated with the concentration of COLL2-1NO2 (a peptide that represents the combination of collagen type II degradation products (Coll2-1) and reactive nitrogen and oxygen species (RNOS), NO and O2âĹŠ and can be measured in urine or serum) at baseline [244]. Finally, adipokines were suggested as predictive biomarkers for early onset post-traumatic knee OA [245]. Lately, together with traditional clinical, biological and chemical approaches, machine learning and computational methods started to make an impact in this area [246]. GLMNET (Generalized Linear Models with elastic NET), a machine learning classifier [247], showed that the combination of plasma citrullinated protein, anti-cyclic citrullinated peptide antibody and 4-hydroxyproline provides specific and sensitive detection and discrimination between early-stage OA and rheumatoid arthritis and between non-rheumatoid arthritis and good skeletal health. Ashinsky et al. [248] evaluated the ability of machine learning to discriminate between MRIs of normal and pathological human articular cartilage. The study employed a multiple linear least-squares regression that successfully predicted OARSI (Osteoarthritis Research Society International) scores and classified plugs with high accuracy (86%). Serum biomarkers have been employed to

train an artificial neural network that discriminates patients affected by osteoarthritis and rheumatoid arthritis 100% correctly [249].

In 2009, an important research in the field of knee OA, the PROOF (PRevention of knee Osteoarthritis in Overweight Females, ISRCTN 42823086) study, was presented [250]. The objective of the PROOF study was the evaluation of the effect of a tailored diet-and-exercise program, focused on reducing weight, and of oral crystalline glucosamine sulphate on the incidence of knee osteoarthritis in a high-risk group of overweight women between 50 and 60 years of age, free of clinical knee osteoarthritis at baseline. The data collected for the PROOF study have been extensively analysed by the D-BOARD consortium, in particular, the WP5 used them via the application of machine learning techniques, to discover new biomarkers for the early detection of knee OA.

This section of the dissertation is focused on the analysis of the PROOF data study (collected after 2.5 and 6.5 years from the baseline) using machine learning-based methods. In particular, it is presented a pipeline for the identification, evaluation and validation of biomarkers, at which core is placed RGIFE. The proposed approach identified small highly predictive models defined by a handful of variables whose relevance was extensively assessed. In addition, the relationship (positive or negative) of each biomarker with the knee OA incidence was studied. Finally, for a subset of the available data, a functional network was generated using the FuNeL protocol presented in Chapter 3. Overall, the analysis revealed the importance of the use of imaging-based information for the prediction of the disease as well as the dietary information of each individual. Furthermore, the results confirmed the influence of some known biochemical markers. When compared with the state-of-the-art, the proposed models showed better performance in predicting the incidence of OA. The knowledge associated with FuNeL networks demonstrated a high overlap with the information provided by RGIFE, confirming the robustness of the proposed methodologies when applied to the same data.

5.2 Material and methods

This section describes the characteristics of the data generated from the PROOF study. Then, each step of the pipeline employed for the identification and evaluation of the biomarkers is presented.

5.2.1 Datasets and individuals

The data used for the presented analysis comes from the PROOF study: a preventive randomised controlled trial including 407 middle-aged women with a BMI $\geq$ 27 kg/m2 free of clinical knee OA at baseline [250]. After 30 months, the preventive effects of a diet and exercise program and oral glucosamine sulphate were evaluated. Five different outcome measures were used to define the presence of knee OA:

- incidence of "combined radiographic and clinical ACR-criteria"
- presence of frequent knee pain:
- lateral JSN (Joint Space Narrowing) of $\geq$ 1.0 mm
- medial JSN (Joint Space Narrowing) of $\geq$ 1.0 mm
- incidence of K&L $\geq$ 2

The American College of Rheumatology (ACR) criteria is a set of rules, defined to determine the presence of knee OA [251]. For example, using history, physical examination and radiographic findings, the condition is present if there is pain in the knee and one of the following statement is true: over 50 years of age, less than 30 minutes of morning stiffness, crepitus on active motion and osteophytes. The chronic knee pain is self-explanatory, while the lateral and the medial Joint Space Narrowing (JSN) measures the loss of lateral (or medial) femorotibial cartilage. The presence of OA is confirmed if, during the analysed time-frame, the space in between joints has decreased more than a certain threshold (1.0 mm in the PROOF study). Finally, the Kellgren & Lawrence (KL) system is a method for classifying the severity of knee osteoarthritis

	Mean ± SD or percentage
Age (yr.)	55.7 ± 3.2
BMI (kg/m^2)	32.3 ± 4.3
Menopausal status	69%
Western ethnicity	96%
Mild symptoms	45%
Physical activity (SQUASH score)	6912 ± 3704
K&L=1	60%
K&L=2	10%

Table 5.1: Baseline characteristics of the included subjects (N = 365).

(OA) using five grades/scores that combines multiple criteria such as: JSN, presence of osteophytes, bony deformity, etc. [252].

Each individual, free of knee OA at the baseline (e.g. K&L grade = 0), was analysed, for the existence of the knee OA after 2.5 years. The baseline characteristics of the subject included in the PROOF study are shown in Table 5.1 The variables measurements were taken at baseline while the presence of the condition was assessed later on. Therefore, the 2.5 years data are considered the result of an *incidence* study and an individual was labelled as *incident* if, after 2.5 years, one of the five outcome measures was present (e.g. K&L grade $\geq$ 2). Five different classification problems were defined, one for each outcome measure. Each data sample was characterised by the value of 186 heterogeneous variables (a list of the variables is available in Appendix C). When some of those variables were used to define the presence of knee OA, they were removed from the analysis (e.g. ACR value at baseline when using the ACR criteria to determine the presence of the condition), see Table 5.2 for the specific number of variables. The information were derived from baseline questionnaires (including demographics, menopausal status, knee complaints, physical activity level, quality of live, habitual nutritional intake, and KOOS (Knee injury and Osteoarthritis Outcome Score) questionnaire), radiographs (for obtaining baseline K&L grade, medial alignment angle, and knee joint shape using active shape modelling), MR images (scored with semi-quantitative MOAKS system [253] and used to define MRI OA [254]), physical examination (including pain upon palpation of knee structures, crepitations, presence of Heberden's nodes, blood pressure, knee laxity and range of

motion, warmth of the knee joint, waist circumference, and skinfolds for fat percentage calculation), and biochemical markers from serum and urine such as the fibulin-3 epitopes (fibulin3-1, fibulin3-2 and fibulin3-3) [243], COLL2-1NO2 [244], and C1M and C2M (collagen type I and II degraded by matrix metalloproteinase) [255]. A detailed description of the acquisition of non-biochemical variables is given in [250].

OA definition	Attributes	Incident/non-incident
ACR criteria	185	39 / 315
Knee pain	186	51 / 300
Lateral JSN	186	41 / 311
Medial JSN	186	38 / 314
KL incidence	184	27 / 294

Table 5.2: Summary of the information for the datasets generated from different knee OA outcome measures defined after 2.5 years from the baseline (beginning of the PROOF study).

For a subset of the PROOF study individuals (74 samples), a lipidomics screening was performed after 6.5 years from the baseline time point. The lipidomics measurements were performed by two different partners of the D-BOARD consortium, namely TNO (Netherlands Organisation for Applied Scientific Research) and UNOTT (University of Nottingham). The platform employed by TNO screened a total of 294 lipids, while UNOTT generated data containing information about 32 lipids. The presence of OA was defined, for the 74 individuals, after 6.5 years from the baseline and using three of the previously described outcome measures: combined radiographic and clinical ACR-criteria, chronic knee pain and K&L ≥ 2. Different than the 2.5 years data, in which the measurements of the variables were taken at the beginning of the study, the lipids abundance was assessed at the same time point in which the presence of OA was determined. This results in a *cross-sectional* study that can be used to extract biomarkers that can predict the presence of OA rather than its incidence (e.g. development).

The distribution of samples in *OA* and *non-OA* is reported in Table 5.3. Unlike the data generated after 2.5 years, the lipidomics dataset did not contain missing values, nor categorical data were present (such as questionnaire answers). Three different classification problems, as for the 2.5 years data, were defined using the listed OA outcome measures.

OA definition	OA/non-OA
ACR criteria	15 / 59
Knee pain	9 / 65
KL incidence	17 / 57

Table 5.3: Distribution of the individuals (class labels) in the datasets generated from different knee OA outcome measures defined after 6.5 years from the baseline (beginning of the PROOF study).

The presented data were employed to identify driving factors (biomarkers) for knee osteoarthritis in overweight women with the help of machine learning methodologies. Separate analyses were performed using the data associated with the two different time points. From now onward, for simplicity, the five knee-OA measures, and the relative analysis, will be referred as: ACR criteria, knee pain, lateral JSN, medial JSN and K&L score.

5.2.2 Extension of RGIFE to analyse the PROOF study data

The aim of WP5 of the D-BOARD consortium is the application of machine learning, and more in general, bioinformatics techniques for the discovery of new knee OA biomarkers. To solve this research problem, the PROOF data study was analysed with RGIFE, the heuristic presented in Chapter 4. RGIFE has been shown to be good at identifying small numbers of predictive relevant features from biomedical data. From a machine learning point of view, the data resulting from the PROOF study represents a difficult task due to (a) the imbalance distribution of the samples (much more *case* than *controls*) and (b) the presence of missing values.

When the number of instances of one class far exceeds the other, problems can arise because the machine learning algorithm might be tempted to treat the minority examples as outliers of the majority class and therefore ignore or mistake them. Several approaches have been proposed to solve the imbalance problem; commonly the aim is to re-balance the distribution of the samples between the classes. By having an equal class distribution, the machine learning algorithms tend to learn better from the data, thus generate more accurate models. The undersampling methods remove samples from the majority class until it contains as many data point as the minority

class. However, by using this approach, important samples and relevant information might be lost. This is particularly problematic when dealing with biomedical data that already contains a limited number of observations. Conversely, the oversampling techniques try to re-sample the minority class until it consists of as many samples as the majority class. This approach was shown, over the years, to be more robust and suitable.

As common in clinical studies, the PROOF data available from the 2.5 years time point were affected by the presence of missing (baseline) values. Missing values appear for different reasons: patients that did not fully completed a form, faulty reading from lab machines, human errors in the collection or recording of the data, etc. Missing values might generate problems when working with machine learning algorithms. In some case, the algorithms are unable to handle data affected by missing values, in others they can result in variance underestimation, distribution distortion, and correlation depression [256]. The overall approach to solve the missing data problem is to impute them by looking at the closest and most plausible values available in the data.

To apply RGIFE to the PROOF data, a new version of the heuristic was implemented. With such modifications RGIFE can deal with data affected by missing values and tackle the problem of imbalanced distribution of the samples. The new implementation includes several algorithms for the application of both imbalance learning (SPIDER, SMOTE, B-SMOTE, etc.) and missing values imputation (mean, K-means, KNN, etc), all of them are executed using the libraries available from the KEEL tool suite [257]. Both techniques are put in place before RGIFE generates the model that evaluate the predictive performance of a feature set. Furthermore, in the updated RGIFE implementation is possible to set a cost-sensitive learning strategy to be used by the classifier. The cost-sensitive learning assumes a higher cost (weight) when the classification errors involve a specific class. By having a large cost when misclassifying the minority samples, the learning algorithm can put more effort in classifying the under-represented samples correctly, thus try to overcome the imbalance distribution issue.

When analysing the 2.5 years data, initial tests were performed trying different methods to tackle both the imbalance class distribution and the imputation of missing val-

ues. The results (not reported here) showed that the best performance was achieved when applying an oversampling approach with SPIDER [258] and imputing the missing values with the K-Means algorithm [259]. SPIDER is an oversampling algorithm that re-balance the class distribution in two steps. In the first step, the samples misclassified by a KNN classifier are labelled as noisy. Then, SPIDER strongly amplifies the minority class instances and it removes the noisy examples from the majority. Conversely, the K-means imputation algorithm is based on the K-means clustering and it can be divided into three phases [259]. First, K complete samples (without missing values) are randomly selected as K centroids. Then, iteratively the partitions are modified trying to reduce the distance of each sample to its centroid. Finally, the missing values are imputed based on the cluster information. The samples that are a member of the same cluster are taken as nearest neighbours of each other, and the nearest neighbour algorithm is used to replace missing data. Both algorithms were used setting their parameters to the default values provided by KEEL.

5.2.3 Discovery and evaluation of small sets of biomarkers

When trying to identify biomarkers from biomedical data, the method employed for their discovery represents the biggest part of the whole process. However, other steps are necessary to validate and interpret the set of proposed factors. Therefore, a pipeline, defined by different methodologies, was put in place to extract, validate and interpret biomarkers. RGIFE is at the core of this pipeline (illustrated in Figure 5.1), it provides small sets of variables that can be potentially used as biomarkers. First, it is necessary to assess the predictive performance of the proposed biomarkers, this is typically done using a cross-fold validation. Then, from the complete set of data, the candidate biomarkers are extracted. Those two operations will be described in Section 5.2.3.1, while in Section 5.2.3.2 it will be discussed how the predictive performance of the proposed biomarkers are statistically assessed. Once the list of biomarkers is defined, a thorough analysis needs to be performed so that their relevance and association with the studied condition, knee OA in this instance, can be appraised. Within the same model, different biomarkers have different roles and importance, Section 5.2.3.3 will present how this can be assessed. In Section 5.2.3.4, it will be described

a methodology to estimate the association (positive or negative) of each variable with the dependence variable, knee OA measure in here. Using literature mining, the proposed prediction models can be compared with the state-of-the-art approaches and check whether they provide better solutions. Finally, by applying the FuNeL protocol, it is interesting to validate the role of the biomarkers in functional networks and evaluate the relationships between them. The presented pipeline is generic enough to be applied to a wide variety of biomedical data, furthermore, although RGIFE has been used in this instance, other biomarkers discovery methods can perfectly fit in it.

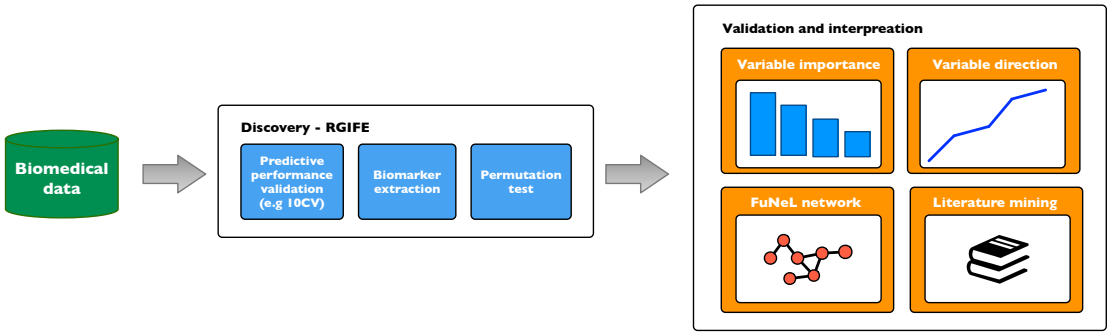


Fig 5.1: The proposed pipeline for the identification, validation and interpretation of biomarkers.

5.2.3.1 Generation and selection of reduced predictive models using RGIFE

RGIFE is a flexible and fine tuneable heuristic whose behaviour can be adjusted based on the type of data being analysed. 30 different configurations were analysed to perform a full search in the space of all the optimal solutions (set of biomarkers) for each OA definition. Having fixed the oversampling and the missing values imputation algorithms (when dealing with the 2.5 years data), the configurations differed in terms of maximum depth of the decision trees within the forest (the number of trees was set to 3000 as in the analysis performed in Chapter 4) and misclassification costs (penalisation when misclassifying *incidence* or *OA* samples during the learning phase).

For each dataset, the best performing configuration was identified using a standard $M \times n$-fold cross-validation. When analysing the 2.5 years data, n was set equal to 10, and the whole cross-validation process was repeated 10 times, that is $M = 10$. The repetitions of the validation scheme were performed to minimise the possible

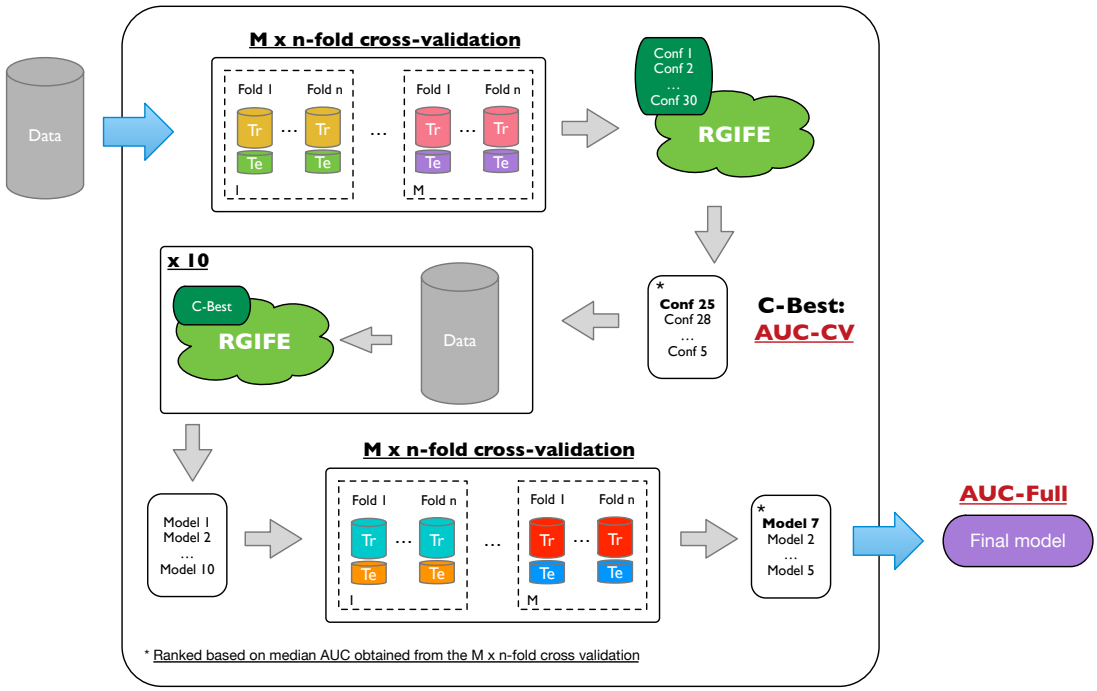


Fig 5.2: The overall pipeline employed for the identification of the best predictive models. When analysing the 2.5 years data $M = 10$ and $n = 10$, while for the analysis of the 6.5 years data $M = 1$ and $n = 74$ (leave-one-out).

bias introduced by the data being split into training and test set when having such a limited number of minority samples (e.g. *incidence* class). On the other hand, when analysing the 6.5 years data (lipidomics information), a leave-one-out validation scheme was preferred (i.e. $n = 74$ and $M = 1$). The use of LOOCV was necessary to obtain a robust validation when using a limited number of samples and a reduced number of OA cases (e.g. 9 using the knee pain OA definition). By ranking the different configurations based on their predictive performance (calculated using the AUC: Area Under the ROC Curve), the best set of parameters, tailored to each different dataset, were selected. Finally, using the just mentioned best configuration, RGIFE was applied to the whole set of samples, to identify the smallest most predictive set of biomarkers. A summary of the overall analytic pipeline employed for the discovery of the best models is illustrated in Figure 5.2.

In the process of biomarkers discovery, the key is to find the correct balance between the size of the inferred model (number of variables) and its performance. Often, the smaller the size of the model, the lower are its predictive performance. For this study, the correct trade-off between size and predictive power is guaranteed by selecting and

analysing the best performing models (higher AUC) containing at most 10 variables. The maximum size of the models, 10 variables, was suggested by the clinician collaborators. Based on the analytic pipeline employed for the identification of the best biomarkers (see Figure 5.2), in the results section, two different AUC values, namely AUC-CV and AUC-Full, are provided. The first one refers to the AUC obtained by the best performing configuration with the first cross-validation. That is, RGIFE is applied to the training sets and the AUC is calculated when predicting the class of the test samples and using only the attributes selected from the training data. On the other hand, to calculate AUC-Full, first RGIFE is applied to the whole dataset, then the selected attributes are kept on the training and test sets generated from a cross-validation scheme (different than the one employed for AUC-CV, when analysing the 2.5 years data). Finally, AUC-Full is calculated by predicting the labels of the test samples across the different folds. AUC-CV gives us an idea about the predictive performance of models when analysing totally new and unseen data, while AUC-Full indicates how well the selected biomarkers discriminate between the two classes of individuals in the PROOF study data. In addition, AUC-CV is a good estimator to detect a possible overfitting occurred during the learning training. If the AUC-CV values are poor (0.5 is the threshold indicating performance worse than a random classifier) while obtaining good AUC-Full values, it is plausible that the selected biomarkers can detect the condition in the analysed data, but are likely to perform poorly if tested on completely unseen samples.

5.2.3.2 Permutation tests

A statistical test, based on permutations, was used to assess whether the performance of each model was statistically significant. The approach is similar to one described in [162]. For each dataset, 100 permutated copies, where the labels (*OA* and *non-OA*) were randomly assigned to each individual, were generated. Then, for each of the permutated dataset, the AUC obtained using the biomarkers selected by RGIFE was calculated. By counting the number of times in which a model provides equal or higher performance, when trained using random data rather than the original data, the statistical significance (p-value) of its predictive power (AUC in this instance) was

calculated. In here, the p-value represents the probability to obtain the original AUC by chance. In addition, the likelihood to draw the AUC associated with the models by chance was estimated with a one-tailed permutation test using the distribution of the permutated AUC values.

5.2.3.3 Variable importance

After having selected the best predictive models, one of the aims was the analysis of the role and the relevance of each component (variable). Given a model, the least "important" variable is the one that, when removed, causes the smallest drop in performance. The additive value was obtained from the generation of decremental sub-models. S tarting from the original set of variables, the less contributing one (causing the largest drop in AUC) was iteratively removed until reaching a single-variable model. More specifically, from a model with m variables, all the m sub-models defined by $(m-1)$ variables were tested to select the one providing the highest AUC (while $m \geq 1$). The AUC was calculated, as in the previous analysis, performing an $M \times n$-fold cross-validation. Overall, this process ranks the variables in an increasing importance order, the later a variable is removed the higher is its contribution.

5.2.3.4 Variable direction

From a clinical point of view, when analysing the variables that define a predictive model, it is fundamental to assess their direction, that is how the value of a variable can influence the presence of the disease (condition). For example, it is useful to know how changes in the values of BMI affect the chance to develop knee OA. The variable direction aims to determine the relations of each variable x with the outcome measure y (presence of knee OA). That is, how the response y, or its expectation, varies with the values x_j assumed by the biomarkers of the predictive models. This problem can be solved by calculating the partial dependence of each variable. Partial dependence is a method to visualise the partial relationship between the outcome and the predictors [260]. The idea is to check the effect of changing the value of a variable x to y while holding the remaining variables constant. This was obtained by generating a copy of the data for each possible value x_j while leaving the other variables as in the original

data. Then, for each of the data copy, a prediction using only the previously selected biomarkers was performed. The process generates a range of predictions, for each value x_j, that when plotted, describes the relationship between x and y. Each point of the plot represents the average probability of the samples (calculated across an $M \times n$-fold cross-validation) to belong to the positive class (presence of OA) when fixing the value of the analysed variable.

5.2.3.5 Inference of functional networks with FuNeL

The primary goal of the D-BOARD project is the identification of novel biomarkers for the prediction of knee OA. However, the analysis was not limited to the inference of small panels of biomarkers using RGIFE but was extended to the inference of functional networks. With the generation of networks that include the relationships among the variables of the PROOF study data, a better understanding of the knee OA condition can be provided to the experts of the field. As already mentioned, many diseases and conditions are the results of complex mechanism and chains of interactions, this can be captured by the networks generated using the FuNeL protocol presented in Chapter 3. For this part of the analysis, only the 6.5 years lipidomics data were considered. Two main reasons are behind this choice. First, the lipidomics data are not affected by missing values, therefore the application of FuNeL becomes simpler. Although BioHEL can deal with datasets with missing values, its imputation is quite trivial (mean values calculated from the training set), therefore it would have been tricky to understand the impact of missing values in the final networks. In addition, given the heterogenic nature of the attributes that characterise the 2.5 years (different types of information such as food habits and protein abundance), the biological interpretation of those functional networks would have been more difficult.

Given the small number of attributes of the lipidomics data, the C_2 configuration of the FuNeL protocol was used: no feature selection and one stage of network generation. Preliminary tests showed that, probably due to the limited number of samples, the networks generated from different executions of FuNeL were quite different regarding edges and top hubs. Therefore, to increase the stability of the inferred networks, it was decided to slightly modify the way in which FuNeL creates the final network.

Rather than applying a statistical permutation test based on the node score (number of times an attribute appear in the BioHEL's rules), the edge score was used (number of occasions two attributes are expressed together in the same rule). The same approach was also tested when analysing FuNeL with transcriptomics data, however, in that instance, the edge score did not contain enough signal. That is, due to the large number of total attributes in the transcriptomics data (at least 7000), very few edges resulted having a significant score (p-value < 0.05). Conversely, when applied to lipidomics data, whose observations are defined in a much lower dimensional space (at most 294 attributes), the signal associated with the edge score was stronger and led to a generation of more robust functional networks.

5.3 Results

The updated version of RGIFE (able to deal with missing values and imbalance distribution of the classes) was employed to identify small sets of predictive biomarkers from the PROOF study data collected at two different time points: 2.5 and 6.5 years. The following section will present the models generated from this analysis. For the sake of readability, the results from data collected at different time points will be separated. Furthermore, the lipidomics results will be divided according to the source (partners) that generated the data.

5.3.1 2.5 years predictive models

Out of all the subjects included in the PROOF study, 365 had a 2.5 years follow-up data and were selected for the present study. A different total number of samples was available for separate knee OA outcome measures. The ACR criteria and chronic knee pain after 30 months occurred in 39 out 354 (11%) and in 51 out of 351 (15%) women respectively. The incidence of lateral JSN ≥ 1.0 mm was assessed in 41 (12%) out of 352 women, while medial JSN was seen in 38 (11%) women out of 352. Finally, the incidence of K&L ≥ 2 was measured in 27 (8%) out of 321 individuals.

When dealing with the 2.5 years data, RGIFE used the K-means imputation algorithm to process the missing values, while SPIDER was employed, during the learning phase,

OA measure	Variables	Cat.	AUC-Full	AUC-CV	p-value
ACR criteria	KL grade $\geq$ 1 in one or both knees	OA	0.788	0.692	< 1e-04
	Maximal isometric quadriceps strength	CI			
	Mode 10 (Active Shape Modelling)	IM			
	Mode 15 (Active Shape Modelling)	IM			
	Mode 11 (Active Shape Modelling)	IM			
	Presence of knee pain in the last month	PQ			
	Difficulties when kneeling	PQ			
	C2M concentration	BM			
Knee pain	KL grade $\geq$ 1 in one or both knees	OA	0.755	0.637	< 1e-04
	KL grade $\geq$ 2 in one or both knees	OA			
	WOMAC function score	OA			
	Maximal isometric quadriceps strength	CI			
	Mode 11 (Active Shape Modelling)	IM			
	Difficulties when jumping	PQ			
	Frequency of biscuits / week	FQ			
Lateral JSN	Fat percentage	CI	0.737	0.549	< 1e-04
	Mode 11 (Active Shape Modelling)	IM			
	Mode 10 (Active Shape Modelling)	IM			
	Frequency fruits / week	FQ			
	Concentration of Coll2-1NO2 adj. for creatinine	BM			
Medial JSN	Quality of life	CI	0.731	0.539	< 1e-04
	Nr. years since menopause	CI			
	Waist circumference	CI			
	Mode 15 (Active Shape Modelling)	IM			
	Frequency bananas / week	FQ			
	C1M concentration	BM			
KL incidence	BMI	CI	0.823	0.699	< 1e-04
	HbA1c concentration	CI			
	Presence of OA on MRI	IM			
	Grinding / clicking sound when moving the knee	PQ			
	Frequency of apples and pears / week	FQ			

Table 5.4: Summary of the models inferred for each knee OA outcome measure. Variables are baseline measures divided according to the type of information provided: OA measures (OA), clinical information (CI), imaging-based information (IM), biochemical marker (BM), pain questionnaire (PQ) and food questionnaire (FQ) answers. The AUC column contains two values: AUC-Full and AUC-CV (in brackets). The last column indicates the permutation test p-value (one tailed).

to address the imbalance distribution of the classes. In Table 5.4 are reported the models generated (using the pipeline illustrated in Figure 5.2) from the data collected after 2.5 years from the start of the PROOF study. Different OA outcome measures were used to define the incidence of knee OA. The variables are grouped based on their source of information: OA measures (OA), clinical information (CI), imaging-based data (IM), biochemical markers (BM), pain (PQ) and food questionnaire (FQ). All the values assumed by those variables are coming from the baseline assessments.

Figure 5.3 shows the ROC curves (AUC-Full values) generated from each of the inferred models. The best predictive biomarkers were inferred using the K&L score incidence

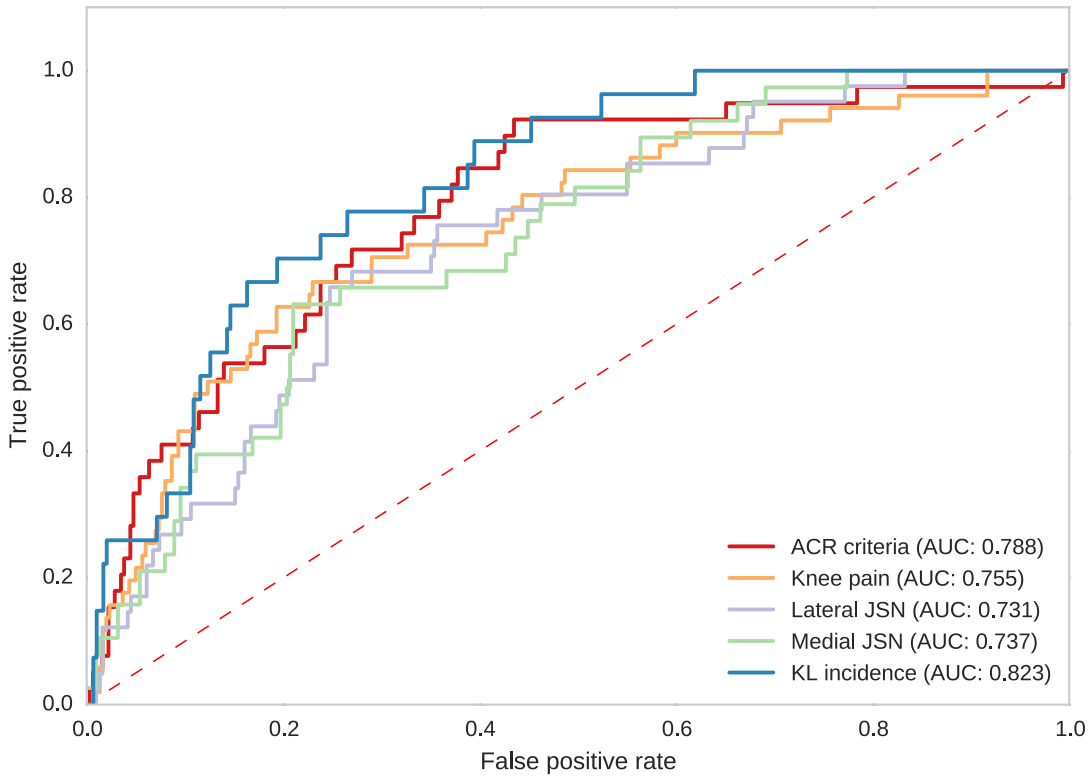


Fig 5.3: The ROC curves generated by the best performing models (2.5 years). The AUC values refer to the AUC-Full.

OA outcome measure, they lead to an AUC-Full of 0.823. It is interesting how the best performing model is also the smallest inferred, with only 5 variables. The second best model is associated with the ACR criteria and provides an AUC-Full of 0.788. However, it also contains the largest number of variables (8 in total). Finally, the JSN outcome measures lead to the two lowest performance, respectively 0.731 for the lateral and 0.737 for the medial compartments. Overall, all the identified models obtained AUC-CV values larger (calculated with a 10 x 10-fold cross validation) than what is obtained with a random classifier, suggesting that some valuable pattern were identified by RGIFE within the data. The rank of the models is equivalent when considering either AUC-CV and AUC-Full. From a pure machine learning point of view, it interesting to notice how the top 3 best models were generated using a random forest classifier with a limited depth (2 for ACR criteria and knee pain, 4 for the K&L score incidence). Conversely, the worse models were identified when RGIFE was employing *deeper* decision trees (depth = 8 for both the JSN data). This might suggest that when RGIFE is dealing with challenging data and is not able to detect any simple

pattern (small decision trees within the forest), the best results are achieved with large trees that are more specific and tight to the data. However, although providing good AUC-Full values, this might lead to an overfitting problem given the modest AUC-CV values.

Table 5.4 shows the heterogeneity of the models in terms of variable types. Every model is defined by at least four different categories of variables. The imaging-based variables are important for the ACR criteria model as well for JSN lateral. On the other hand, to obtain good predictions using the knee pain criteria, OA measures assume a relevant role. Food related questionnaire data appear in almost every predictive model (ACR criteria model is the exception). Interestingly, most of them are related to the fruit intake per week, while in the knee pain model the data are associated with the number of biscuits (sugar) consumed per week. Finally, it is important to notice the presence of biochemical markers already associated, from the literature, with the incidence of knee OA such as C1M [255] and C2M [261] and the concentration of Coll2-1NO2 [244].

To statistically assess the goodness of the AUC associated to each of the inferred models, a statistical test was used (see Section 5.2.3.2 for details). For all the presented models, the permutation tests showed that their performance was statistically significant. None of the models obtained an AUC higher than the values provided in Table 5.4 when applied to the 100 permutated datasets (empirical p-value = 0). The statistical significance was also confirmed by the one-tailed permutation test (p-value $<$ 1e-04 for all the tested models).

5.3.1.1 Additive values of the biomarkers

The additive value of each variable, within the predictive models, was assessed performing a decremental analysis. The later a variable disappears from the models, the higher is its contribution when solving the classification task. Figures 5.4 – 5.8 illustrate the submodels generated using the five different OA outcome measures. The y-axis of the figures represents the AUC values, the result of a 10-fold cross-validation repeated 10 times, of each submodel. The variables defining the submodels are described over each bar, in white is highlighted the variable that is removed in the following decremental

step. The first bar represents the original model, while the last one shows the single best performing variable.

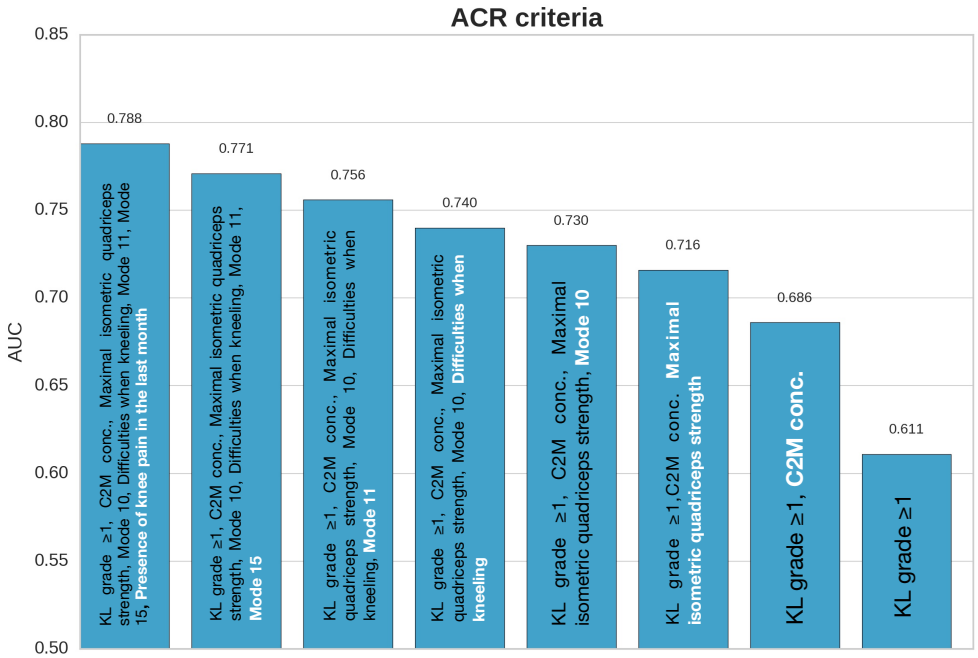


Fig 5.4: Decremental models generated using the ACR criteria for the knee OA incidence.

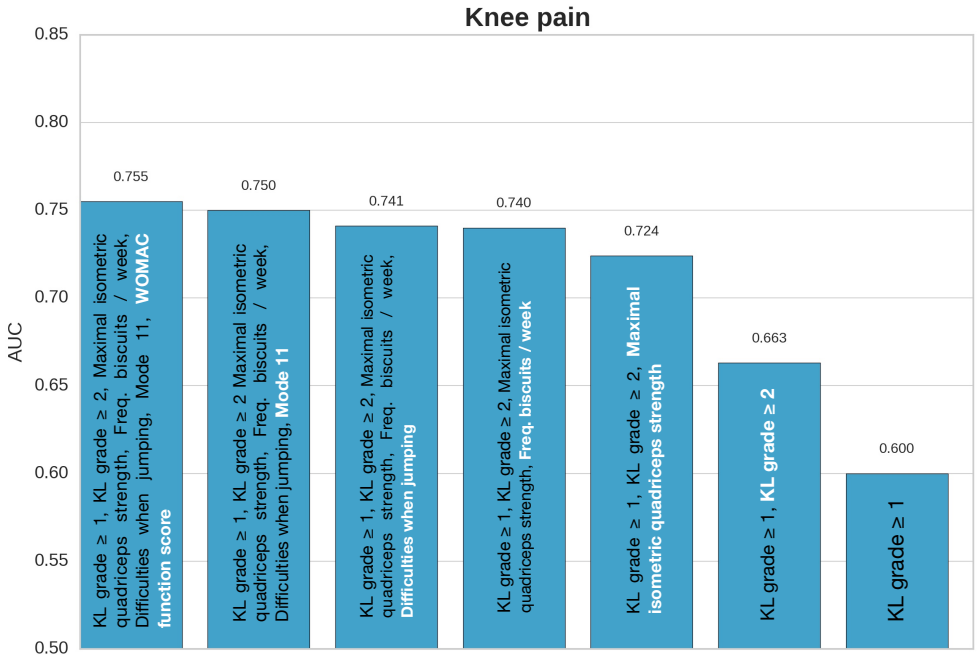


Fig 5.5: Decremental models generated using the knee pain for the knee OA incidence.

In Figure 5.4, when using the ACR criteria, the largest drop in AUC (0.075) is associated with the removal of C2M, a biomarker whose concentration levels were found to be different between OA and healthy subjects [261]. Furthermore, C2M is also the second last variable to be removed, indicating its important role in the prediction task. The K&L grade (whether at baseline is ≥ 1 or ≥ 2) represents the most valuable information in the knee pain model (Figure 5.5), by themselves, the two variables can lead to a good AUC of 0.663.

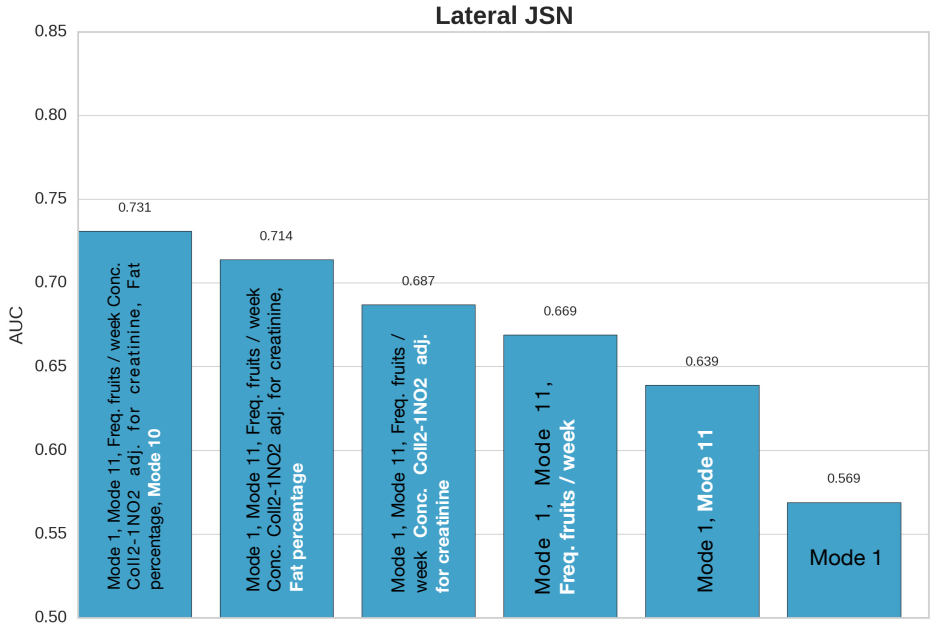


Fig 5.6: Decremental models generated using the lateral JSN for the knee OA incidence.

The analysis performed on the lateral JSN model, presented in Figure 5.6, highlights the relevance of the image-associated variables when trying to assess the incidence of OA. Mode 11 and Mode 1 (Active Modelling Shaping, a statistical model of shapes which iteratively deform to fit an image, e.g. knee image), are the last dropped variables, with the former one leading to a decrease of 0.07 regarding the AUC. The plot in Figure 5.7 suggests a relevant role, in the medial JSN model, for the information about the weekly intake of bananas. Bananas are high in magnesium and potassium which help in increasing the bone density, thus their consumption might contribute in decreasing the probability of developing knee OA. This seems also confirmed by the negative association between the "Freq. bananas / week" variable and the outcome

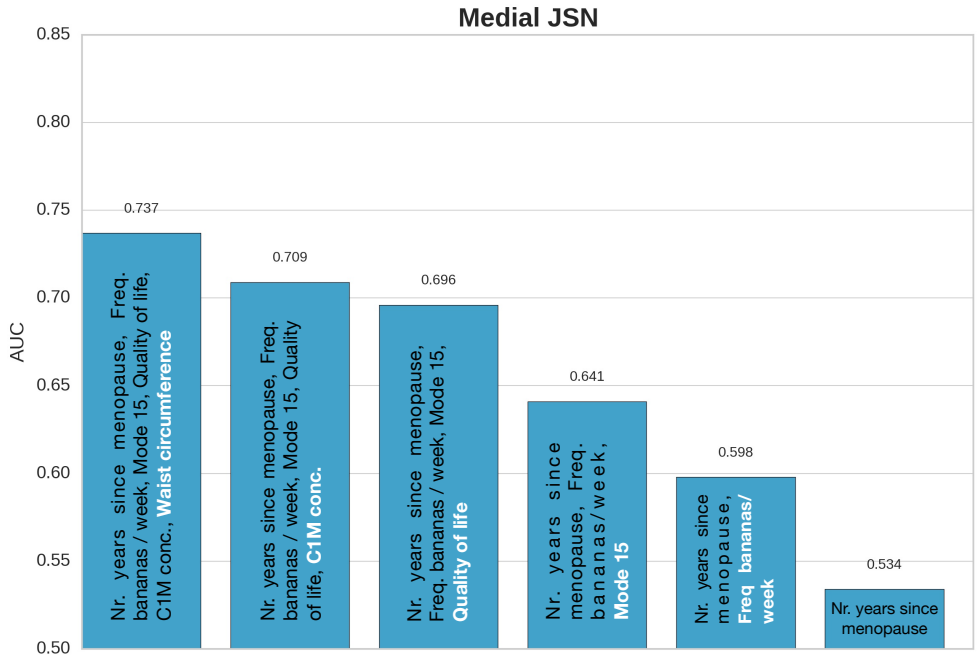


Fig 5.7: Decremental models generated using the medial JSN for the knee OA incidence.

measure depicted in Figure 5.12. Finally, in Figure 5.8, two large decreases in performance are associated to food and pain questionnaire derived information, namely the fruit intake per week and the grinding/clicking of the knee. As observed for the ACR criteria, the one variable model (presence of patello femoral OA on MRI at baseline) can lead to a reasonable AUC of 0.617.

5.3.1.2 Biomarkers association with knee OA

The analysis performed to tackle the variable direction question allows us to determine if the change in intensity (value) of a variable corresponds to an increase of the probability to be affected by knee OA (for the studied population). In order words, whether a variable has a positive or a negative association with the OA outcome measure. In Figures 5.9– 5.13, are illustrated the directions the variables for all the inferred models. Each data point corresponds to the average probability (calculated across all the samples via a 10 x 10-fold cross-validation) to belong to the positive (incidence) class. For binary variables (e.g. K&L score ≥ 1) are shown the distribution of the probabilities of the two possible values.

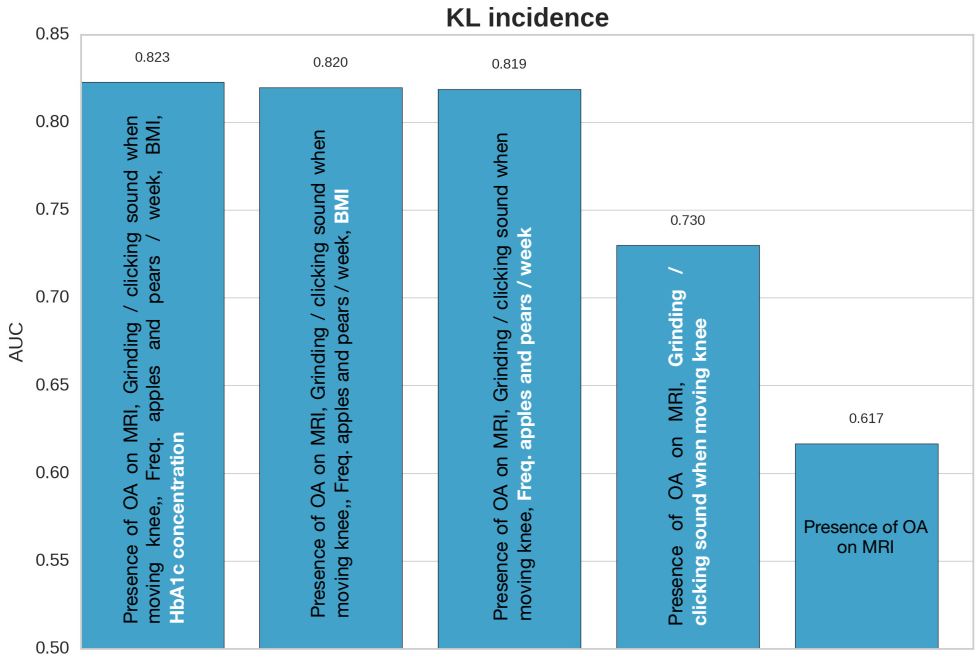


Fig 5.8: Decremental models generated using the K&L score for the knee OA incidence.

Figure 5.9 shows that as expected, the KL1 grade and the variable associated with the difficulties affecting the women when kneeling, have a positive relationship with the outcome. That is, large values (e.g. high pain) increase the chance to be affected by the condition. On the other hand, *C2M*, a biochemical marker whose levels were found to differ between OA and healthy subjects [261], have a negative relationship with the output. Similarly, it seems that (muscular) weakness might lead to a higher probability of knee OA in overweight women. For variables such as *Mode 10* is not possible to define a direct relation with the outcome measure as there is no monotonic behaviour. The predictive model is likely to contain those variables because, when coupled together with others, they provide good discrimination of the samples. Finally, *Mode 11* and *Mode 15* (Active Shape Modelling), X-ray based variables, show opposite association with the knee OA incidence.

Analogous to what was observed for the ACR criteria, the Figure 5.10 suggests that a baseline KL2 grade (boolean variable, therefore KL2 grade = 1) increases the probability for the knee OA incidence. In addition, difficulties when jumping (pain) result in a higher chance to become affected by the condition. The plot associated with the number of biscuits eaten per week illustrates a positive relationship with the incidence

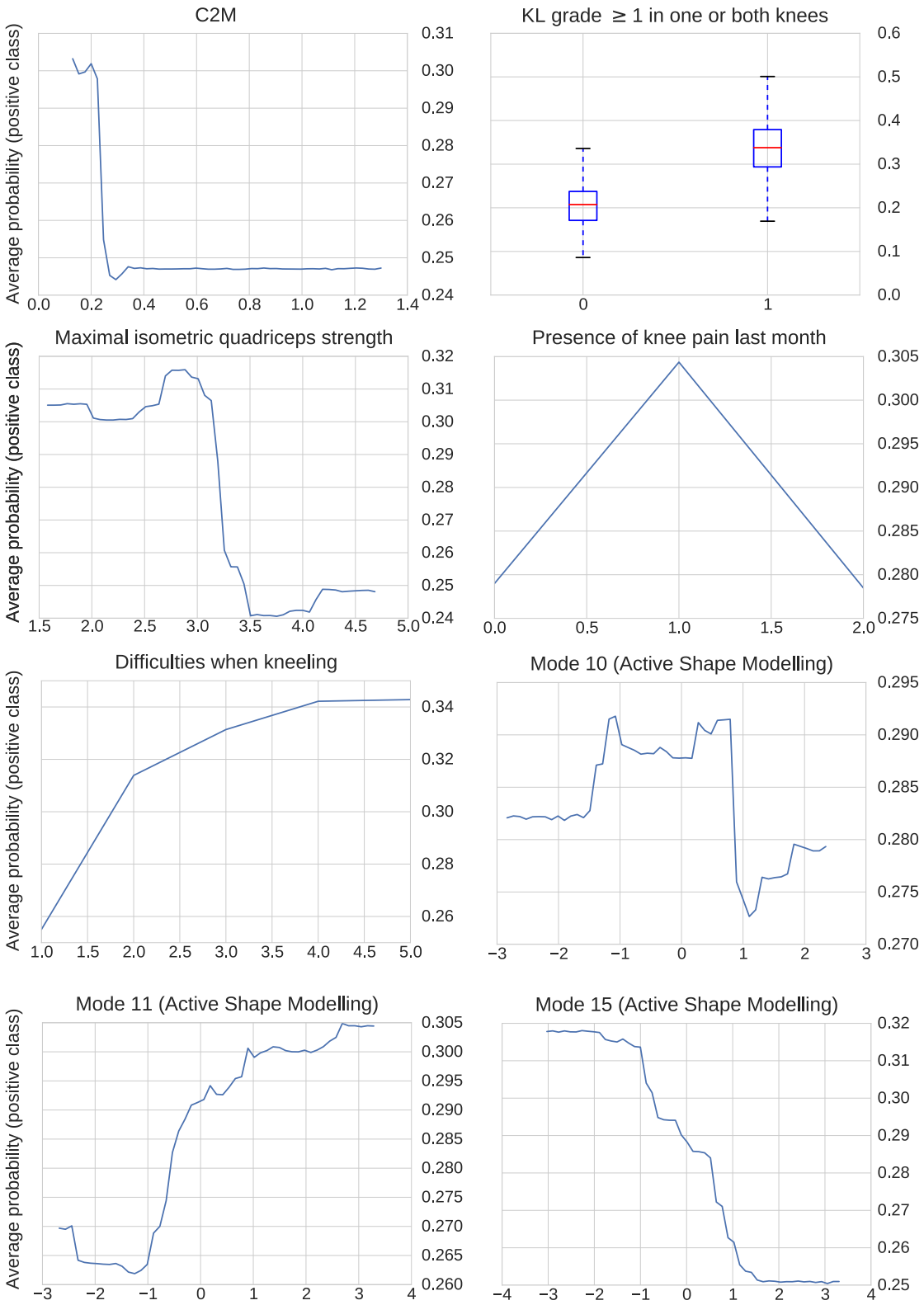


Fig 5.9: Direction of each variable in the ACR criteria model for the incidence of knee-OA. The x-axis shows the possible values of each variable, the y-axis reports the average probability to be associated with the positive class (incidence of OA).

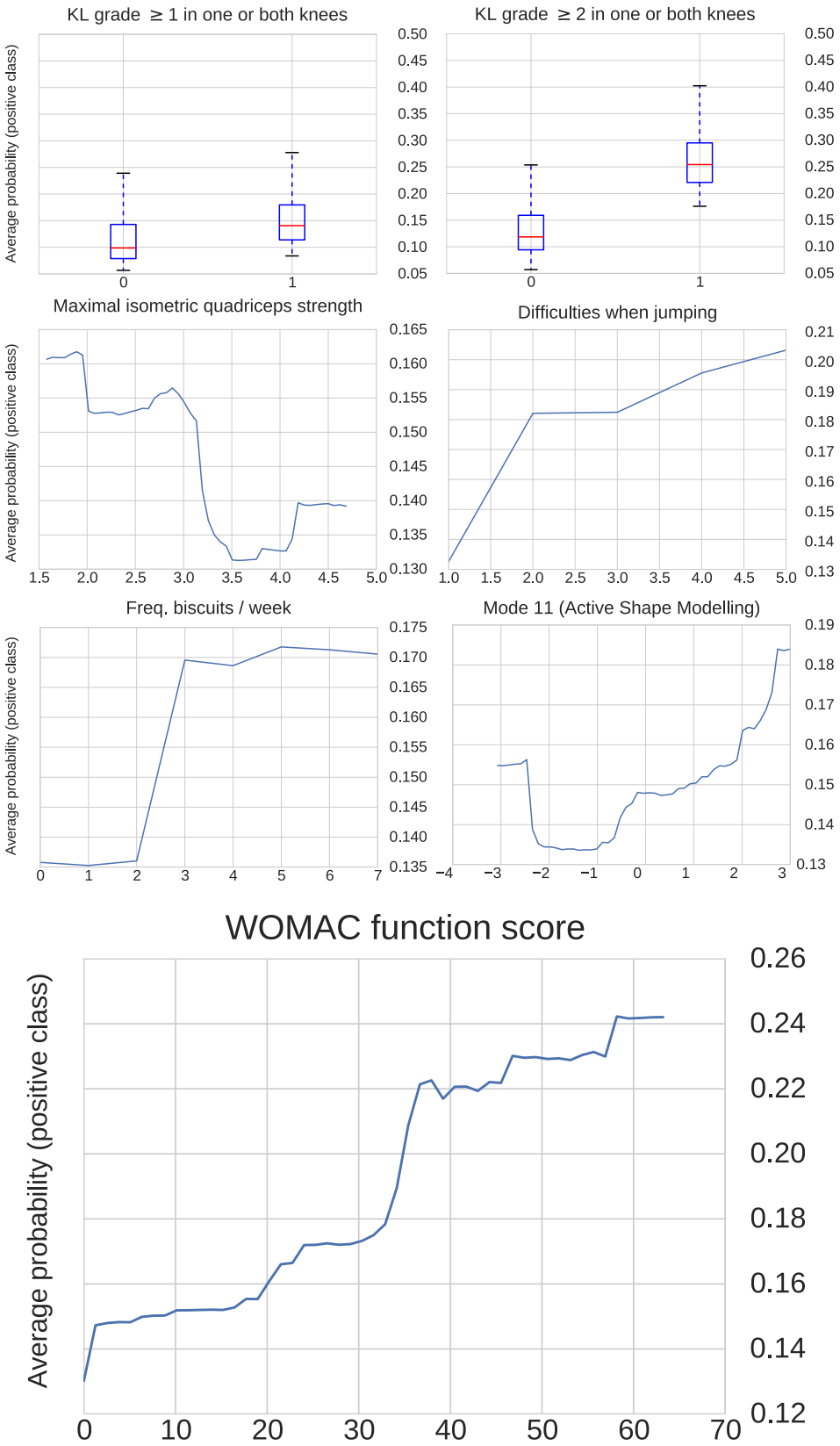


Fig 5.10: Direction of each variable in the knee pain model for the incidence of knee-OA. The x-axis shows the possible values of each variable, the y-axis reports the average probability to be associated with the positive class (incidence of OA).

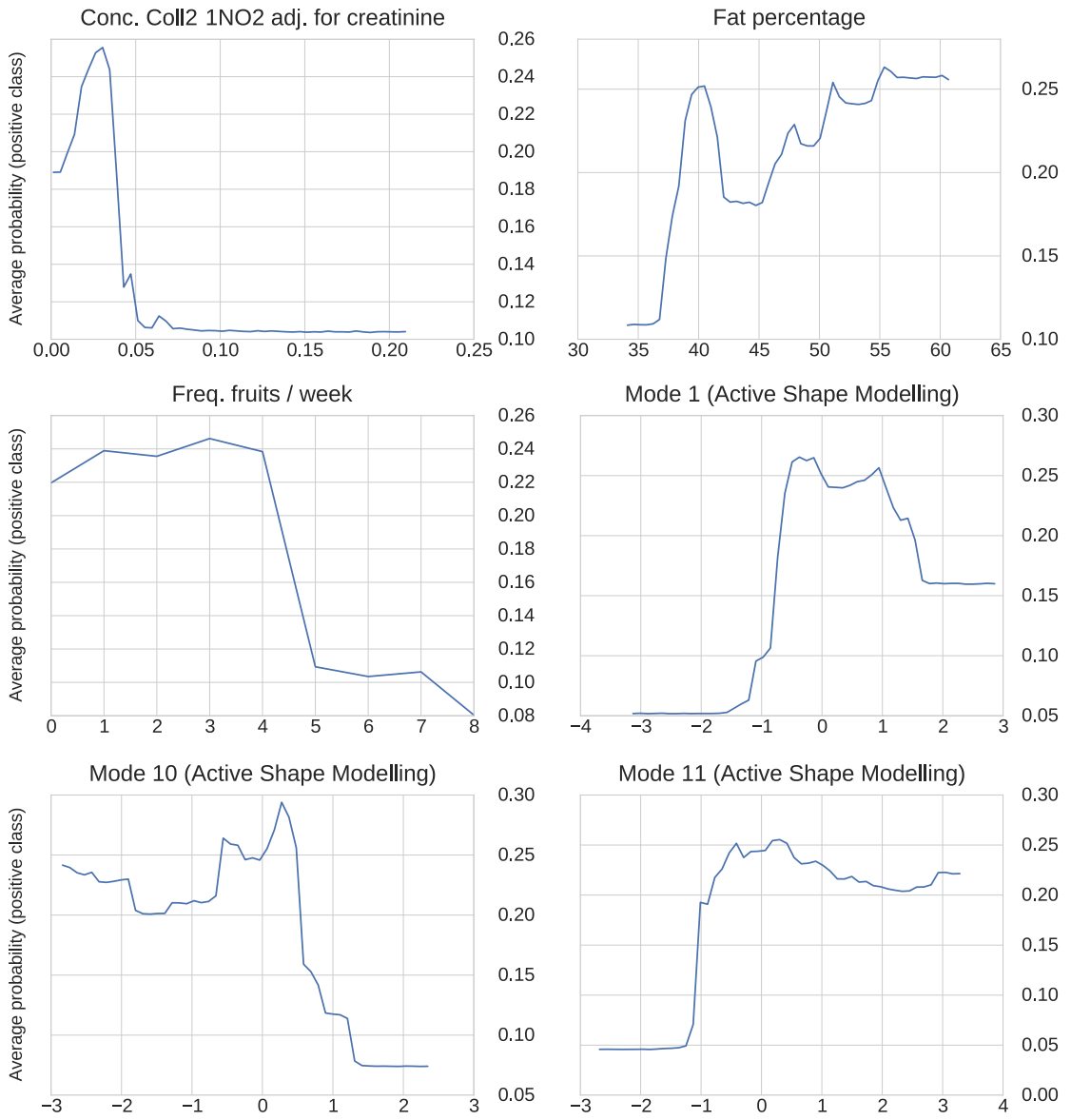


Fig 5.11: Direction of each variable in the lateral JSN model for the incidence of knee-OA. The x-axis shows the possible values of each variable, the y-axis reports the average probability to be associated with the positive class (incidence of OA).

of knee OA. This might imply that a low intake of biscuits, therefore a low quantity of consumed sugar, can help in preventing the development of the condition. Finally, intuitively, a higher WOMAC score (the Western Ontario and McMaster Universities Osteoarthritis Index that assesses pain stiffness and physical function in individuals with knee osteoarthritis), corresponds to higher probability for the presence of knee OA in the individual.

Opposite than what observed for the biscuits variable, Figure 5.11 tells that the frequency of fruits has a negative relation with the outcome variable, indicating how the assumption of fruits might mitigate the presence of knee OA. Two of the Active Shape Modelling variables, namely *Mode 1* and *Mode 11* have a positive association with the lateral JSN incidence outcome, while *Mode 10* shows an overall negative relationship. Additionally, as expected, women with an elevated fat percentage give the impression to be more affected by the osteoarthritis in the joints. Finally, the most interesting remarks can be made on the association between the concentration of Coll2-1NO2 with the incidence of knee OA. The identified negative relationship is confirmed by the findings of Landsmeer et al. [244], where the PROOF study data analysis, performed with a binary logistic regression, showed that small values of Coll2-1NO2 were related to a greater incidence of knee OA.

The *Waist circumference* plot, created from the medial JSN model and available in Figure 5.12, can be associated with *Fat percentage* plot in Figure 5.11. Large values for both variables hint a prominent overweight condition, that can lead to higher chances for the joint inflammation to appear. Interesting, the PROOF data suggests how the knee OA tends to be more developed in the time frame close to the menopause. In fact, the partial dependence analysis shows that as the number of years since the occurring of menopause increases, the overall probability decreases. Finally, as already previously highlighted in this section, higher fruit intake, bananas in this instance, might lead to a reduction in the chance of developing knee OA.

In Figure 5.13, the BMI and the frequency of apples and pears per week confirm all the conclusion drawn so far. The plot indicates that a higher amount of fruit that can be seen as an overall healthier diet can decrease the possibility of knee inflammation. The frequency of grinding or clicking sounds when moving the knee, from the PROOF study samples, seems to be negatively associated with the incidence of knee OA. In addition, the analysis of MRI (at baseline) might provide insights about the development of knee OA in the individuals (overweight women in this instance).

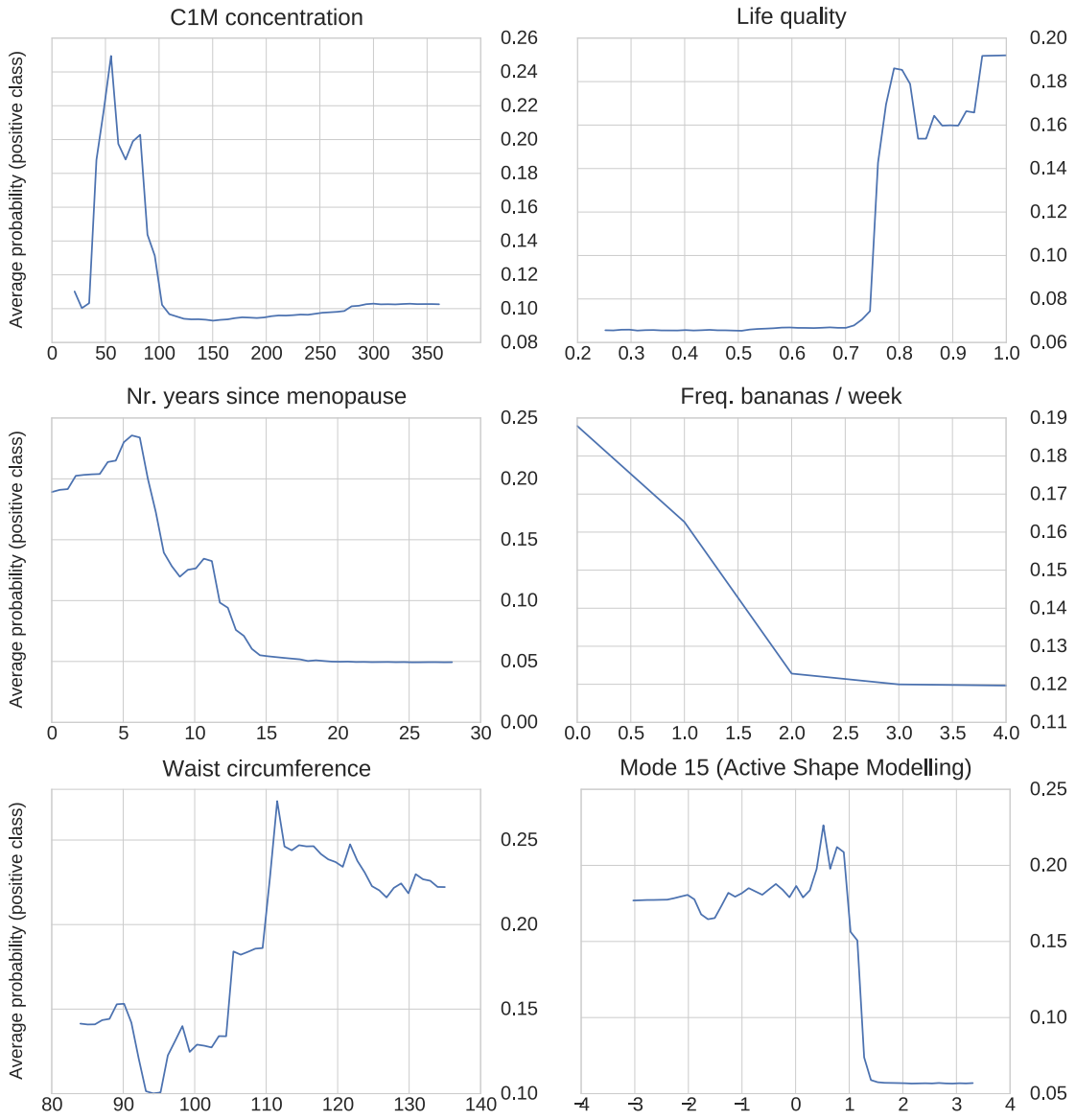


Fig 5.12: Direction of each variable in the medial JSN model for the incidence of knee-OA. The x-axis shows the possible values of each variable, the y-axis reports the average probability to be associated with the positive class (incidence of OA).

5.3.1.3 Comparison with literature findings

To check if the proposed biomarkers can improve the state-of-the-art solutions, the specialised literature was searched. The aim was to compare the inferred models with what is currently reported in the literature. After a thorough mining, only two models (knee pain and K&L score incidence) were found comparable with the literature findings. For a fair evaluation, for each identified literature study, only the AUC calculated using an internal validation (some studies also used external data to assess their performance) was considered. If multiple models (having a different subset of

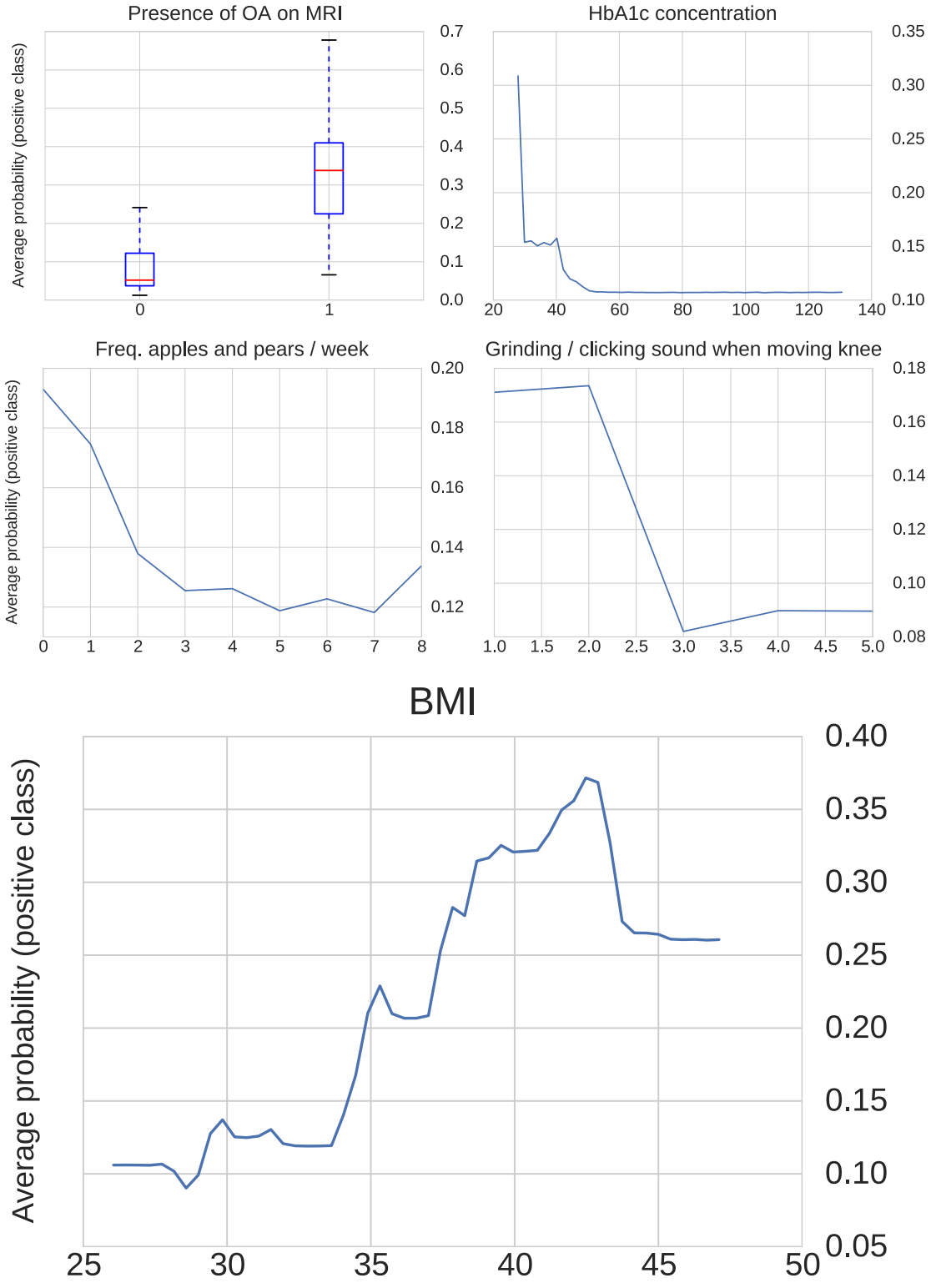


Fig 5.13: Direction of each variable in the K&L score incidence predictive model for the incidence of knee-OA. The x-axis shows the possible values of each variable, the y-axis reports the average probability to be associated with the positive class (incidence of OA).

Ref.	AUC	OA defintion	Attributes
[262]	0.790	KL grade < 2 B/L and KL ≥ 2 at F.U.	Gender Age BMI Knee pain KL score 1 at B/L
[263]	0.690	KL grade < 2 B/L and KL ≥ 2 at F.U.	Gender Age BMI Occupational risks Family osteoarthritis Previous knee injury
[264]	0.740	KL grade < 2 B/L and KL ≥ 2 at F.U.	Gender Age BMI Minimum JSW Osteophyte

Table 5.5: Summary of the K&L score incidence models found in the specialised literature. JSW: Joint Space Width, B/L: baseline, F.U.: follow-up.

variables) were available, the best performing one selected. All the identified models were generated using the same learning approach: a univariate analysis followed by a multivariate logistic regression method. A summary of the literature finding is reported in Table 5.5 for the K&L score incidence, while Table 5.6 shows the models associated with knee pain. In [262–264] there was no description of the validation protocol employed to generate the reported AUC values. Therefore, it is assumed, as standard practice for clinical studies, that the published AUCs are equivalent to the AUC-Full values. In [265] is mentioned that a *"10-fold cross-validation strategy has been used as a feature selection strategy"*, however, it is not specified if the reported AUCs were calculated with such procedure or using the whole set of samples for learning. The internal validation (AUC-Full from 10 x of 10-fold cross-validation) of the RGIFE selected models indicated an AUC of 0.823 with the K&L score incidence and an AUC of 0.755 when using the knee pain (see Figure 5.3).

The model for the prediction of K&L score incidence has higher predictive performance than all the models available in the specialised literature while using the same or fewer variables, see Table 5.5. A superior performance can also be clearly noticed

Ref.	AUC	OA defintion	Attributes
[264]	0.600	Painful knee at B/L; painful knee at F.U	Age Pain intensity Minimum JSW Osteophyte
[265]	0.623	Chronic right knee pain	Osteophytes (OARSI grades 0–3) femur medial compartment Chondrocalcinosis (grades 0-1) medial compartment
[265]	0.740	Chronic right knee pain	Osteophytes (OARSI grades 0–3) femur medial compartment Osteophytes (OARSI grades 0–3) femur lateral compartment Chondrocalcinosis (grades 0-1) medial compartment

Table 5.6: Summary of the knee pain models found in the specialised literature. JSW: Joint Space Width, B/L: baseline, F.U.: follow-up.

when comparing the knee pain model with the literature, Table 5.6. Different than for the K&L score incidence, the knee pain model is larger (7 attributes) and more heterogeneous (imaging-based information, food and pain questionnaire data, OA and clinical information). The PROOF study analysed overweight middle-aged women, therefore information such as gender or age, that are consistently used in the literature, would not be emerge as relevant. Nevertheless, it is interesting to see how *BMI* constantly appears across the K&L score incidence literature models as well as in the 5-variables model. Overall, this comparison showed how the proposed biomarkers, for two knee OA outcome measures, perform better than the current models available in the literature while being similar or smaller in terms of size.

5.3.2 6.5 years predictive models

Out of all the subjects included in the PROOF study, only 74 had a 6.5 years follow-up lipidomics data. A different total number of subjects was available for different knee OA outcome measures. The OA assessed with the ACR criteria and chronic knee pain after 6.5 years occurred in 15 (20%) and in only 9 out of 74 (12%) women respectively. The incidence of K&L ≥ 2 was measured in 17 (23%) out of 74 individuals. Given the small number of OA samples available in the lipidomics data, the 10 x 10-fold cross-validation scheme (as employed in the analysis of the 2.5 years data) was ruled out. Instead, a leave-one-out validation was preferred (the number of folds is equal to the number of samples, 74 in this instance). Initially, the performance of the predictive models was tested using a 10 x 5-fold cross-validation. However, this provided an elevated instability of AUC when duplicating the experiments with different repe-

titions of the validation process (large standard deviation values). Therefore, the only remaining option was the LOOCV, which if on the one hand provides more robust estimations, it also tends to over-estimate the prediction performance of the models [30]. Furthermore, different than for the analysis of the 2.5 years data, the lipidomics data were not affected by missing values, thus RGIFE was only coupled with SPIDER to tackle the imbalance class distribution (together with a cost-sensitive learning). To identify the best set setting of RGIFE (across 30 different configurations) for each of the analysed data, the same pipeline used for the 2.5 years data was employed (Figure 5.2). For the sake of readability, the results section will be divided according to the source that generated the lipidomics data.

OA measure	Variables	AUC-Full	AUC-CV	p-value
ACR criteria	C16:0-Cer C24:2-Cer C18:0-ChE	0.921	0.757	$<$ 1e-04
Knee pain	C22_3-2 C38:4-DG C40:6-DG C60:6-TG	0.855	0.443	0.0001
KL incidence	C18_3-w C36:1-PC C42:7-PC C38:5-PCplas C22:4-ChE	0.910	0.561	3.00e-04

Table 5.7: Summary of the models inferred for each knee OA outcome measure when using TNO lipidomics data. The last column indicates the permutation test p-value (one tailed).

5.3.2.1 Selected models from TNO data

Table 5.7 reports the models extracted from the analysis of the TNO lipidomics data using three knee OA outcome measures. For each model are provided both the AUC-Full and the AUC-CV, with the latter useful to identify possible overfitting during the learning phase.

Figure 5.14 shows instead the ROC curves generated by the models of Table 5.7. All the inferred sets of lipids provide high AUC-Full values, however, when observing the AUC-

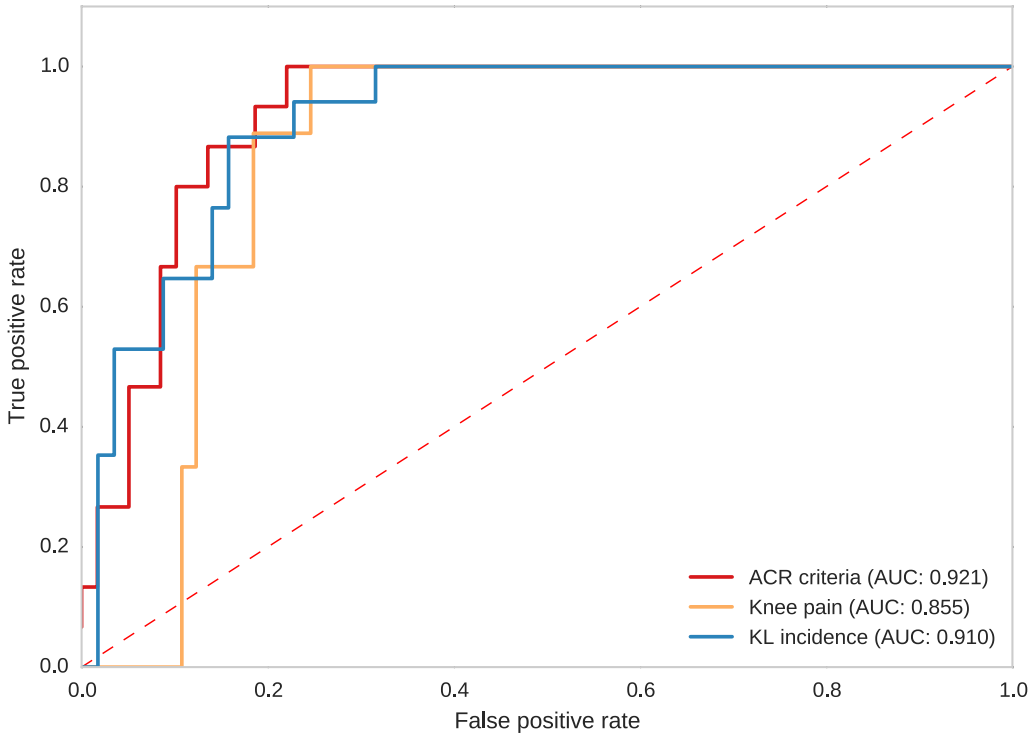


Fig 5.14: The ROC curves generated by the best performing models (6.5 years). The AUC values refer to the AUC-Full.

CV, bad performance emerge for the knee pain models. Having an AUC-CV lower than 0.5 (random classification) means that the learning phase was not able to extract any significant pattern from the data that could be later used to correctly predict the category of unseen data (test samples). The reduced AUC values for this model are likely due to the few positive individuals (9 against 65 non-OA) available when defining the presence of knee OA using the chronic pain measure. With such a small number of positive observations, it is almost impossible for the machine learning methods to identify meaningful patterns from the data. Conversely, the model generated using the ACR criteria performed well providing an AUC-CV of 0.757 and an AUC-Full of 0.921 by using only three lipids. Finally, the lipids selected using the K&L score incidence criteria lead to an interesting AUC-Full of 0.910 coupled by a lower AUC-CV of 0.561. All the inferred models were defined by a small number of lipids. This suggests that many of the original 249 variables do not contain any useful information for the prediction of the knee-OA presence. When using the permutation test, none of the inferred models was outperformed, in terms of AUC, by models generated using random datasets (empirical p-value = 0). Similarly, the one-tailed test also provided

low p-values when contrasting the original AUC-Full values with the performances obtained with the random datasets. The p-values calculated for each model with the one-tailed test, are shown in Table 5.7, all of them are lower than the classical standard α value of 0.01, indicating significant performance for the proposed models.

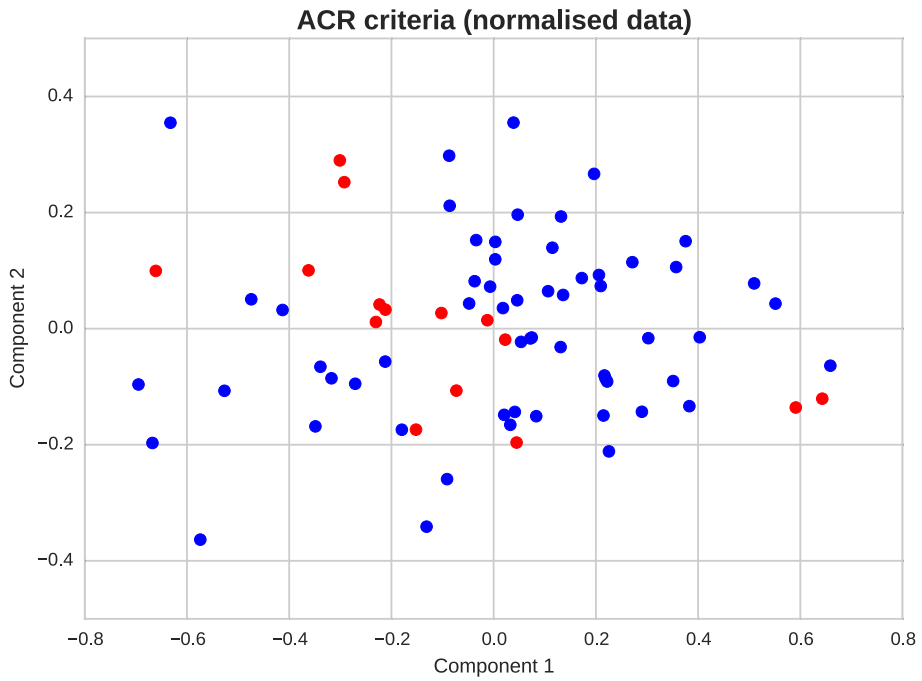


Fig 5.15: Plot of the first two components of the PCA performed using the biomarkers extracted from the (normalised) ACR criteria lipidomics data.

The good discriminative power of the biomarkers extracted using the ACR criteria measure is also shown by the plot associated to the Principal Component Analysis (PCA) in Figure 5.15. The PCA is a method that aims to emphasise variation and bring out strong patterns in a dataset. When considering only the first two components, it is possible to generate a plot and assess how well they discriminate and cluster samples that belong to different classes. In Figure 5.15 the majority of the OA-affected individual are clustered together (centre of the plot) when using the two main components extracted from the selected biomarkers. Two (positive) outliers are also clearly visible on the right-hand side of the plot. In the PCAs generated from the other two datasets, the individuals of different classes tended to overlap making difficult the identification of "pure" clusters of data points, thus they are not reported.

5.3.2.2 Additive values of the biomarkers from TNO data

To assess the relative importance of each biomarker, within the predictive models, decremental models were composed by removing the variable whose dismiss led to the smallest drop in performance. The AUC of each decremental model was calculated via a LOOCV. Figures 5.16 – 5.18 are shown the submodels associated to each OA outcome measure. When using the ACR criteria, a single lipid (*C16_0_Cer*) guarantees a high AUC of 0.713, while the removal of *C18_0_ChE* bring the largest drop in performance. This is not true for the Figure 5.17 and Figure 5.18, where both single-lipid models obtain low performance. The largest drop in AUC, for all three OA outcome measures, occurs when removing the second last important variable. Finally, for both K&L score and knee pain, using only two lipids it is possible to obtain and AUC-Full of 0.711 and 0.651 for knee pain and K&L respectively. Overall, the submodels analysis indicates that with the abundance of very few lipids, selected from the original set of 294, is possible to achieve a good prediction of knee OA presence in overweight women.

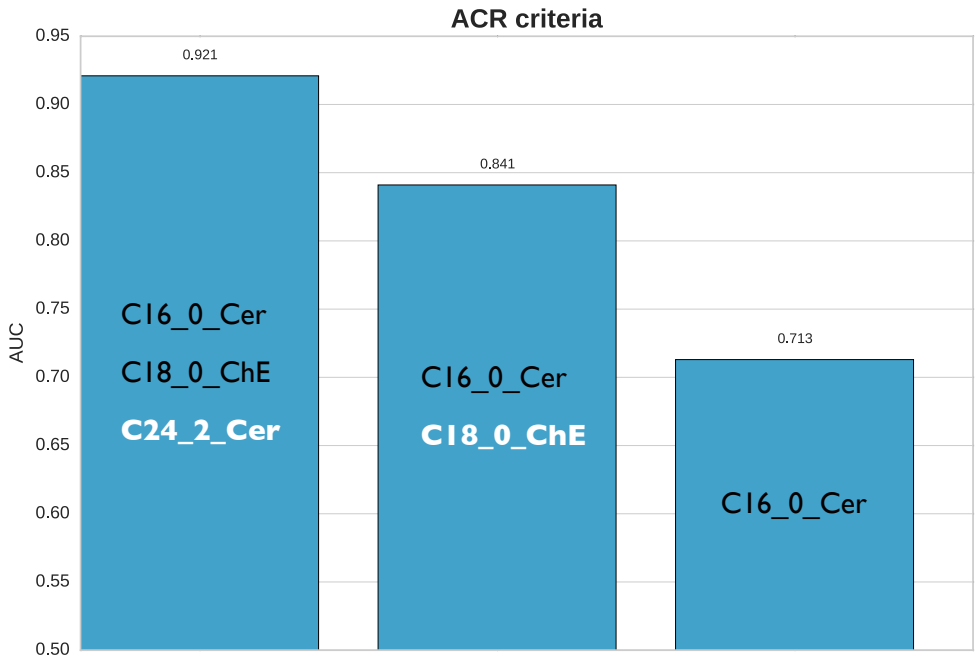


Fig 5.16: Decremental models generated using the ACR criteria to define the knee OA presence.

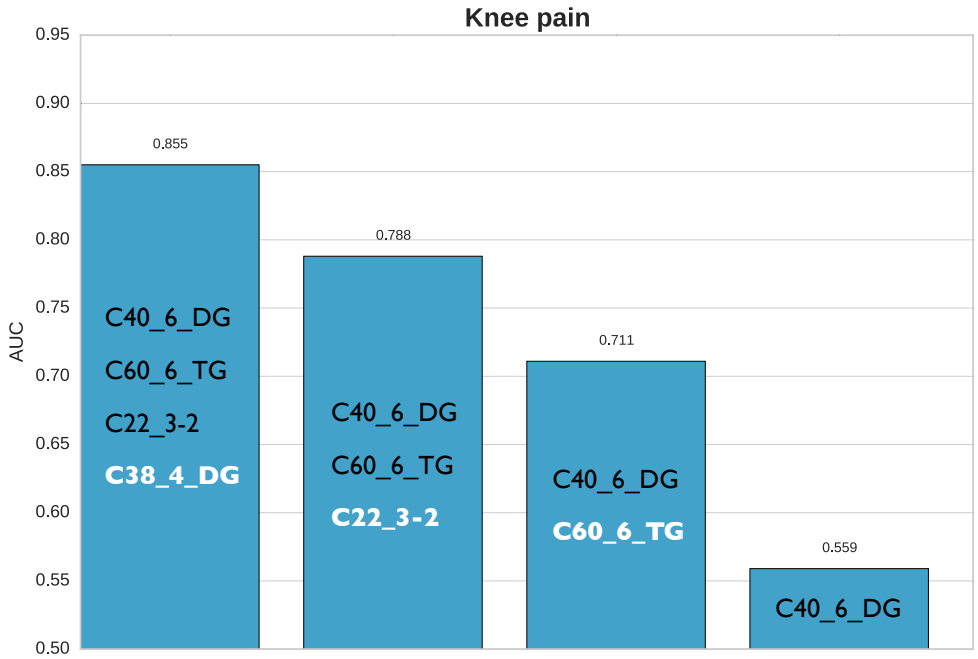


Fig 5.17: Decremental models generated using the knee pain to define the knee OA presence.

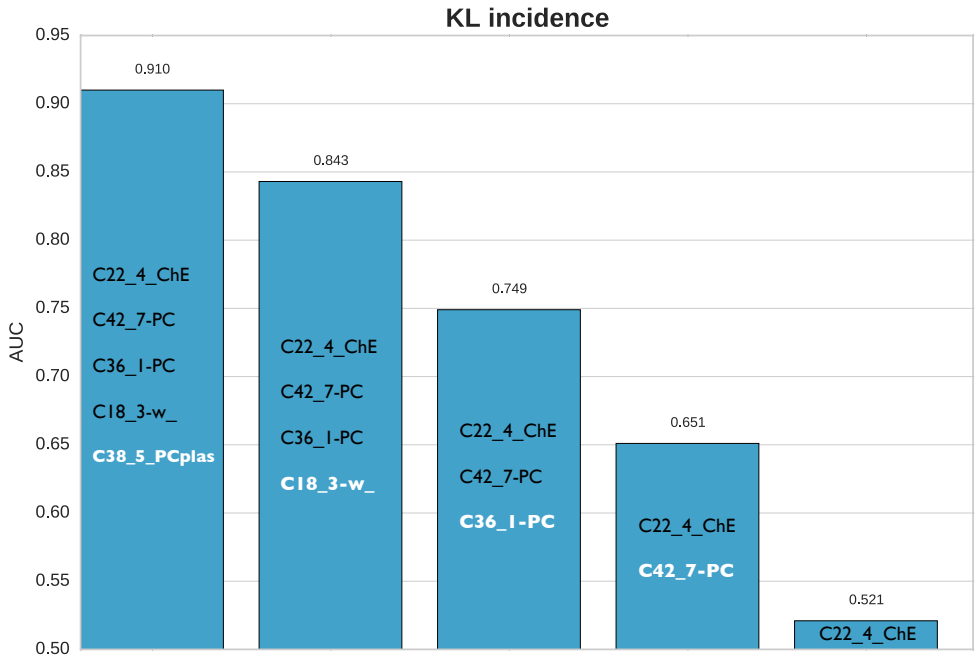


Fig 5.18: Decremental models generated using the K&L score to define the knee OA presence.

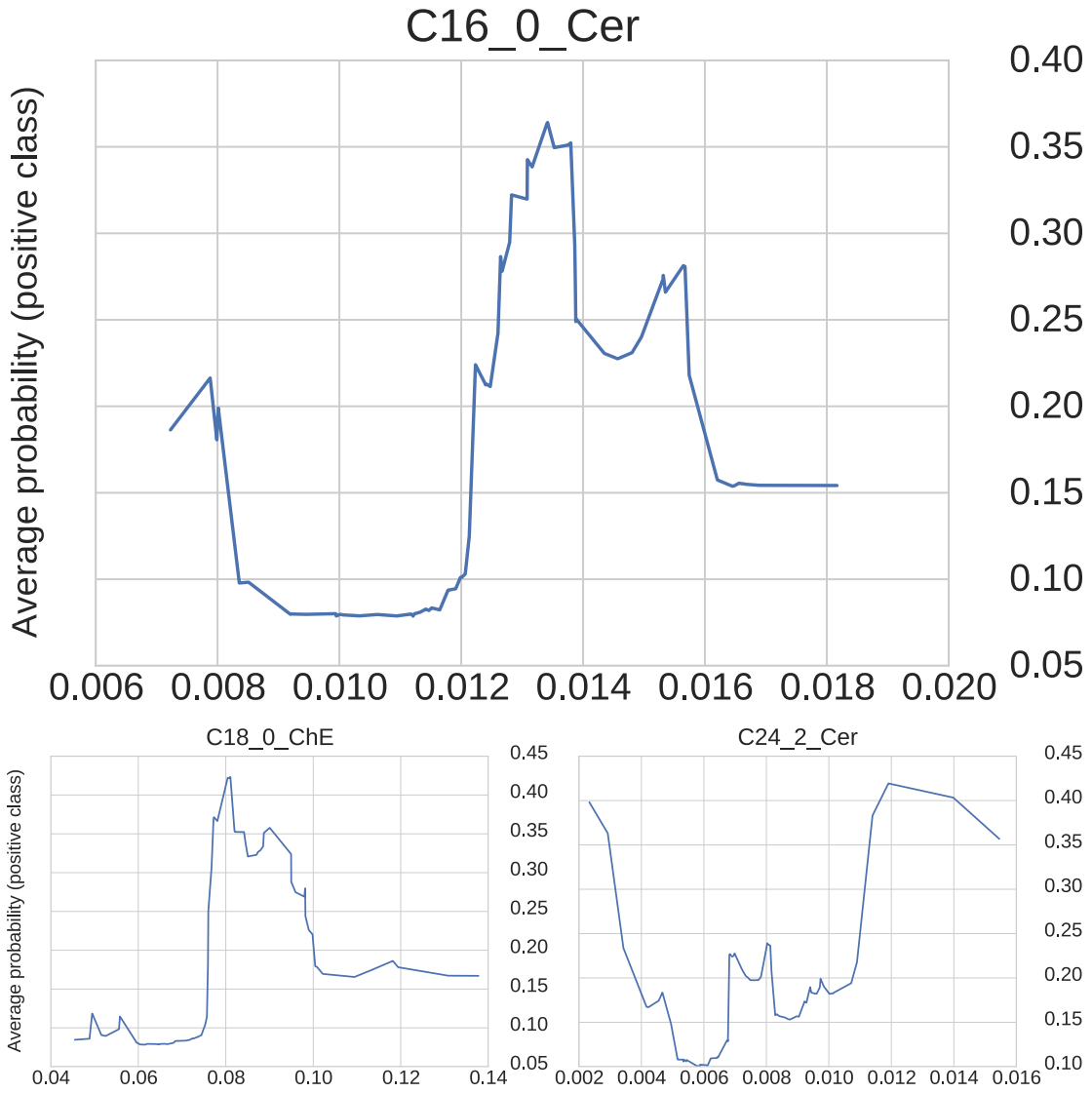


Fig 5.19: Direction of each lipid in the ACR criteria predictive model for the presence of knee-OA. The x-axis shows the abundance of each lipid, the y-axis reports the average probability to be associated with the positive class (presence of OA).

5.3.2.3 Biomarkers association with knee OA from TNO data

In contrast with the analysis of the biomarkers extracted from the 2.5 years data (see Section 5.2.3.4), for several lipids it was not possible to identify a direct association with the knee OA outcome measure. In Figures 5.19 – 5.21 are provided the partial dependency plots for the variables of all the inferred models. Two of the three lipids extracted with the ACR criteria definition show a $\bigwedge$-shape relationship, where an higher probability of developing the knee OA is associated with mid-range values of the lipid. In other words, the lipids have a positive association up to a threshold

values after which it becomes negative (large values). Conversely, *C24_2_Che*, shows an association in the form of a $\bigvee$ where the chance of knee OA decreases to a minimum, usually associate with mid-range values for the lipid, before rising again. This type of relationship can be split into a negative association followed, after a threshold value, by a positive one.

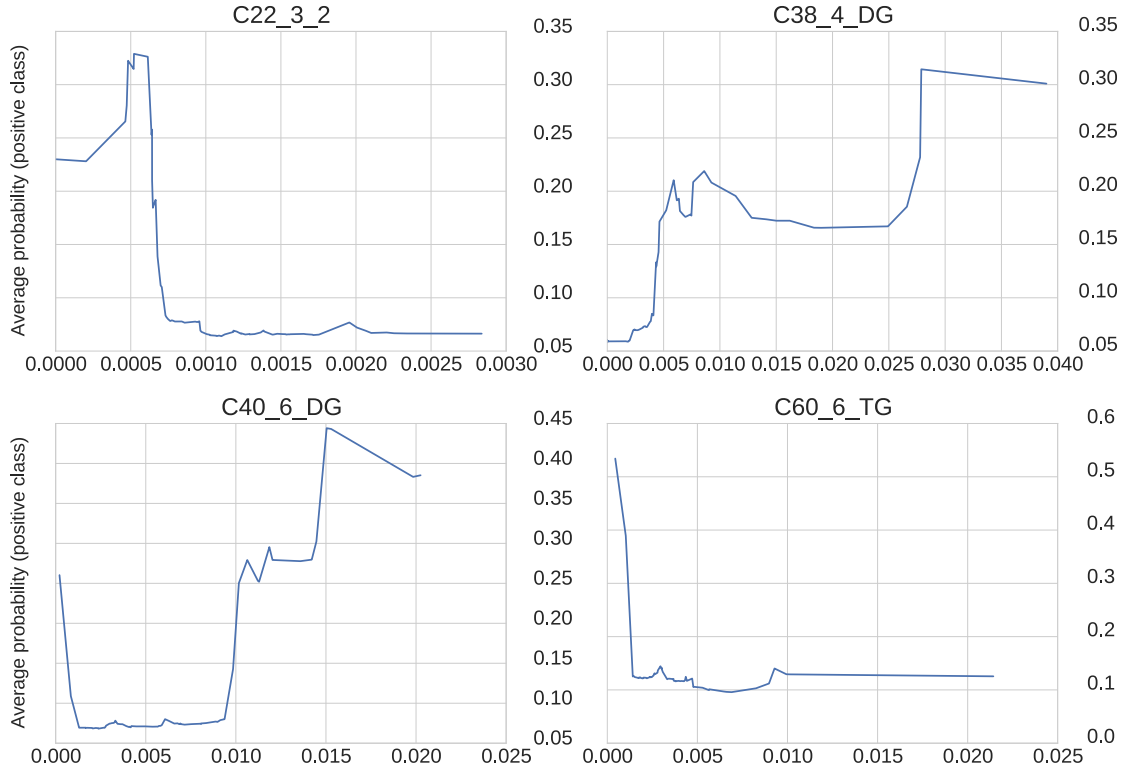


Fig 5.20: Direction of each lipid in the knee pain predictive model for the presence of knee-OA. The x-axis shows the abundance of each lipid, the y-axis reports the average probability to be associated with the positive class (presence of OA).

A clear association is instead visible for the lipidomics selected from the chronic knee pain OA definition, Figure 5.20. A well defined negative relationship is plotted for *C60_6_TG* and *C22_3_2*, while a positive dependence emerges for *C38_4_DG*. Finally, *C40_6_DG* seems to provide a $\bigvee$ association, in fact, there is a big increase in with the probability of incurring in the condition for values higher than 0.010.

Figure 5.21 illustrates the partial dependence plots for the lipids in the K&L score incidence model. The majority of them have a negative association with the outcome measure, while only one, *C38_5_PCplas*, shows a positive relationship. Finally, a peak at the value of 0.8 characterises the $\bigwedge$-shape dependence of *C36_1_PC*.

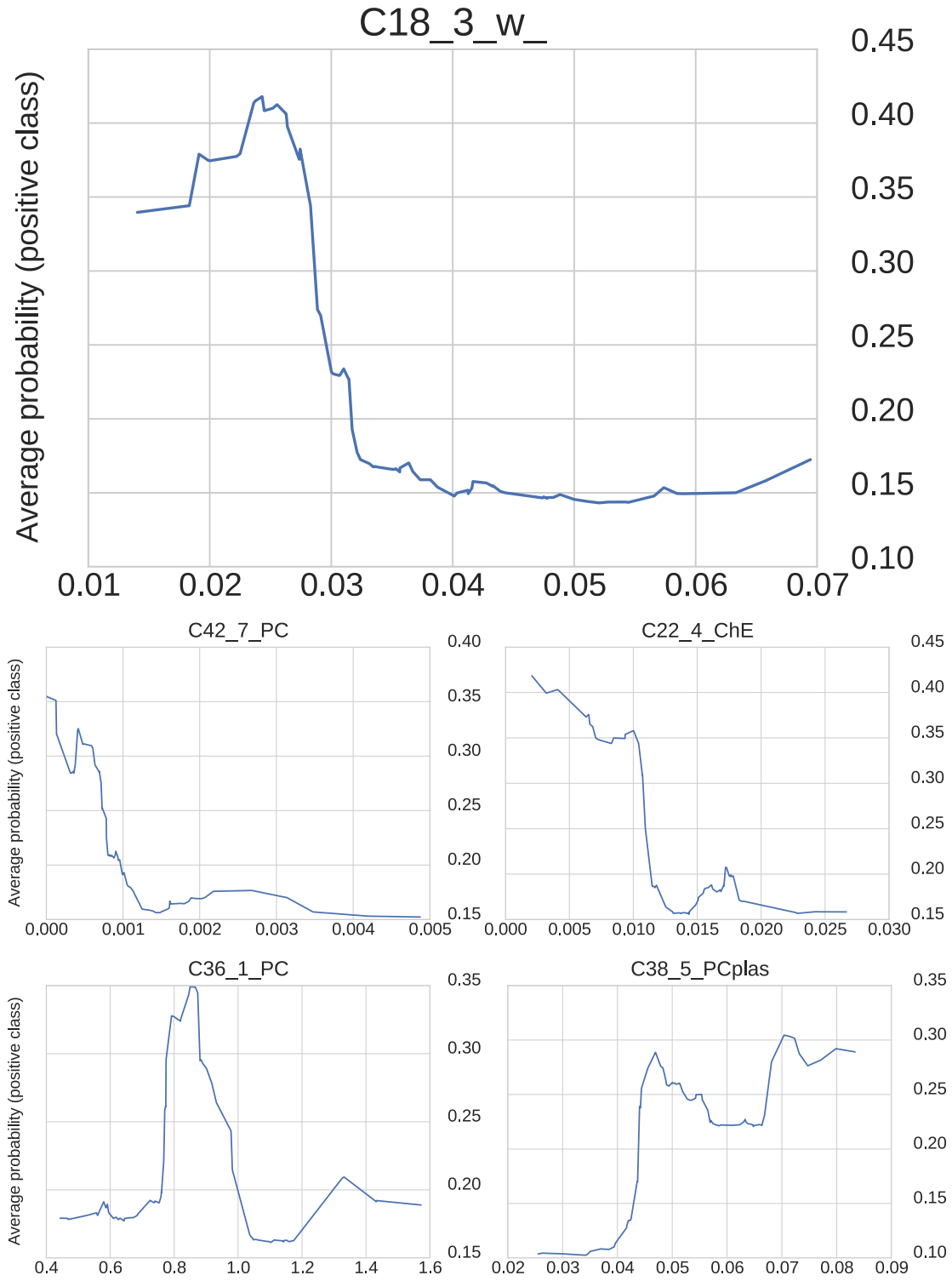


Fig 5.21: Direction of each lipid in the K&L score incidence predictive model for the presence of knee-OA. The x-axis shows the abundance of each lipid, the y-axis reports the average probability to be associated with the positive class (presence of OA).

5.3.2.4 Functional networks from TNO data

Using the FuNeL protocol, functional networks were generated from the lipidomics datasets. The goal was to go beyond the pure list of biomarkers and generate some networks that can help in better understanding the mechanism behind the characterisation of knee OA. Before the generation of the networks, it was necessary to assess the predictive value of the BioHEL's rules generated from the lipidomics data. In fact, a network generated from classification rules that cannot well discriminate the samples, *OA* and *non-OA* in this instance, simply risks to be overfitting. Therefore, the AUC of 10 000 rule sets generated by BioHEL was calculated using a leave-one-out validation scheme. The resulting values were: 0.696 for the ACR criteria, 0.602 for the knee pain and 0.529 for K&L score incidence. When comparing those values to the performance of RGIFE, in Table 5.7, lower, but still good, performance were obtained when using the ACR criteria data. Conversely, by using all the attributes (no feature selection was performed before BioHEL), a better AUC was achieved for knee pain, while a slightly worse classification occurred with the K&L score incidence measure. Overall, the AUCs obtained by BioHEL represent a good result considering the difficulty of the data and that neither cost-sensitive learning or oversampling was performed during the learning process.

Having assessed the predictive value of BioHEL's classification rules, three functional networks were generated from the TNO lipidomics data. A summary of the main topological properties of the functional networks is provided in Table 5.8. Each network is illustrated in the Appendix C. FuNeL contains an embedded feature selection process based on the classification rules generated by BioHEL. Attributes that are not relevant for the classification task do not emerge in BioHEL's rule and consequently in the network. Moreover, attributes (lipids) that do not belong to statistically significant

OA definition	Nodes	Edges	Diameter	Clust. Coeff.	Density	Avg. Degree
ACR criteria	222	1032	6	0.359	0.042	9.29
Knee pain	239	958	8	0.225	0.034	8.01
KL incidence	261	1361	5	0.359	0.040	10.42

Table 5.8: Summary of the main topological properties of the networks generated from the TNO lipidomics.

OA measure	Biomaker	Degree	Network avg. degree	$\overline{SPL}$ (biomarkers)	$\overline{SPL}$ (others)
ACR criteria	C16:0-Cer	116	9.29	1.33	2.57
	C24:2-Cer	1			
	C18:0-ChE	72			
Knee pain	C22_3-2	2	8.01	1.50	2.80
	C38:4-DG	91			
	C40:6-DG	17			
	C60:6-TG	8			
KL incidence	C18_3-w	24	10.43	2.00	2.44
	C36:1-PC	138			
	C42:7-PC	1			
	C38:5-PCplas	58			
	C22:4-ChE	9			

Table 5.9: Role of the RGIFE selected lipids within the FuNeL networks. For each biomarker is reported the node degree. In addition is shown the average shortest path length between the RGIFE-selected lipids and between all the other nodes in the network.

edges are not represented by a node. As a consequence, all the inferred networks contain fewer lipids than the original set available in the data. The *density* is the property that measures how many edges are in a network compared to the maximum possible number of edges between all the nodes and ranges between 0 (no edges) and 1 (complete graph). The low values reported in Table 5.8 show that, although the representation of the networks might suggest that FuNeL generated some sorts of hair-balls, where all the nodes are connected to each other, the number of edges between the lipids is limited (in the range of thousands). The *clustering coefficient* is relatively low for all the networks and indicates the presence of few triangular relationships between the lipids. Finally, the *diameter* reveals small networks for ACR criteria and K&L score incidence, where the two farthest nodes are separated by 5 and 6 edges respectively. Conversely, the network generated from the knee pain data is slightly larger with a diameter of 8.

The networks were then used to assess the role and the position of the lipids selected by RGIFE. In Table 5.9 are reported, for each lipid, the number of neighbours in the network (i.e node degree). Most of the proposed biomarkers are central hubs in the FuNeL networks, that is they are connected with many other lipids. Particular importance is assumed by *C16_0_Cer*, *C38_4_DG* and *C36_1-PC* respectively in the ACR criteria, knee pain and K&L score incidence network. When comparing their degree with the average node degree of the networks reported in Table 5.8, a clear difference

emerges, up to 10 times more connections. In addition, it was checked "how close" the proposed biomarkers are, from each other, within the FuNeL network. Similarly to what measured with the *proximity* of the disease-associated genes in Chapter 3, the average distance (shortest path length) between the RGIFE-selected lipids was contrasted with the average distance between the remaining lipids of FuNeL networks. The values of Table 5.9 indicates a higher proximity of the RGIFE-selected biomarkers if compared with the average distance between any other pair of lipids. Greater proximity represents a stronger functional relationship between the proposed biomarkers, in other words, they are at the core of the co-prediction networks. From a machine learning point of view, the importance of those results is two-fold: (1) it shows that the proposed biomarkers are identified as influential in the PROOF study data by two different machine learning methods employing a diverse knowledge representation and (2) it proves the robustness of the proposed methods (RGIFE and FuNeL) as similar knowledge, although represented in different form, is extracted from the same data.

As already mentioned earlier, it is difficult to extract some insights by looking at the whole network (see Appendix C). Therefore, a clustering algorithm was applied to identify a set of strongly connected nodes. MCODE, a Cytoscape plugin that finds clusters (highly interconnected regions) in a network [131], was used for this purpose. In Figure 5.22 are illustrated the highest ranked clusters, found by the MCODE algorithm (used with default parameters), for the three inferred TNO networks. With a red circle are highlighted the RGIFE-selected lipidomics within the major clusters. In the knee pain cluster, no biomarkers were present. However, they were present in the lower ranked clusters (not shown here). Overall, Figure 5.22 visually shows the large number of connections that involves *C16_0_Cer* and *C18_0_ChE* in the ACR criteria network and *C36_1-PC* and *C38_6_PCplas* in the K&L score incidence network. Once again, this reveals that several of the proposed lipids seems to play a relevant role in different models extracted from the analysis of the PROOF data study.

5.3.2.5 Selected models from UNOTT data

When analysing the data generated from UNOTT, the biomarkers selected by RGIFE obtained poor values of AUC-CV, respectively:

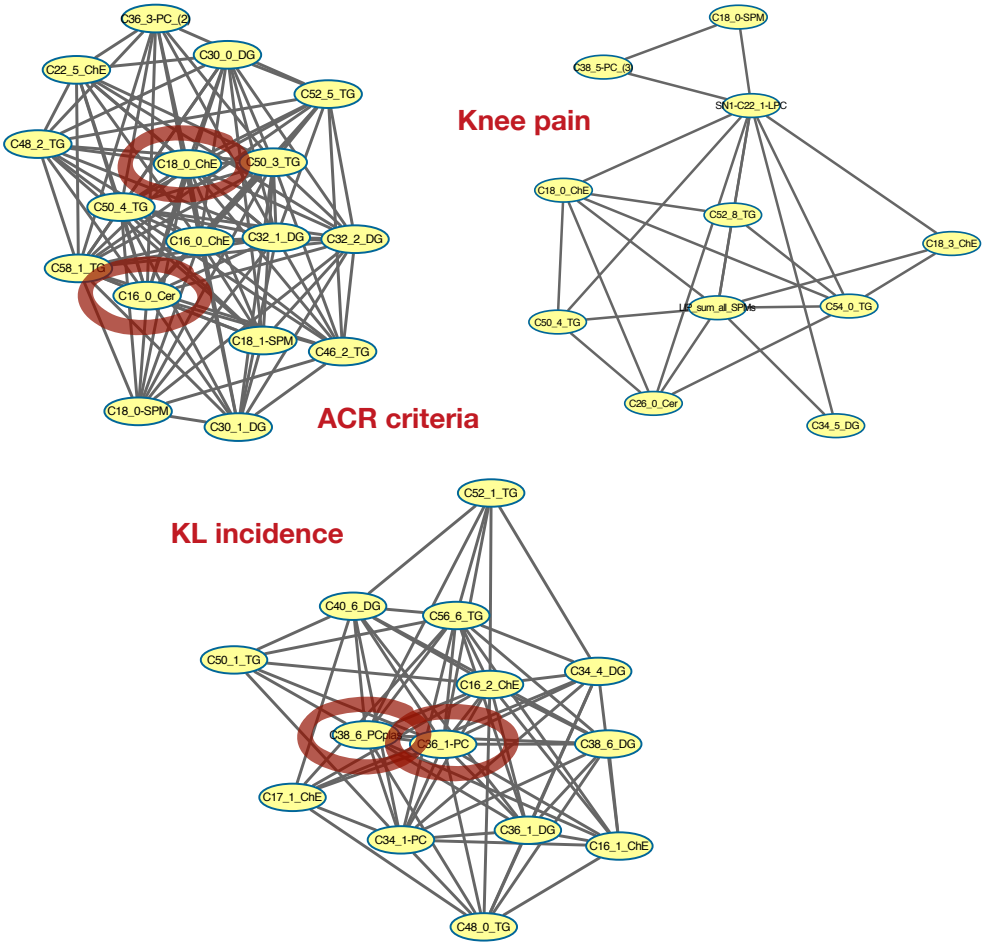


Fig 5.22: The main clusters identified by the MCODE algorithms within the FuNeL networks generated from the TNO lipidomics data. In red are highlighted the lipids selected by RGIFE.

- ACR criteria: 0.363
- Knee pain: 0.282
- K&L score: 0.463

The low values tell that RGIFE was not able to identify good patterns, within the data, that can be helpful when predicting the label (*OA* / *non-OA*) of unseen samples. Given the bad performances, it was necessary to verify that the low AUC values were not due to a poor behaviour of RGIFE. Thus, two other feature extraction algorithms were tested: SVM-RFE [111] and CFS [104]. As introduced in Chapter 4, SVM-RFE requires setting upfront the number of features to select. This value was set to 10 because this represents the largest model that the collaborators in the D-BOARD

projects aim to analyse. The AUCs values when using CFS were: 0.395, 0.456, 0.550 respectively for ACR criteria, knee pain and KL score incidence. Lower performance were obtained by SVM-RFE: 0.369, 0.165 and 0.431. Overall, even if the lipids selected by CFS seemed to predict slightly better the presence of the knee OA with the UNOTT data, the general values were still quite low and suggest a potential overfitting during the learning phase. Given that not only RGIFE but other machine learning approaches were unable to extract relevant biomarkers from the UNOTT data, no further analysis was performed (as presented for the TNO data). The rationale of this choice is that the signatures extracted from the UNOTT data would simply be relevant for the small (74) set of samples within the PROOF study, quite unlikely they would result powerful in determining the presence of the condition in other individuals from different cohorts.

5.3.2.6 Merging TNO and UNOTT data

The lipidomics analysis performed by the TNO and UNOTT were carried out on the same set of individuals. More importantly, both lipid analyses were performed using the plasma samples collected at the same data point. Therefore, it was interesting, especially from a machine learning point of view, to check whether by merging both sources of information, better predictive models could be identified. The resulting datasets contained a total of 329 lipids. RGIFE was applied using the same leave-one-out validation scheme employed in the previously proposed experiments. The summary of the TNO+UNOTT models is shown in Table 5.10.

The AUC-CV values for the ACR criteria and knee pain models are essentially the same that were obtained when considering only the TNO lipids. Furthermore, the biomarkers extracted using the TNO+UNOTT data are identical to the ones identified when studying only TNO data (see Table 5.7). This result further highlights the irrelevant information, in terms of knee OA presence prediction, encapsulated within the UNOTT lipidomics data. Moreover, the best performing RGIFE configurations were similar when extracted from the TNO and TNO+UNOTT data. For ACR criteria it slightly varied from $(depth = 8, cost = 2)$ to $(depth = Unlimited, cost = 2)$, while for the knee pain it changed from $(depth = 4, cost = 1)$ to $(depth = 2, cost = 2)$ confirming the robustness of RGIFE. On the other hand, a bit counter-intuitive, is

OA measure	Variables	AUC-Full	AUC-CV	p-value
ACR criteria	C16:0-Cer C24:2-Cer C18:0-ChE	0.927	0.760	1.38e-05
Knee pain	C22_3-2 C38:4-DG C40:6-DG C60:6-TG	0.863	0.450	0.0001
KL incidence	C18_3-w C40:4-PC-(2) C38:5-PCplas C36:6-DG C50:0-TG Resolvin-D1	0.954	0.447	0.0008

Table 5.10: Summary of the inferred models for each knee OA outcome measure when using TNO+UNOTT lipidomics data. The last column indicates the permutation test p-value.

visible a decrease in performance when using the K&L score incidence measure. The drop of AUC-CV by 0.12 can be probably due to an overfitting of the heuristic, the additional information brought by the UNOTT lipids are likely to have "destroyed" the patterns that characterised and were found by RGIFE within the TNO data. The K&L score incidence model presented in Table 5.10 shares only two lipids with the model generated from the TNO data. Out of the 6 biomarkers, only one belongs to the UNOTT data: *Resolvin-D1*. Nevertheless, it is interesting to observe how the family of this lipid can be relevant for the knee OA. In fact, other researchers have recently shown that targeting the *D-series resolvin* receptor system provides robust analgesics effects for the treatment of osteoarthritic pain [266].

5.4 Discussion

This chapter presented the analysis performed for the collaboration with the D-BOARD consortium. The work consisted in the definition of a generic pipeline for the identification, validation and evaluation of novel biomarkers for knee osteoarthritis (OA) in overweight women. The analysed data were derived from the PROOF study [250], a project that aimed to assess the effect of a diet-and-exercise program, com-

bined with the oral assumptions of crystalline glucosamine sulphate, on the incidence of knee osteoarthritis overweight women aged between 50 and 60 years of age. Data were collected at two different time points: after 2.5 and 6.5 years from the baseline (start of the study).

Out of all the subjects included in the PROOF study, 365 had a 2.5 years follow-up where the incidence (development) of knee OA was defined using five different outcome measures. Each individual was characterised by variables containing: clinical and imaging-based information, food routine, pain presence and biochemical markers concentration. Different than the traditional approaches, mostly based on the combination of univariate filtering and multivariate logistic regression, machine learning techniques were employed to identify knee OA-related biomarkers and generate predictive models. The multivariate machine learning approach, represented by the use of RGIFE, resulted in five small (at most 8 variables) highly predictive (AUC > 0.7) models. Overall, the best model was inferred with the K&L scale for the definition of the presence of (patello femoral) OA, probably one of the most adopted measure. The worst performance was obtained when using the data labelled with JSN (Joint Space Narrowing) criteria. Possibly, the bad results can be explained by the difficulty in assessing the JSN from the X-ray that can bring to errors when defining the presence of OA, thus provide misleading data from which is hard to learn.

All the inferred models contained variables covering most of the different categories of information collected for the PROOF study. In contrast to many reports about knee OA, the PROOF study includes, together with "classic" clinical, imaging-based and pain information, variables that describe the daily food routine of the subjects. Those variables appeared in almost all the models, suggesting an important role when trying to discriminate *incident* and *non-incident* samples. Fruit intake information were present in many models, in particular they were valuable for the K&L score incidence model where their removal caused a significant drop in AUC (almost 10%).

The variable direction analysis, also known as partial dependence analysis, allowed to assess the association between each biomarker and the OA incidence. From the outcome of this investigation some expected output emerged. Both BMI and waist circumference were positively related to the outcome, the higher the value assumed by

those variables, seen as a significant overweight, the greater the chance to develop the condition. Similar, a large intake of fruit, that can be interpreted as an overall healthier diet, help in decreasing the chance to develop knee OA. Conversely, less obvious was the negative association between the years since the menopause and the incidence of OA as well as the grinding sounds when moving the knee. The variable direction analysis was another example of how machine learning models contain relevant knowledge that can be exploited to gain new insights. In this case, the models were built under certain constraints (fixing the value assumed by each biomarker) and looking on how their structure, and consequently their predictions changed, it was assessed if the concentration of a specific biomarker might increase or decrease the chance of developing the condition.

The partial dependence analysis also helped in confirming the importance of certain "wet" biomarkers of extracellular matrix tissue turnover. Both ACR criteria, medial and lateral JSN based prediction models contained blood or urine-based biomarkers. C1M and C2M seemed to be negatively associated with the incidence OA defined by ACR criteria and medial JSN, in a similar way, COLL2-1NO2 showed a negative relationship with incidence OA defined by lateral JSN. Previous studies have found correlations between OA severity and C1M [255] and a difference in C2M levels between OA and healthy subjects [261]. Furthermore, the negative association between COLL2-1NO2 and OA incidence is in line with a previous study also performed in the PROOF cohort [244], thus providing some degree of validation. These findings suggest that assessment of structural degradation products from the extracellular matrix in body fluids may provide valuable information on the development OA and prediction of disease incidence in high-risk groups.

The analysis of the inferred models revealed that the knowledge associated with the imaging-based variables provides valuable information. Each model uses at least one variable related to imaging-based techniques. In addition, a positive association emerged between the presence of OA on MRI at baseline and the incidence of knee OA when using the K&L grade definition. Overall, this highlights a need in re-evaluating a proper use of imaging information in primary care settings, especially when treating subjects at risk for future knee OA development. Nevertheless, it needs to be

stressed that not all the imaging-based variables included in the final models are easy to obtain in primary care, such as the outcomes of statistical shape modelling ("Mode x" variables). Hopefully, the relevant role of MRI features for the prediction of knee OA incidence in the model might direct the design of new studies that focus on early detection or early treatment of knee OA among a high-risk group of overweight and obese women.

For two of the OA outcome measures (knee pain and K&L score incidence), comparable models were found in the literature. The performance of the models inferred with RGIFE resulted superior to all the models proposed as state-of-the-art. The number of biomarkers was lower or equal with the K&L score incidence while slightly higher for the chronic knee pain model. All the selected studies the same approach: univariate analysis followed by logistic regression. As already noticed in Chapter 4, RGIFE with its multivariate approach offers better predictive performance than the traditional univariate methods (Chi-Square in Chapter 4). Better results are likely to occur because simple univariate methods seem unable to capture the complex mechanism behind knee OA; by checking the association of each single factor with the outcome variable, interactions that might trigger the presence of the condition are missed. This recommends a need in pushing for the adoption of appropriate computational techniques when dealing with clinical, and more in general, biomedical studies. Traditional statistical methods, often due to their simple approaches, seem to discover only a reduced amount of knowledge and this might limit the finding of new research insights. The thorough validation performed in this dissertation confirms machine learning as a valid and powerful alternative that can lead to better scientific discoveries.

Some of the PROOF study individuals, 74 in total, were available for follow-up analysis after 6.5 years from the baseline. Using this data, result of a cross-sectional study, biomarkers can be detected to predict the existence of knee OA (at a specific time point rather than incidence as for the 2.5 years data). The 74 subjects were used for a lipidomics screening, the presence of knee OA was determined with three outcome measures: ACR criteria, chronic knee pain and K&L grade. Two different lipidomics screenings were performed by two D-BOARD partners: TNO and UNOTT. The application of RGIFE, and two other feature extraction methods, to the UNOTT data, did

not lead to good predictive models. The AUC of the selected biomarkers was below the random classification threshold of 0.5. The poor results, confirmed by different biomarker discovery methods, hints that the lipids targeted by UNOTT are not involved in the developing of knee OA. However, low AUC values could also be caused by the small amount of information available for each sample (only 32 lipids). When having both a limited number of features and samples, it is hard for machine learning algorithms to extract patterns that can generalise the phenomena described within the data. On the contrary, when using the TNO data, better predictive models could be generated by RGIFE. The best predictive lipids were selected using the ACR criteria. Better performance from the TNO data might due to a the different set of lipids target by the screening technology.

Given that both TNO and UNOTT data were generated from samples collected at the same time point, the information was merged into a single dataset and analysed. The biomarkers extracted with the ACR criteria and knee pain measures led to the same signatures inferred when using only the TNO data. Conversely, RGIFE selected different, and worse performing, biomarkers from the K&L score data. In this instance, the addition of the UNOTT data probably inhibited RGIFE from finding the same patterns extracted when dealing only with the TNO data. Overall, the study of TNO+UNOTT data has been two-fold. On one side, the same biomarkers emerged from the analysis of different (incremental) data, suggesting an important role of the selected lipids for knee OA among overweight women. On the other hand, it has proven the robustness of RGIFE and its capacity to identify the most predictive attributes even when irrelevant information (UNOTT lipids) are added. This robust behaviour can be the result of the multiple repetitions (10 times) of the cross-fold validation performed in each iteration of the heuristic. The repetitions assure that possible flukes, due to the data being split into training and test set (possibly accentuated when having only a few samples as in this case) are mitigated. This translate in a more robust attribute ranking that guides to a more accurate filtering of the variables.

Different than the 2.5 years, the lipidomics data were not affected by missing values and were only characterised by continuous values (lipids abundance). Therefore, (lipid) functional networks were generated using FuNeL. The classification rules inferred by

BioHEL, that is at the core of FuNeL, showed low performance when using UNOTT data, thus the networks were only created from the TNO data. Most of the lipids selected by RGIFE resulted being hubs of the FuNeL networks and part of the main cluster. Furthermore, the proposed biomarkers were at the core of the inferred networks, as they tended to be closer (shorter path between each other) than any other pair of nodes. From a machine learning perspective, this result showed that, although using different knowledge representations and providing the output in different forms, RGIFE and FuNeL, when analysing the same data, extract consistent information. In this thesis, only two types of knowledge representation were analysed (rule-based and random forest). However, the results hint that, regardless how it is presented, relevant information is always available within the models. Nevertheless, appropriate techniques are fundamental to exploit the structure of the models and maximise the knowledge extraction.

5.5 Future work

The biomarkers extracted from the 2.5 years data contain two types of variables: early signs (e.g. pain while jumping) and risk factors (e.g. BMI or waist circumference). The presented analysis was performed using the information collected for the PROOF study, all the available variables were included without distinction. In the future, a separate analysis could be performed considering only one type of variable. This would allow to generate either a set of markers for early OA or a predictive model that would provide a risk score for the development of knee OA within a few years. In addition, this would provide different models to be applied based on the available information for each tested subject.

In this dissertation, the evaluation of the lipidomics models has been limited to the study of the predictive performance and the partial dependence analysis. Similarly, the networks inferred from the TNO data were only used to study topological properties and analyse the relationships between the variables selected by RGIFE. In future, the developed models deserve to be studied more in details to assess the role of the selected lipids better, especially involving a close collaboration with the clinical expert of knee

OA. For example, the association between the lipid abundance and the presence of knee OA requires a further study. The non-linear associations ($\bigwedge$ and $\bigvee$ shapes) need to be better understood and evaluated so that the role of each lipid can be properly interpreted. There is a limitation in the biomedical and clinical evaluation that machine learning experts can provide. Moreover, the interpretation of lipidomics data, different than transcriptomics, is fairly new and not many tools are available to clinically and biologically evaluate the role of specific lipids. Given the limited availability of methods that can perform enrichment analysis from sets of lipids, a simple over-representation test (e.g. using an hyper-geometric statistical test) could be implemented to check whether the selected biomarkers share some biological characteristics.

Overall, all the inferred models performed quite well for the prediction of knee OA. The performance can be considered as "fair" (AUCs is between 0.70 - 0.80) and "good" (AUCs between 0.80-0.90). Compared to the models from the specialised literature, similar or better performance were observed. However, this validation process is limited and is necessary to evaluate the predictive power using an independent set of individuals (external validation). Given that the new validation data will unlikely include all the variables employed by the inferred models, the role of the decremental analysis is fundamental. Based on the new variables available, the performance of the best fitting sub-model could be easily extracted from the decremental analysis and compared with the AUCs obtained from the new samples.

Looking at the analysis performed in this chapter, it is quite straight forward to ask whether the merging of the 2.5 years and 6.5 years data might generate a dataset from which better models can be extracted. From a machine learning point of view, the union of different data into a single one is a common process that often leads to the generation of better models because more information become available for each sample. However, from a clinical point of view, the merging of information obtained at different time points generates meaningless data. In fact, models inferred from such data would not be usable on a daily basis as they would require information collected at different time points. Nevertheless, in the near future, the collaborators of the D-BOARD might be able to provide new data (obtained after 6.5 years from the baseline) containing the same set of variables available in the 2.5 years data. Those

data could be analysed on their own but also merged with the lipidomics information to determine if a better predictive performance could be achieved. The 6.5 years clinical data could also be used as validation of the current biomarker signatures to check whether their performance decrease if tested with measurements collected at a later time point. Furthermore, it would be worth checking if RGIFE would select the same biomarkers.

Summary

This chapter has presented the application of machine learning-based methodologies for the analysis of knee osteoarthritis data. While in Chapter 3 and Chapter 4, RGIFE and FuNeL have been mainly tested with transcriptomics data, in here they have demonstrated their ability in dealing with different types of biomedical data. The results revealed that, although presenting the extracted knowledge in different forms, FuNeL and RGIFE, when applied to the same data, unveil consistent information. Overall, the analysis proposed in this chapter have further confirmed the importance of exploiting the information encapsulated within machine learning models to gain new relevant biomedical knowledge.

6

Conclusions

Contents

6.1 Summary

In the last decade, thanks to the constant reduction of the bio-technologies costs, we have seen that biomedical data accumulates at an increasing speed. Appropriate computational approaches are necessary to make sense of this large abundance of information. In the biomedical field, decades-old statistical-based methods are still commonly used to analyse the data and extract meaningful knowledge. However, due to the design simplicity of these methods, the information that they can extract from biomedical data is often limited. Machine learning represents an important alternative to statistical-based methods, with a rich and versatile knowledge representation, different and more interesting patterns can be found in the data.

This thesis was focused on the analysis of biomedical data with machine learning methodologies. In particular, the presented work expanded and improved the typical use of machine learning in biomedical context. While in many applications the generation of (predictive) computational models represents the end point, in here it became the input for the process of knowledge discovery. The structure of the inferred models was mined to gain new insights about specific biomedical problems. The rationale is that by understanding how machine learning algorithms can solve analytical problems, we can learn how to better address biomedical tasks. In particular, the thesis tried to verify the following research hypothesis:

Research hypothesis

Can we extract relevant knowledge from the analysis of machine learning models generated from biomedical data?

This research hypothesis was tested focusing on two main biomedical analytical tasks: (1) the inference of biological networks and (2) the discovery of biomarkers. Both research topics require the ability to correctly identify and interpret complex interactions between different factors present in the data. The first problem focuses on the inference of functional associations between biological entities (genes, proteins, lipids, etc.), the second one demands the identification of sets of entities (biomarkers) that together can drive/influence a specific biomedical condition. Machine learning,

using complex knowledge representations, is particularly suited for both tasks. This dissertation presented different methods to mine machine learning models, generated from biomedical data, and solve both problems. In addition, given the large variety of biomedical data (from -omics to clinical data, from image to questionnaire information) that are continuously generated, a further aim of the thesis was the proposal of flexible methodologies able to deal with such an assortment of data and not tailored for a particular format.

6.2 Evaluation of the research question

The research question was tested with thorough analysis presented in Chapter 3, Chapter 4 and Chapter 5. Machine learning models were exploited, with different approaches, to generate functional networks and propose biomarkers from biomedical data.

In Chapter 3 the problem of the network inference was tackled presenting FuNeL. FuNeL is a protocol that identifies functional associations by mining the classification rules generated with BioHEL [39]. It employs the co-prediction inference paradigm where biological entities that participate in the same classification rules (generated to discriminate between different samples such as controls vs. cases) are hypothesised to be functionally related. Different than the commonly employed similarity-based methods, within machine learning models the entities are related not because they are similar (e.g. have similar expression profiles), but because together they detect strong patterns. The success in the test performed using synthetic data provided a first hint on the ability of FuNeL to identify relevant associations between biological entities. When tested with eight real-world datasets, FuNeL networks were shown to be complementary, in terms of biological characteristics, to the networks generated with three state-of-the-art similarity-based methods. In addition, the co-prediction was proven to better capture the concept of functional relationship when using gene-disease associations. In FuNeL networks, disease-associated genes were found more closely connected and present in functional units. The research hypothesis considered in the thesis was further evaluated using a prostate cancer dataset as a case study.

The biomedical relevance of the knowledge associated with the FuNeL networks was confirmed by (a) the specialised literature and (b) the analysis of an independent set of data.

Chapter 4 introduced significant improvements in RGIFE, a heuristic for the identification of small sets of highly predictive biomarkers. Originally presented in [208], RGIFE was extensively revamped to address two main drawbacks: large computational time and undesired local optimums. Different than in FuNeL, within RGIFE the machine learning models are exploited to guide the search for the optimal set of biomarkers. Based on an iterative feature elimination paradigm, RGIFE mines the structure of machine learning models, generated to solve a classification task, to define a feature ranking and remove the irrelevant ones. First, the newly introduced features (see Section 4.2.1 for details) were shown to improve the performance of the original RGIFE both in terms of computational time and number of selected biomarkers. When compared with well-known approaches used for biomarker discovery, RGIFE offered statistically similar predictive performance while constantly using fewer features. The use of a sophisticated multivariate knowledge representation (random forest) led to better performance when contrasted with a simple univariate method (i.e. Chi-Square). Furthermore, when applied to synthetic datasets, RGIFE showed the ability to identify relevant features among irrelevant and redundant information. The genes extracted by RGIFE from a prostate cancer dataset revealed higher relevance, in a disease context, to that of other methods for biomarker discovery. This chapter provided an answer to the research hypothesis showing that, not only the genes extracted by RGIFE were enriched for biological pathways known to be associated with prostate cancer (according to specialised literature), but also were highly genomically altered in other independent datasets (not used for the inference of the signature).

In Chapter 5 biomedical data were analysed trying to identify biomarkers for knee osteoarthritis (OA). The presented work was part of the collaboration with the D-BOARD consortium, that aims to discover new biomarkers to predict the presence of knee osteoarthritis in overweight women. RGIFE was used to generate small different highly predictive models (with less than 8 biomarkers). In addition, using the information extracted from the machine learning models, the identified biomarkers were

extensively characterised by studying their importance (additive value to the predictive model) and their association with the presence of the condition. The result of this analysis highlighted the importance of image-based biomarkers and, even more importantly, confirmed the relevance for the incidence prediction of knee OA, of well-known biochemical markers. Contrasted with the models available in the literature (generated with a statistical approach based on a univariate analysis followed by a multivariate logistic regression), the proposed models were shown better performing. This further confirmed the importance of using machine learning techniques when dealing with a complex condition such as OA. Simple statistical approaches are not powerful enough to provide valuable and strong solutions. Finally, the chapter illustrated the robustness of the methods proposed in this thesis. FuNeL was applied to a subset of data (lipidomics data) and most of the topologically relevant nodes (hubs) of the networks coincided with the variables (lipids) selected by RGIFE. These findings suggest that, although using a different knowledge representation and providing information in a diverse form, when applied to the same data the methods extract similar insights.

Overall, the thesis' research hypothesis was **validated** in the three research chapters. The results obtained clearly showed that machine learning models can be mined to infer **relevant** knowledge. Moreover, the use of machine learning unveils information that would be **totally missed** when using traditional statistical-based approaches. This thesis was focused on methods that generate biological networks and discover predictive biomarkers. However, the same "mining" approach can be used to address other challenging problems in the biomedical fields (protein structure prediction, phylogenetic tree construction, etc. see Figure 1.1).

6.3 Contribution to the area of bio-data mining

The main contributions of this dissertation to the area of biomedicine and bio-data mining are:

- development of the FuNeL protocol for the inference of functional networks from the analysis of rule-based machine learning models (in Chapter 3)

- demonstration that machine learning-based inference methods generate networks that are topologically different than co-expression and capture dissimilar and complementary biomedical knowledge. The difference between networks have also been assessed using a novel evaluation approach based on gene-disease associations (in Chapter 3)

- improvement of a machine learning-based heuristic, called RGIFE, to select small panels of highly relevant biomarkers. Demonstration that methods specifically designed to solve the problems of biomarker discovery perform better than more generic machine learning approaches (in Chapter 4)

- identification and characterisation of knee-osteoarthritis biomarkers, via the use of a machine learning-based methods, from different types of biomedical data. In addition, confirmation of the relevance of established biochemical markers (in Chapter 5).

Furthermore, it is important to highlight that the two main methods proposed in this thesis, FuNeL and RGIFE, are generic enough to be easily modified and used with other machine learning algorithms. The network inference stage of the FuNeL protocol (see Section 3.2.2) can use other (rule-based) machine learning classifiers while maintaining the other steps for the generation of different functional networks. Similarly, the RGIFE heuristic presented in Chapter 4, although tested only with a random forest, can be coupled with another classifier that provides a feature ranking. Moreover, despite the fact that the proposed approaches were tested mainly using transcriptomics data, other types of biomedical data can be employed. This was partially shown in Chapter 5 where both FuNeL and RGIFE were applied to lipidomics and clinical data. The only requirements, for both methods, is that data points are assigned to different categories and can be used for a classification problem.

6.4 Limitations

The work presented in this dissertation has some limitations that will be covered in the next sections.

6.4.1 Computational time

The methods proposed in this thesis, although able to generate meaningful outputs and discover relevant knowledge, have not been designed aiming for execution speed. FuNeL uses BioHEL at the core of its network inference, a classifier based on evolutionary learning whose execution is computationally expensive. The generation of a single co-prediction network (in terms of rule sets inference) on itself is relatively fast, but the permutation test, requiring many iterations, represents a time-consuming element. However, given the independence of each BioHEL run, FuNeL can be trivially parallelised to reduce its overall computational time (see Appendix A.4 for the complete analysis). Optimisation in the core rule learning process of BioHEL or the use of a faster classifier would also be a solution to tackle this limitation.

The improved version of RGIFE uses a random forest instead of BioHEL as proposed in its original form [208]. As expected, the choice of a faster classifier dramatically decreased the computational time required by the heuristic. This makes RGIFE now comparable with CFS in terms of speed, when dealing with large datasets. However, other methods such as ReliefF and SVM-RFE are still faster. Similar to FuNeL, RGIFE has an independent component in the execution of each iteration (based on $M \times n$-fold cross-validation), therefore its runs can also be parallelised. In addition, a speed-up in performance could be obtained using an even faster classifier (e.g. decision tree), only if this would not decrease the performance of the heuristic.

6.4.2 Co-prediction paradigm

The co-prediction paradigm defines functional associations between entities that are used in the same classification rule. If more than two attributes are present in the same rule, associations are inferred between all the possible pairs of attributes. When the rules contain many attributes, the co-prediction infers a large clique per each rule. This might lead to functional networks defined as large sets of connected cliques. Such a problem is partially mitigated by the use of the permutation test that removes spurious edges. In the analysis performed in Chapter 3, when BioHEL was applied to transcriptomics dataset, the number of attributes per rule was hardly more than

4. Thus, the mentioned problem of having networks defined by multiple cliques did not occur. However, if applied to other types of data that require complex rules (with many attributes) to perform a correct classification, the co-prediction approach might need to be adjusted. One solution could be to assess how often subsets of attributes (more than 2) appear together in other rules. An alternative could be the adoption of a pruning procedure similar to the one used by ARACNE (see Section 2.4) to remove spurious connections.

6.4.3 Data pre-processing

The transcriptomics data used for the analysis of FuNeL and RGIFE (Chapter 3 and Chapter 4) were found in public repositories, they were already pre-processed and "ready-to-use". On the other hand, the D-BOARD data (2.5 years) required the imputation of missing values. A limited preliminary analysis was performed to establish the effect of different strategies to solve this problematic, as it was out of the scope of this thesis. The missing values were imputed using the K-Means algorithm. Similarly, the class imbalance problem was addressed with the SPIDER oversampling algorithm. However, many other imputation methods are available [256] and different strategies can be used to tackle the imbalance problem [267]. Although it is unlikely that different pre-processing approaches can substantially alter the overall predictive performance of the models, other biomarker sets could emerge.

6.4.4 Lack of ground truth and field limitations

One of the main limitations in bio-data mining is the lack of established knowledge or a ground-truth that could be systematically and automatically used to assess the *correctness* of computational approaches. Without ground truth, the correctness is limited to special cases for which the expected output is known or can be related to established knowledge. A common alternative is the validation on synthetic data such as partly employed in Chapter 3 and Chapter 4. Unfortunately, there is a limit to how well synthetic data can represent the complex characteristics of biomedical phenomena. An example is given by the analysis of the SD datasets [218]. Although the RGIFE-Union policy was not able to identify the entire optimal subset of features, it

still obtained the highest accuracy when tested with two classifiers. Given this restrictions, the community should start putting more effort into the identification of new strategies for the evaluation of new proposed methods. For example, the analysis of biological networks should not be limited to the study of the main clusters. Additional validation, based on the analysis of the relationships between disease-associated genes could be performed (as described in Chapter 3 and Chapter 4). This would provide a better understanding and interpretation of novel network inference methods. Overall, although it is understandable that the proposed problems are difficult and challenging, the design of new validating solutions seems a necessary step on the way to refining most of the methods proposed in the fields of bio-data mining.

6.5 Future work

Different future research steps can be pursued from the work presented in this dissertation. In the next sections, some future research directions will be presented.

6.5.1 Integration of FuNeL and RGIFE

The dissertation was focused on the analysis and the problematics associated with two biomedical analytical tasks: the inference of functional networks and the discovery of biomarkers. This resulted in the presentation of two novel approaches namely FuNeL and RGIFE. Chapter 5 showed that the information extracted by the two methodologies, although presented in different forms, are mutually confirmed. Because FuNeL and RGIFE infer related knowledge, their union could lead to the discovery of more robust and relevant insights. A straightforward combination would use RGIFE instead of SVM-RFE in the feature selection step of FuNeL (see Figure 3.2). However, given that RGIFE tends to select a small number of features, the resulting networks could be very small and dense. If a larger network would be required, RGIFE could implement an "early stopping condition" so that more attributes could be used for the inference stage.

Another approach could use both knowledge representations during the inference of the functional associations. Specifically, the feature importance ranking (performed

by RGIFE using a random forest) could guide the generation of the classification rules (with BioHEL). In this instance, the most important attributes could be preferred to form the classification rules or could be less likely to be removed during the generalisation step of the genetic algorithm (the process of randomly remove attributes from the rules to avoid over-fitting). Alternatively, the FuNeL knowledge could guide the search performed by the RGIFE heuristic. The topological properties of the inferred network, such as node degree or centrality, could help in defining the importance ranking used by RGIFE. Key nodes, such as hubs and central nodes, would have lower chance to be removed during the iterative process.

6.5.2 Knowledge integration for a better learning

An emerging research path involves the integration of biological knowledge during the model generation process [114, 115, 268]. That is, the learning process is biased/guided by some established information (received externally). This prior knowledge can be expressed in multiple forms such as cellular pathways or biological and molecular networks. Using patterns present in external sources of information, it is easier to identify the artefacts in the data (e.g. spurious structures). The inference process can then focus on the features that are consistent with the established knowledge. For example, the relationship between two genes with common biological characteristics (e.g. GO terms) or involved in the same pathways, could be used directly in FuNeL or could be "preferred" during the rule learning process. Similarly, disease-associated genes could be used as the seed for the generation of the classification rules in BioHEL, that will be lately served to create co-prediction networks.

When it comes to RGIFE, an approach would be to weight more heavily the variables sharing many biological features, or associated with the same disease, to make their combined removal harder. In the Section 6.5.1 it has been suggested to use the (topological) information extracted from co-prediction networks to guide the RGIFE removal. Alternatively, the feature removal could be influenced by one (or the combination) of the many molecular networks available in the literature such as STRING, HIPPO, I2D, etc.

6.5.3 Exploring the role of different knowledge representations

Another unexplored avenue is the use of machine learning algorithms that adopt a different knowledge representation. Other rule-based classifiers can be placed at the core of FuNeL, the Apriori algorithm [54] is the simplest example. However, FuNeL is not exclusive of rule-based machine learning, and the co-prediction approach could be extended to other types of knowledge representation. The co-prediction paradigm defines associations within features that participate in the same classification model, thus features that "co-operate" in solving a classification problem. Therefore, associations could be extracted from a random forest linking the attributes present in the same tree. If using a single decision tree, relationships could be identified analysing the paths from the root to the leaves of the tree. Possibly, even linear models, if trained on different subsets of the data (to obtain multiple models covering different parts of the solution space), could be explored for the co-prediction principle. Finally, it would be interesting to assess how different types of knowledge representation used by BioHEL can affect the resulting FuNeL networks. BioHEL uses rules that define a hyper-rectangle in the feature space, however different kind of representations are suitable for the classifier (e.g. ratio of predicates). This research topic has started to be explored in [269] where the effects of four knowledge representations have been evaluated within FuNeL networks. The results show that different representations generate networks of varying sizes and with relatively low overlap between important nodes. On the other hand, the overlap between enriched terms of different networks is much higher. Overall, this work suggests the importance of not restricting the biological data analytic process to a reduced/specific type of knowledge representation, because it will only be able to provide partial knowledge. Nevertheless, further and more detailed (e.g. focused on a specific biomedical problem) analysis need to be performed to fully understand the effect of different knowledge representations. Similar remarks can be expressed for RGIFE. The heuristic is generic enough to be implemented with any classifier that estimates a feature importance. Many approaches could be a target for future analysis: from the traditional standard algorithms (e.g. decision trees or SVM) to the new largely embraced methods such as XGBoost [270] or deep neural networks.

6.5.4 Application to other fields

In this dissertation, it was extensively demonstrated how the proposed methods can be successfully utilised with various data from the biomedical domain. However, given the flexibility and the generality of the introduced methods, they could be applied to extract knowledge in different domains. For example, the process of drug discovery and repositioning is currently attracting a lot of interest. One of the recent DREAM challenges, in which I participated [16], asked the researchers to predict the individual variability in cytotoxic response based on genomic and transcriptional profiles. If the input data are categorised according to the treatment (e.g. placebo vs. drug), one could use RGIFE to identify the most affected factors (e.g. genes or lipids). Additionally, a process of drug repositioning could be seeded from functional associations between biological entities (e.g. genes). Mainly, a drug targeting a gene could also be used to target other genes functionally associated with it.

Biomedicine is not the only field in which is important to identify driving factors from large-scale datasets. With the advent of the *big data* era, many other research areas require computational approaches to make sense of the collected data. For example, in the field of economics, Hal R. Varian used different machine learning methods to examine a dataset of 72 countries to see which variables were significant predictors for the economic growth [271]. Using RGIFE, a similar approach could be successfully applied to the huge amount of data produced by the stock market, trying to discover new predictors for the market trends. Experts in economics might also be interested in the identification of new associations between predictive variables. By generating FuNeL networks from economics datasets one might discover, perhaps without much surprise, that the *GDP level* is linked with the *life expectancy*. On the other hand, unexpected relationships might be identified and provide new hints to the field experts.

This section contains only few example, however, overall, I believe that the proposed approaches, but in general the analysis of machine learning models, have the potential to extract interesting knowledge from various types of data and contribute to the generation of new research hypothesis in a vast variety of fields.

A

Appendix

A.1 Enrichment score analysis

In this section are reported the network average rankings, based on the Enrichment Score, across the 8 datasets for every inferring method. The networks are ranked between 1 and N (where $N = 4$ for FuNeL and $N = 8$ for Pearson, ARACNE and MIC: 4 $SE(C_i)$ + 4 $SN(C_i)$). For this analysis were considered Gene Ontology terms (biological process (BP), molecular function (MF) and cellular component (CC)) and biological pathways. The last row of each table represents the average rank across different biological categories, in bold are highlighted the best performing networks.

Cat.	**C1**	**C2**	**C3**	**C4**
GO BP	3	4	2	**1**
GO MF	4	2.5	**1**	2.5
GO CC	2	4	**1**	3
Pathways	4	2	3	**1**
Average	3.25 ± 0.96	3.125 ± 1.03	**1.75 ± 0.96**	1.88 ± 1.03

Table A.1: Average ranks based on ES across the 8 datasets for the networks generated with FuNeL.

	Pearson (SE)				**Pearson (SN)**			
Cat.	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4
GO BP	2	4	**1**	3	6	7	5	8
GO MF	8	3.5	5	6	7	3.5	**1**	2
GO CC	2	5	4	6	**1**	8	3	7
Pathways	6	5	4	3	7	**1**	8	2
Average	4.5 ± 3.00	4.38 ± 0.75	**3.5 ± 1.73**	4.5 ± 1.73	5.25 ± 2.87	4.88 ± 3.22	4.25 ± 2.99	4.75 ± 3.20

Table A.2: Average ranks based on ES across the 8 datasets for the networks generated with Pearson.

	ARACNE (SE)				**ARACNE (SN)**			
Cat.	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4
GO BP	4	7	5.5	8	2	5.5	**1**	3
= **GO MF**	4	8	4	7	2	6	**1**	4
GO CC	3	7	5	8	2	6	**1**	4
Pathways	**1**	4	2	6	7	5	8	3
Average	3.00 ± 1.41	6.50 ± 1.73	4.13 ± 1.55	7.25 ± 0.96	3.25 ± 2.50	5.63 ± 0.48	**2.75 ± 3.50**	3.50 ± 0.58

Table A.3: Average ranks based on ES across the 8 datasets for the networks generated with ARACNE.

	MIC (SE)				MIC (SN)			
Cat.	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4
GO BP	2	6	4	5	3	8	**1**	7
GO MF	**1**	6	4	7	3	8	2	5
GO CC	3	5.5	4	5.5	2	8	**1**	7
Pathways	2.5	4	**1**	5.5	8	5.5	7	2.5
Average	**2.13 ± 0.85**	5.38 ± 0.95	3.25 ± 1.50	5.75 ± 0.87	4.00 ± 2.71	7.38 ± 1.25	2.75 ± 2.87	5.38 ± 2.14

Table A.4: Average ranks based on ES across the 8 datasets for the networks generated with MIC.

A.2 Disease association analysis

In this section are presented the network average rankings across the 8 datasets for every inferring method based on the gene-disease association (G-D) properties: participation in triangular relationship and proximity. Two sources of disease associations were used: Malacards [139] (a meta-database of human maladies consolidated from 64 independent sources) and manually curated databases (OMIM [143], Orphanet [146], Uniprot [145] and CTD [144]). The networks are ranked between 1 and N (where $N = 4$ for FuNeL and $N = 8$ for Pearson, ARACNE and MIC: $4\ SE(C_i) + 4\ SN(C_i)$). The number of disease-associated genes participating in a triangle is denoted as 1A, 2A and 3A. The last row of each table represents the average rank across different metrics, in bold are highlighted the best performing networks.

Malacards

Cat.	**C1**	**C2**	**C3**	**C4**
1A	3	**1**	4	2
2A	4	**1**	2	3
3A	3	3	**1**	3
Proximity	3	**1**	4	2
Average	3.25 ± 0.50	**1.50 ± 1.00**	2.75 ± 1.50	2.50 ± 1.49

Table A.5: Average ranks based on Malacards G-D associations across the 8 datasets for the networks generated with FuNeL.

	Pearson (SE)				Pearson (SN)			
Cat.	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4
1A	8	5	6	3	4	**1**	2	7
2A	7	4	5	3	8	2	6	**1**
3A	6.5	6.5	6.5	6.5	3	2	4	**1**
Proximity	**1.5**	5	3	**1.5**	7	6	8	4
Average	5.75 ± 2.90	5.13 ± 1.03	5.13 ± 1.55	3.50 ± 2.12	5.50 ± 2.38	**2.75 ± 2.22**	5.00 ± 2.58	3.25 ± 2.87

Table A.6: Average ranks based on Malacards G-D associations across the 8 datasets for the networks generated with Pearson.

	ARACNE (SE)				ARACNE (SN)			
Cat.	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4
1A	4	8	7	5.5	2.5	2.5	5.5	**1**
2A	7	6	8	**1**	4	3	2	5
3A	5.5	**1**	5.5	2	5.5	5.5	5.5	5.5
Proximity	7	3	5	**1**	6	2	8	4
Average	5.88 ± 1.44	4.50 ± 3.11	6.38 ± 1.38	**2.38 ± 2.14**	4.50 ± 1.58	3.25 ± 1.55	5.25 ± 2.47	3.88 ± 2.02

Table A.7: Average ranks based on Malacards G-D associations across the 8 datasets for the networks generated with ARACNE.

	MIC (SE)				MIC (SN)			
Cat.	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4
1A	6	3	7.5	**1**	5	2	7.5	4
2A	8	3	5	2	7	**1**	6	4
3A	7	2	8	6	5	**1**	3	4
Proximity	6	**1**	2	5	7	4	8	3
Average	6.75 ± 0.96	2.25 ± 0.96	5.63 ± 2.75	3.50 ± 2.38	6.00 ± 1.15	2.00 ± 1.41	6.13 ± 2.25	3.75 ± 0.50

Table A.8: Average ranks based on Malacards G-D associations across the 8 datasets for the networks generated with MIC.

Curated databases

Cat.	C1	C2	C3	C4
1A	3	**1**	4	2
2A	3	4	**1**	2
3A	**1**	2.5	2.5	4
Proximity	4	**1**	3	2
Average	2.75 ± 1.26	**2.13 ± 1.44**	2.67 ± 1.25	2.50 ± 1.00

Table A.9: Average ranks based on curated G-D associations across the 8 datasets for the networks generated with FuNeL.

	Pearson (SE)				Pearson (SN)			
Cat.	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4
1A	8	4	6.5	**1.5**	6.5	**1.5**	3	5
2A	2	7	5.5	3	4	5.5	**1**	8
3A	5	7	4	8	**1**	6	3	2
Proximity	4	8	2.5	7	5	2.5	**1**	6
Average	4.50 ± 2.50	6.50 ± 1.73	4.63 ± 1.75	4.88 ± 3.12	5.13 ± 2.32	3.88 ± 2.21	**2.00 ± 1.15**	5.25 ± 2.50

Table A.10: Average ranks based on curated G-D associations across the 8 datasets for the networks generated with Pearson.

	ARACNE (SE)				ARACNE (SN)			
Cat.	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4
1A	3	5	8	6	2	4	**1**	7
2A	5	**1**	2	3	7.5	6	7.5	4
3A	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5
Proximity	6	5	4	**1**	7.5	3	7.5	2
Average	4.63 ± 1.25	3.88 ± 1.93	4.63 ± 2.50	**3.63 ± 2.14**	5.38 ± 2.66	4.38 ± 1.25	5.13 ± 3.09	4.38 ± 2.06

Table A.11: Average ranks based on curated G-D associations across the 8 datasets for the networks generated with ARACNE.

	MIC (SE)				MIC (SN)			
Cat.	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4
1A	6	2	4	3	8	**1**	7	5
2A	7	3	8	**1.5**	5.5	**1.5**	5.5	4
3A	4	**1**	8	2	6	5	7	3
Proximity	5.5	**1**	3	4	7	2	8	5.5
Average	5.63 ± 1.25	**1.75 ± 0.96**	5.75 ± 2.63	2.63 ± 1.11	6.63 ± 1.11	2.38 ± 1.80	6.88 ± 1.03	4.38 ± 1.11

Table A.12: Average ranks based on curated G-D associations across the 8 datasets for the networks generated with MIC.

A.3 Case study: prostate cancer dataset

In this section are reported the additional results from the analysis performed using the prostate dataset [170] as a case study that were not included in the main sections of the thesis. In particular is shown: (1) the unique enriched terms of co-prediction and co-expression networks, (2) the overlap between GO terms associated to the hubs of networks generated with different methods and (3) the average percentages of alteration for key nodes of both co-prediction and co-expression networks in an independent dataset.

A.3.1 Overlap of networks enriched terms

In Figure A.1 and A.2 are shown the non-overlapping terms between FuNeL networks and, respectively, ARACNE and MIC networks. For the sake of readability the generic GO terms (with depth < 9 in the GO hierarchical structure) are filtered out.

When comparing ARACNE and FuNeL, 16 unique pathways were found for co-prediction networks and 8 for co-expression. In terms of unique GO terms, the overlap was more balanced, 7 for co-prediction networks and 9 for co-expression networks. C_2 and C_4, generated without feature selection, had the largest number of unique pathways, while $SE(C_2)$ was related with the highest number of terms for ARACNE. The comparison of FuNeL networks with MIC co-expression generated many empty columns for the GO terms because several networks resulted having no unique enriched terms. All the unique GO terms (15) associated to MIC were related to $SN(C_2)$ (and with $SN(C_4)$ in two cases), while FuNeL had more networks having paradigm-specific terms (12). As noticed for in the ARACNE comparison, FuNeL networks are more enriched in biological pathways: 16 against 8 unique terms for MIC co-expression.

Fig A.1: Unique enriched GO terms (biological process) for each network configuration (generated from the prostate cancer dataset). The x-axis shows the 12 investigated networks. The y-axis shows the names of enriched terms unique to co-prediction or ARACNE co-expression networks. Red terms are associated with co-expression networks, blue with co-prediction. Empty columns indicate networks with no unique terms.

Fig A.2: Unique enriched GO terms (biological process) for each network configuration (generated from the prostate cancer dataset). The x-axis shows the 12 investigated networks. The y-axis shows the names of enriched terms unique to co-prediction or MIC co-expression networks. Red terms are associated with co-expression networks, blue with co-prediction. Empty columns indicate networks with no unique terms.

A.3.2 Genomic alteration in independent dataset

In this section are included additional information from the analysis of an independent prostate cancer study [170] available in the cBioPortal for Cancer Genomics [195]. In particular are reported the full list of alterations for the topologically important genes analysed in the Section 3.3.7. The Figures A.3–A.10 show the percentage of altered tumour samples for top 10 hubs (nodes with highest degree) and top 10 central nodes (with highest betweenness centrality) in the best performing networks according to the gene-disease association analysis (using the information from the curated databases). The selected networks are C_2 for FuNeL, $SN(C_3)$ for Pearson, $SE(C_4)$ for ARACNE and $SE(C_2)$ for MIC. For all of them the alterations of both hubs and central nodes are shown.

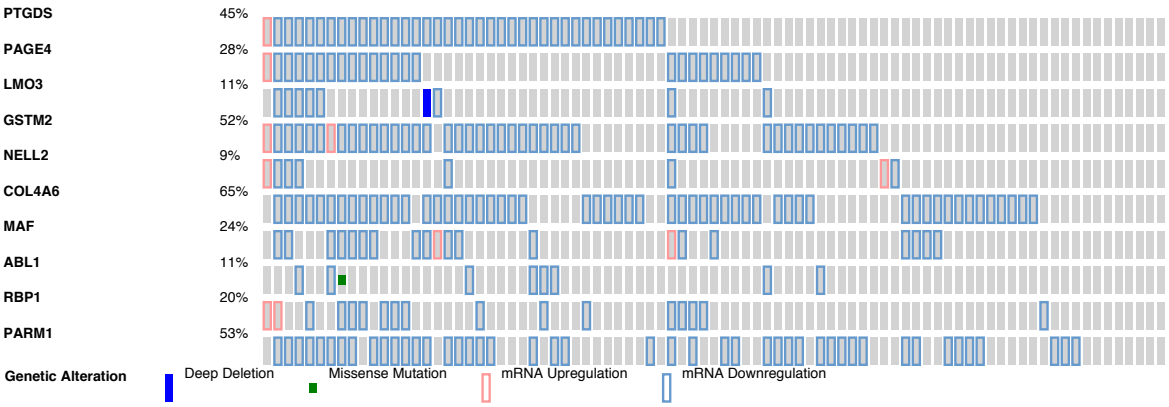


Fig A.3: Percentage of alterations in tumour samples from an independent cancer genomic study. Genes with highest degree (hubs) in C_2 network are shown.

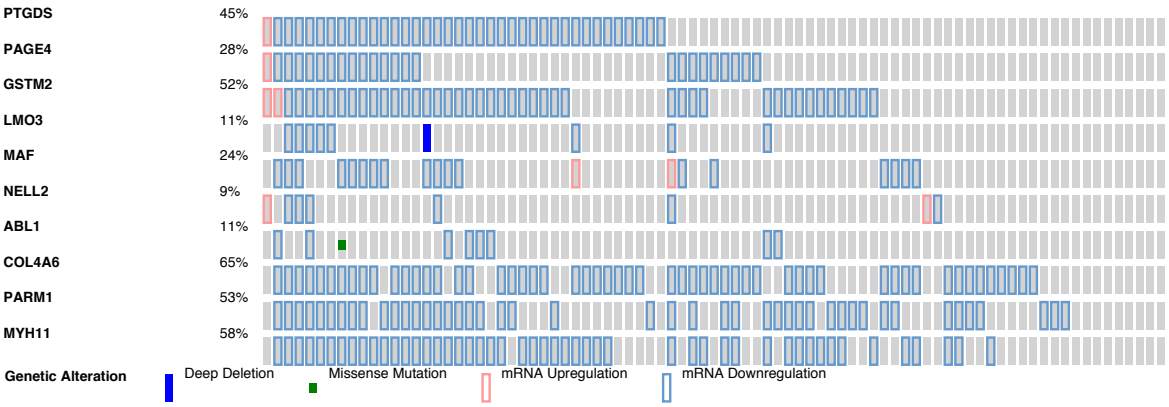


Fig A.4: Percentage of alterations in tumour samples from an independent cancer genomic study. Genes with highest betweenness centrality (central nodes) in C_2 network are shown.

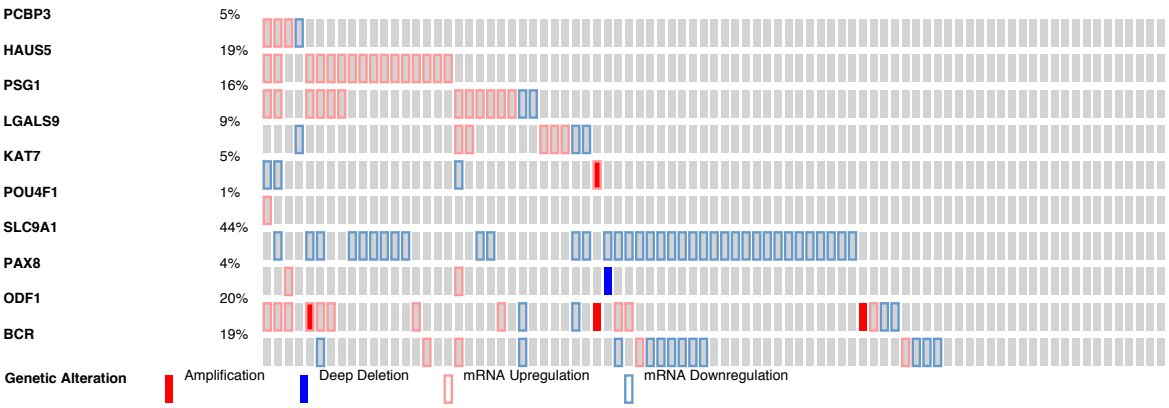


Fig A.5: Percentage of alterations in tumour samples from an independent cancer genomic study. Genes with highest degree (hubs) in Pearson $SN(C_3)$ network are shown.

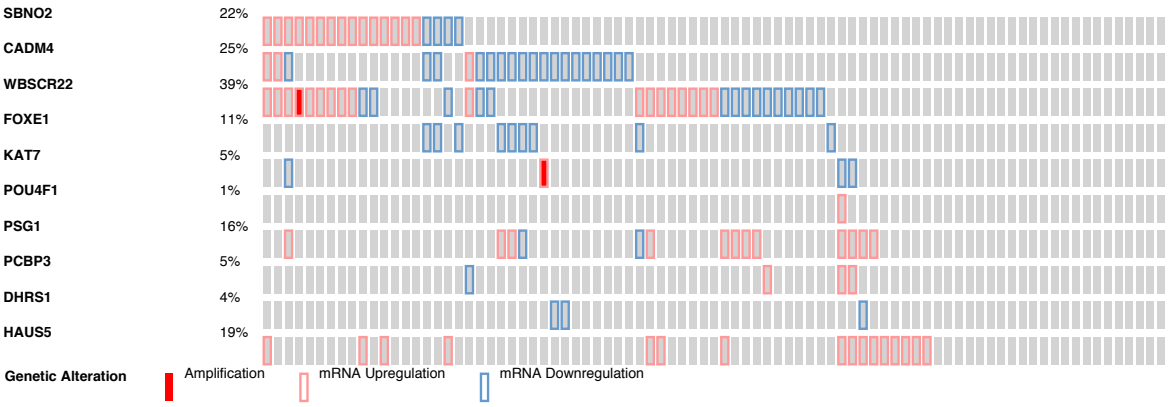


Fig A.6: Percentage of alterations in tumour samples from an independent cancer genomic study. Genes with highest betweenness centrality (central nodes) in Pearson $SN(C_3)$ network are shown.

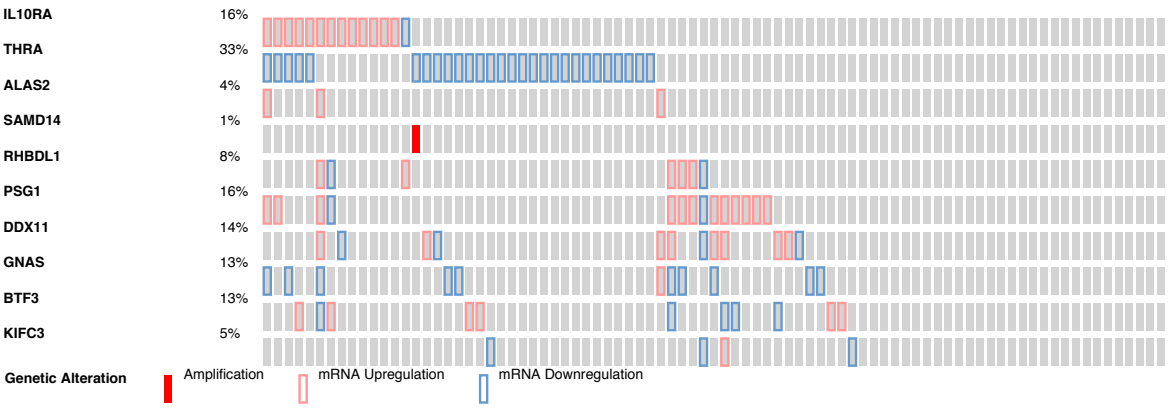


Fig A.7: Percentage of alterations in tumour samples from an independent cancer genomic study. Genes with highest degree (hubs) in ARACNE $SE(C_4)$ network are shown.

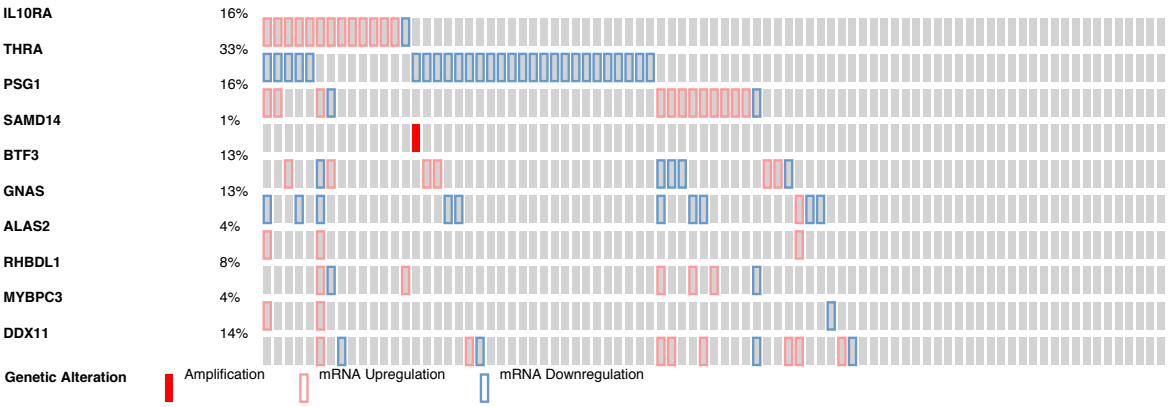


Fig A.8: Percentage of alterations in tumour samples from an independent cancer genomic study. Genes with highest betweenness centrality (central nodes) in ARACNE $SE(C_4)$ network are shown.

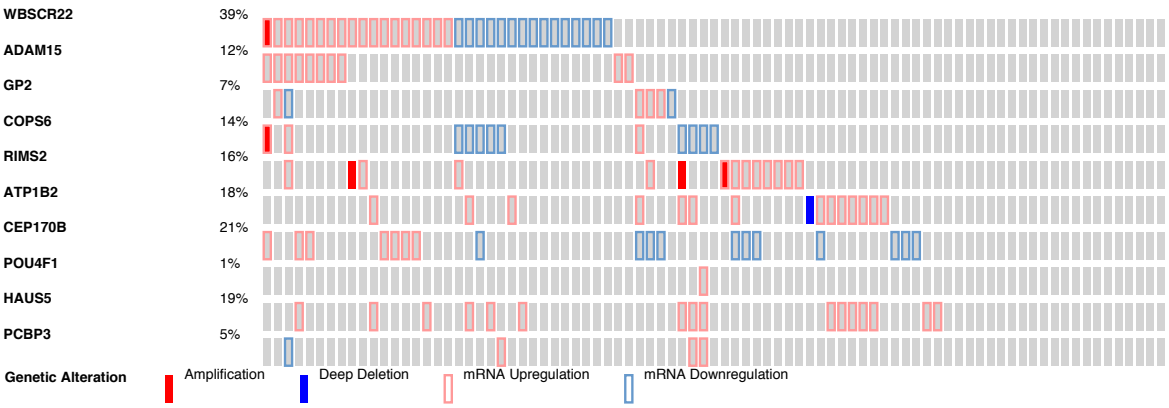


Fig A.9: Percentage of alterations in tumour samples from an independent cancer genomic study. Genes with highest degree (hubs) in MIC $SE(C_2)$ network are shown.

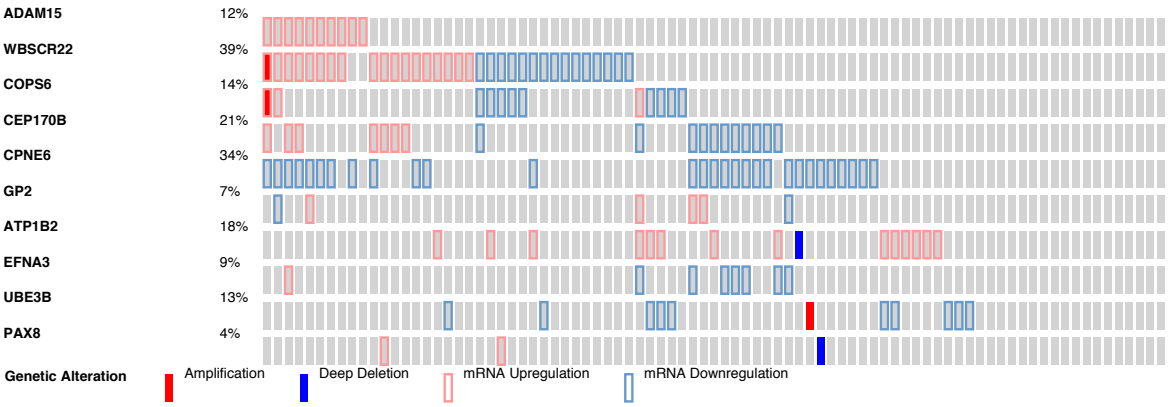


Fig A.10: Percentage of alterations in tumour samples from an independent cancer genomic study. Genes with highest betweenness centrality (central nodes) in MIC $SE(C_2)$ network are shown.

A.4 Time complexity analysis

The FuNeL protocol has four stages (see Figure 3.2): (1) feature selection (optional), (2) rule-based network generation, (3) permutation test and (4) second rule-based network generation (optional).

The running time for the whole pipeline depends on the rule set generation time (execution time of BioHEL), as the optional feature selection stage can be seen as running in constant time. Two main factors that influence the rule set generation time are: (1) the number of attributes and (2) the number of samples.

An execution time analysis of BioHEL was performed using the largest (in terms of number of attributes) Colon-Breast dataset [172]. In the feature selection stage were retained: 20, 200, 2000, 10 000 and 20 000 attributes. From each of these 5 datasets 100 random subsets of 50, 40, 30, 20 and 10 samples were generated. Finally, BioHEL was executed 1000 times to obtain 1000 rule sets for each dataset. Figure A.11 shows the running times averaged across 100 000 runs (1000 runs for each of the 100 datasets).

The total execution time of FuNeL configurations C_1 and C_2 is calculated as:

$$T_1 = (rule\_sets \times t(atts_1, samples)) + (permutation\_runs \times t(atts_1, samples))$$

where $rule\_sets$ is the number of inferred rule sets, $permutation\_runs$ is the number of randomised datasets used in the permutation test and $t(atts_1, samples)$ represents

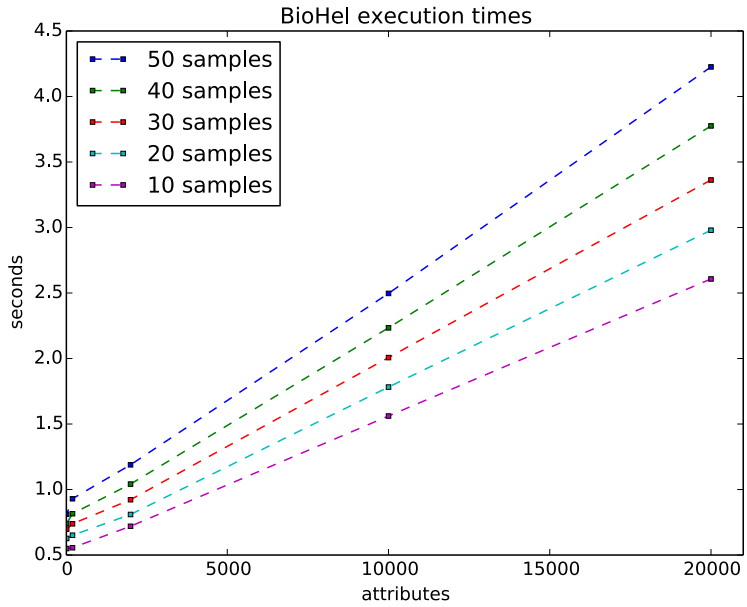


Fig A.11: Average execution times of a single BioHEL run for a given number of samples and attributes.

execution time of a single BioHEL run, that linearly depends on the size of a dataset measured in number of attributes and samples.

Configurations C_3 and C_4 require an additional run of BioHEL (step 4), and their total execution time is:

$$T_2 = T_1 + (rule\_sets \times t(atts_2, samples))$$

where $atts_2$ is the number of attributes after the permutation test ($atts_1 \leq atts_2$).

It is important to notice that each run of BioHEL is independent, thus the generation of the rule sets can be trivially parallelised without any extra overhead. Given n computational cores, the total execution times could be reduced to:

$$T_{real_1} = \frac{T_1}{n} \qquad T_{real_2} = \frac{T_2}{n}$$

B

Appendix

B.1 Predictive performance with synthetic datasets

The predictive performance of the attributes selected by each method were tested using different synthetic datasets. Table B.1 shows the accuracies, obtained from a 10-fold cross-validation, using the datasets described in [103]. N/A is used for SVM-RFE when tested with the *Monk3* dataset because the method cannot deal with categorical attributes. In Table B.2 are reported the accuracies calculated from the analysis of the *madsim* data [220]. Each row includes the average values associated to the analysis of datasets having 1%, 2% and 5% of up/down regulated attributes (genes).

Class.	Dataset	RGIFE-Min	RGIFE-Max	RGIFE-Union	CFS	Relief	SVM-RFE	Chi-Square	L1
RF	**CorrAL**	0.675	0.725	0.758	0.675	0.758	**0.783**	0.658	0.733
	XOR-100	**1.000**	**1.000**	**1.000**	0.500	0.480	0.500	0.420	0.580
	Parity3+3	**1.000**	**1.000**	**1.000**	0.521	0.933	0.429	0.474	0.502
	Monk3	**0.935**	**0.935**	**0.935**	**0.935**	0.910	N/A	**0.935**	**0.935**
	Madelon	0.869	0.868	**0.874**	0.805	0.866	0.787	0.835	0.744
	SD1	0.240	0.319	0.333	0.414	0.452	**0.478**	0.437	0.421
	SD2	0.389	**0.639**	0.635	0.521	0.456	0.466	0.477	0.458
	SD3	0.317	**0.626**	**0.626**	0.428	0.476	0.473	0.487	0.526
SVM	**CorrAL**	0.633	0.625	0.658	0.608	0.642	**0.725**	0.600	0.658
	XOR-100	0.598	0.700	**0.707**	0.500	0.400	0.480	0.500	0.360
	Parity3+3	0.348	0.348	0.348	**0.550**	0.319	0.502	0.500	0.505
	Monk3	0.828	0.828	0.828	0.813	0.820	N/A	**0.837**	0.789
	Madelon	0.598	**0.600**	**0.600**	0.557	**0.600**	0.593	0.595	0.562
	SD1	0.238	0.293	0.281	**0.437**	0.386	0.376	0.369	0.398
	SD2	0.371	0.349	0.351	0.395	0.626	0.459	0.473	0.473
	SD3	0.306	0.358	0.393	0.353	0.469	0.461	0.492	**0.515**
KNN	**CorrAL**	0.575	0.600	0.625	**0.758**	0.733	**0.758**	0.625	0.608
	XOR-100	**0.987**	0.962	0.973	0.560	0.460	0.460	0.500	0.520
	Parity3+3	0.219	0.219	0.219	0.550	**0.936**	0.486	0.560	0.543
	Monk3	0.887	0.887	0.887	**0.902**	0.894	N/A	0.877	0.878
	Madelon	0.698	0.694	0.699	0.868	**0.913**	0.828	0.894	0.805
	SD1	0.292	0.350	0.352	0.423	**0.453**	0.442	0.414	0.374
	SD2	0.436	0.393	0.419	0.421	**0.487**	0.470	0.476	0.446
	SD3	0.352	0.375	0.441	0.462	0.510	0.545	0.520	**0.546**
GNB	**CorrAL**	0.608	0.600	0.633	0.650	0.708	**0.717**	0.600	0.683
	XOR-100	0.602	0.689	**0.691**	0.480	0.420	0.480	0.480	0.420
	Parity3+3	**1.000**	**1.000**	**1.000**	0.567	0.233	0.486	0.500	0.488
	Monk3	**0.894**	**0.894**	**0.894**	0.887	0.887	N/A	**0.894**	0.887
	Madelon	0.698	0.694	0.699	0.699	**0.703**	0.688	0.699	0.675
	SD1	0.21	0.278	0.249	0.437	0.463	**0.477**	0.411	0.382
	SD2	0.283	**0.666**	**0.666**	0.451	0.533	0.458	0.443	0.474
	SD3	0.293	**0.667**	**0.667**	0.346	0.494	0.473	0.499	0.498

Table B.1: Accuracies obtained by each method across the synthetic datasets using four classifiers. The highest accuracies are shown in bold. RF: random Forest, KNN: K-nearest neighbour, GNB: Gaussian Naive Bayes.

Class.	Attributes	RGIFE-Min	RGIFE-Max	RGIFE-Union	CFS	Relief	SVM-RFE	Chi-Square	L1
RF	**5 000**	0.997	0.997	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	10 000	0.977	0.980	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	20 000	0.993	0.993	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	40 000	0.983	0.993	**1.000**	**1.000**	**1.000**	**1.000**	0.997	**1.000**
SVM	**5 000**	0.990	0.987	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	10 000	0.983	0.987	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	20 000	0.990	0.990	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	40 000	0.987	0.993	**1.000**	**1.000**	**1.000**	**1.000**	0.997	**1.000**
KNN	**5 000**	0.997	0.997	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	10 000	0.977	0.987	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	20 000	0.987	0.990	**1.000**	**1.000**	**1.000**	**1.000**	0.997	**1.000**
	40 000	0.997	0.987	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
GNB	**5 000**	0.997	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	10 000	0.983	0.993	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	20 000	0.990	0.993	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	40 000	0.987	0.997	**1.000**	**1.000**	**1.000**	**1.000**	0.997	**1.000**

Table B.2: Accuracies obtained by each method across the madsim datasets using four classifiers. The highest accuracies are shown in bold. RF: random Forest, KNN: K-nearest neighbour, GNB: Gaussian Naive Bayes.

B.2 Signatures analysed in the case study

In here are reported the signatures (list of genes) extracted by each method when analysing the Prostate-Singh [170] dataset within the case study. SVM-RFE, Chi Square and Relief were set to select as many genes as extracted by RGIFE-Union (21).

- **RGIFE-Min**: EPB41L3, HPN, HSPD1, PTGDS, NELL2, TGFB3, GSTM2
- **RGIFE-Max**: TNN, KCNN4, CELSR1, KIAA1109, PEX3, HPN, MFN2, ATP6V1E1, HSPD1, PTGDS, SLC9A7, NELL2
- **RGIFE-Union**: ANXA2P3, TGFB3, CRYAB, NELL2, MFN2, TNN, KIAA1109, PEX3, ATP6V1E1, HPN, HSPD1, LMO3, PTGDS, SLC9A7, SERPINF1, KCNN4, EPB41L3, CELSR1, GSTM2, EPCAM, ERG
- **SVM-RFE**: HPN, HSPD1, MAF, S100A4, JUNB, SERPINB5, C7, TBC1D2B, SDC1, IPO5, SFRP1, PGCP, PEX3, SPTB, FOXO1, GSTA4, CD38, RBBP6, SERINC5, VCAN, C5orf13
- **Relief**: HPN, HSPD1, NBL1, MAF, DPYSL2, C7, PEX3, TGFB3, CFD, TARP, PAGE4, XBP1, PTGDS, PDLIM5, RBP1, LMO3, SERPINF1, DPT, FAM107A, SERINC5, TACSTD2

- **Chi-Square**: HPN, NELL2, HSPD1, RBP1, PTGDS, CALM1, CDKN1C, PDLIM5, CFD, SERPINF1, TARP, COX7A1, GSTM2, CRYAB, RPLP0, TGFB3, ANGPT1, EPCAM, VCL, TMSB15A, LMO3

- **CFS**: RPL13, RPLP0, HBB, RPL6, HOXB3, TSPAN2, MCF2L2, PHEX, CNKSR2, CPA3, PLA2G7, SCGN, COL13A1, CHD9, EPB41L3, MEIS2, CREB3L1, ZFP161, ADORA2A, GLCE, SLC35A2, DDHD2, WIF1, HEPH, TMSB15A, DIXDC1, KIAA0427, PEX3, ZNF146, TRIM23, HPN, PITX1, SLC1A1, PENK, RBP1, C14orf2, TUBB2A, MAP1LC3B, CALCOCO2, CYP1B1, SLC25A6, ORAI2, GSTA4, AHR, SERPINF1, COBLL1, STK38L, SLC7A5, MRPL40, DST, JUNB, GSTP1, LGALS1, SPTAN1, ABI1, SPON1, ROCK2, AKR1B1, TSC22D3, GPM6A, PLAGL1, PLA2G2A, CKS1B, PDLIM5, HSPD1, LMO3, S100A4, PKD2, PTGDS, CDKN1C, CRMP1, CFD, CALR, NELL2, RGS10, ABL1, SERINC5, PMS2L5, MAPK10, GTF2B, RGN, ERG, SERPINB5, NAP1L3, LAMB1, GSTM2, IL11RA, CYP21A2

- **L1-based**: AVPR1B, TGM2, TSC22D3, ACTG1, ACTG2, MYH11, LYPLA2, BGN, HBB, SBF1, B2M, PRB1, MROH5, IGKC, CLSTN1, MYL9, ST5, GRK6, GADD45B, LYZ, PTGER3, ANXA2P3, PTP4A3, EDN2, ZNF337, MSMB, IFITM3, P4HB, SLC25A6, IFI30, ATP1B1, KLK2, KLK3, RPL10, RPL13, CYP3A5, COX6A1, RPL19, LOC91316, ORM1, NME2, CCND1, SFI1, SFN, NPY, UBB, MAF, ACTB, ACTA2, GRIN2C, RPL8, RPL9, HLA-C, PABPC1, RPL5, GAPDH, SEPT9, TUBB4B, NDRG1, PAGE4, RPS2P5, C21orf2, UBE3B, NBL1, ZFP36, MT1H, C4A, TACSTD2, MT1G, C1QL1, NACA, TPT1, FOS, VCL, UBC, IGL@, IGFBP5, COX7A1, FTO, LGALS3BP, PMP22, ALDH4A1, SDC1, KRT17, KRT15, KRT13, FLNA, LUZP1, CCL2, RPLP1, RPLP0, RPL18A, RPS6, RPS3, TXNIP, RPS17, LUM, TMED2, RPL6, TPM1, RPL13A, FASN, RPL7, CST3, DUSP1, TNFRSF6B, MARCKSL1, RPS24, ZFP36L1, ZFP36L2, TOP3B, PLA2G2A, LTF, S100A4, RPS4X, CLU, LRP3, HDGF, ACPP, RPSA, C7, GSTM2, ID1, CTGF, HSP90AA1, PSCA, COX7C, RPL36A, RBM3, RPS14, TMSB4X, EEF1A1, JUNB, JUND, TARP, ATP11A, PTGDS, XBP1, HLA-DRA, SERPINA3, RPL29, CEBPD, HSPD1, LDHA, AMD1, GALNS, PDIA2, IGH@, AAK1, ARR3, HPN, AP2A2, IGHM, VAMP1,

SORD, E2F4, HAP1, C1QTNF3, CFD, RPL32, MAP3K11, GSTP1, TSPAN1, PTRF, SYN1, EEF2

B.3 Time complexity analysis

The time complexity of each feature extraction method was tested across ten different datasets. The time, measured in second, required to identify the optimal subset of features was calculated for each method presented in the Chapter 4. Figure B.1 shows the running times averaged across the experiments performed for the 10-fold cross-validation. When plotting the times required by RGIFE, for each fold was calculated the average time obtained by three executions of the heuristic.

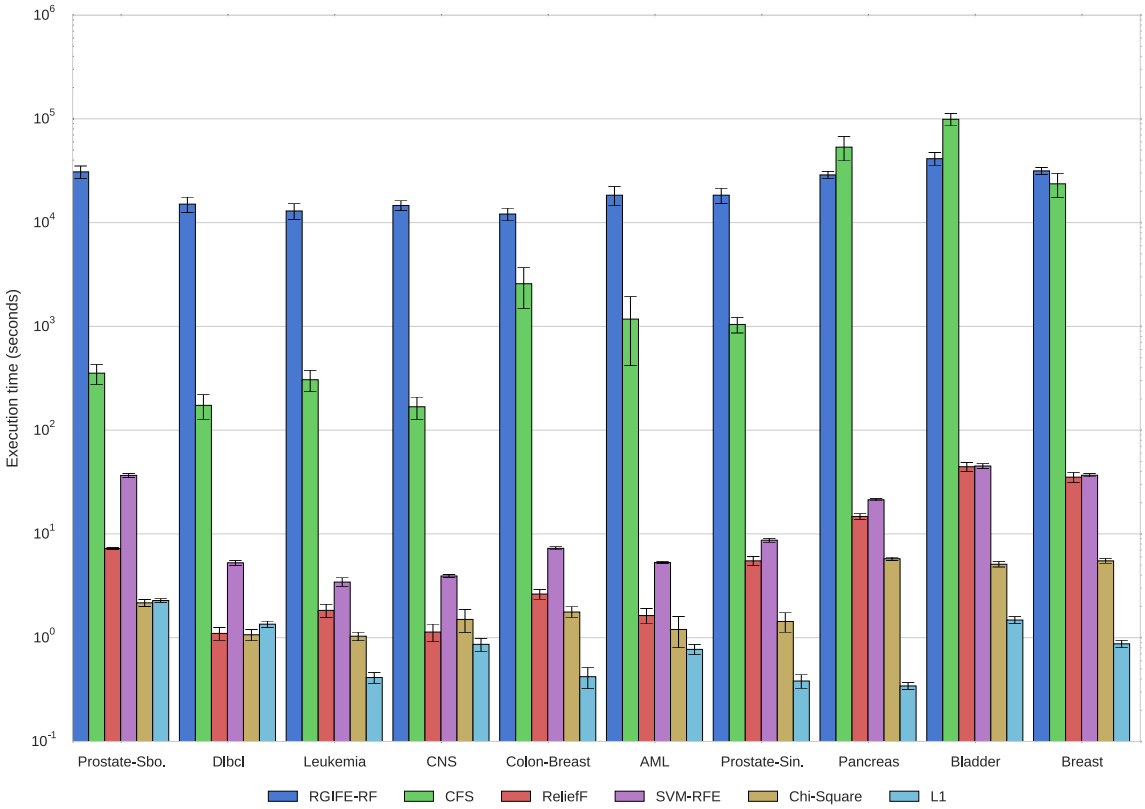


Fig B.1: Average execution times (calculated using a 10-fold cross-validation) of each methods across different datasets. The datasets are sorted by increasing number of attributes.

Overall, the methods more time consuming are CFS and RGIFE, they performed similarly with large datasets (in Figure B.1 the datasets are ranked by increasing

number of total attributes), while RGIFE was clearly slower for smaller dataset. The other four methods in general required less computational time with the L1-based approach that appeared to be the fastest one.

C

Appendix

C.1 PROOF study information

Clinical information
Age
BMI
Ethnicity; Western vs. others
Total cholesterol
Fat percentage
Waist circumference
Concentration of HbA1c
Physical activity level
Quality of life
Menopause status
Years since menopause
Varus laxity left knee in extension
Varus laxity right knee in extension
Valgus laxity left knee in extension
Valgus laxity right knee in extension
Varus laxity left knee in 20° flexion
Varus laxity right knee in 20° flexion
Valgus laxity left knee in 20° flexion
Valgus laxity right knee in 20° flexion
Ateriorpostrior hypermobility left knee
Ateriorpostrior hypermobility right knee
Hypermobility left knee
Hypermobility right knee
Randomized groups to diet and exercise intervention
Randomized groups to placebo-controlled glucosamine intervention
OA information
Presence of Heberden's nodes in one or both hands

KL grade ≥ 1 in one or both knees
Varus malalignment (<178°) in one or both knees
Presence of knee injury in one or both knees
Presence of knee pain in one or both knees
Clinical and radiographic ACR-criteria in one or both knees
WOMAC pain score
WOMAC function score
WOMAC stiffness score
Selfreported osteoarthritis in other joints
KL grade ≥ 2 in one or both knees
Maximal isometric quadriceps strength
Crepitus on passive or active knee flexion
Pain upon palpation of joint margin
Pain upon palpation of the patellar margin
Number of affected subregions with bone marrow lesions
Number of affected subregions with cartilage defects
Number of affected subregions with osteophytes
Presence of bone marrow lesions (yes vs. no)
Presence of cartilage defects (yes vs. no)
Presence of meniscal abnormalities (yes vs. no)
Presence of osteophytes grade 2 (yes vs. no)
Biochemical markers
Fibulin3-1 concentration
Fibulin3-1 log(concentration)
Fibulin3-1 Zlog(concentration)
Fibulin3-2 concentration
Fibulin3-2 log(concentration)
Fibulin3-2 Zlog(concentration)
Fibulin3-3 concentration
Fibulin3-3 log(concentration)
Fibulin3-3 Zlog(concentration)

C1M concentration
Log(C1M) concentration
Zlog(C1M) concentration
C2M concentration
Log(C2M) concentration
Zlog(C2M) concentration
Concentration Coll2-1NO2 adj. for creatinine
Zlog(Concentration) Coll2-1NO2 adj. for creatinine
Imaging-based information
Mode 0 (Active shape modelling)
Mode 1 (Active shape modelling)
Mode 2 (Active shape modelling)
Mode 3 (Active shape modelling)
Mode 4 (Active shape modelling)
Mode 5 (Active shape modelling)
Mode 6 (Active shape modelling)
Mode 7 (Active shape modelling)
Mode 8 (Active shape modelling)
Mode 9 (Active shape modelling)
Mode 10 (Active shape modelling)
Mode 11 (Active shape modelling)
Mode 12 (Active shape modelling)
Mode 13 (Active shape modelling)
Mode 14 (Active shape modelling)
Mode 15 (Active shape modelling)
Mode 15 (Active shape modelling)
Presence of OA on MRI
Presence of tibiofemoral OA on MRI
Presence of patellofemoral OA on MRI
Pain questionnaire

Mode 0 (Active shape modelling)
Mode 1 (Active shape modelling)
Mode 2 (Active shape modelling)
Mode 3 (Active shape modelling)
Mode 4 (Active shape modelling)
Mode 5 (Active shape modelling)
Mode 6 (Active shape modelling)
Mode 7 (Active shape modelling)
Mode 8 (Active shape modelling)
Mode 9 (Active shape modelling)
Mode 10 (Active shape modelling)
Mode 11 (Active shape modelling)
Mode 12 (Active shape modelling)
Mode 13 (Active shape modelling)
Mode 14 (Active shape modelling)
Mode 15 (Active shape modelling)
Mode 15 (Active shape modelling)
Presence of OA on MRI
Presence of tibiofemoral OA on MRI
Presence of patellofemoral OA on MRI
Food questionnaire
Number of days with breakfast
Number of days with lunch
Number of days with dinner
Frequency of milk/ buttermilk per week
Frequency of chocolate per week
Frequency of yogurt per week
Frequency of custard/pudding per week
Number of bread slices/crackers per week
Number of bread slices/crackers with cheese per week
Number of bread slices/crackers with meat per week

Number of bread slices/crackers with salad spread per week

Number of bread slices/crackers with chocolate per week

Number of bread slices/crackers with sweet filling per week

Frequency of cooked vegetables per week

Unities (50 grams) of cooked vegetables per week

Frequency of raw vegetables per week

Unities (50 grams) of raw vegetables per week

Frequency of fruit juice per week

Glasses of fruit juice per week

Frequency of tangering glasses per week

Pieces of tangering glasses per week

Frequency of citrus pieces per week

Pieces of citrus pieces per week

Frequency of apples/pears per week

Pieces of apples/pears per week

Frequency of bananas per week

Pieces of bananas per week

Frequency of other fruits per week

Pieces of other fruits per week

Frequency of apple sauce per week

Tablespoons of apple sauce per week

Pieces of fruit per day

Frequency of snack per week

Frequency of peanuts/nuts per week

Frequency of cheese per week

Frequency of pastry/cake per week

Frequency of candy bars per week

Frequency of chocolate per week

Frequency of biscuits (raisins) per week

Frequency of biscuits (others) per week

Frequency of coffee cups per day

Coffee with milk/sugar
Frequency of tea cups per day
Tea with milk/sugar
Frequency of soda glasses per day
Soda with/without sugar
Week days drinking alchool
Number of alchool glassess per week day
Weekend days drinking alchool
Number of alchool glassess per weekend day
Frequency of more than 4 drinks per drinking time

Table C.1: Complete list of attributes available from the PROOF data study

C.2 Lipidomics functional networks

This section shows the FuNeL networks generated from the 6.5 years lipidomics data. The networks have been generated using a "modified" version of the FuNeL protocol presented in Section 3.2.2 (configuration C_2). The permutation test was applied using the edge score (number of times two lipids appear together in the same rule) rather than the node score (number of time a lipid is used in a classification rules), more details can be found in Section 5.2.3.5

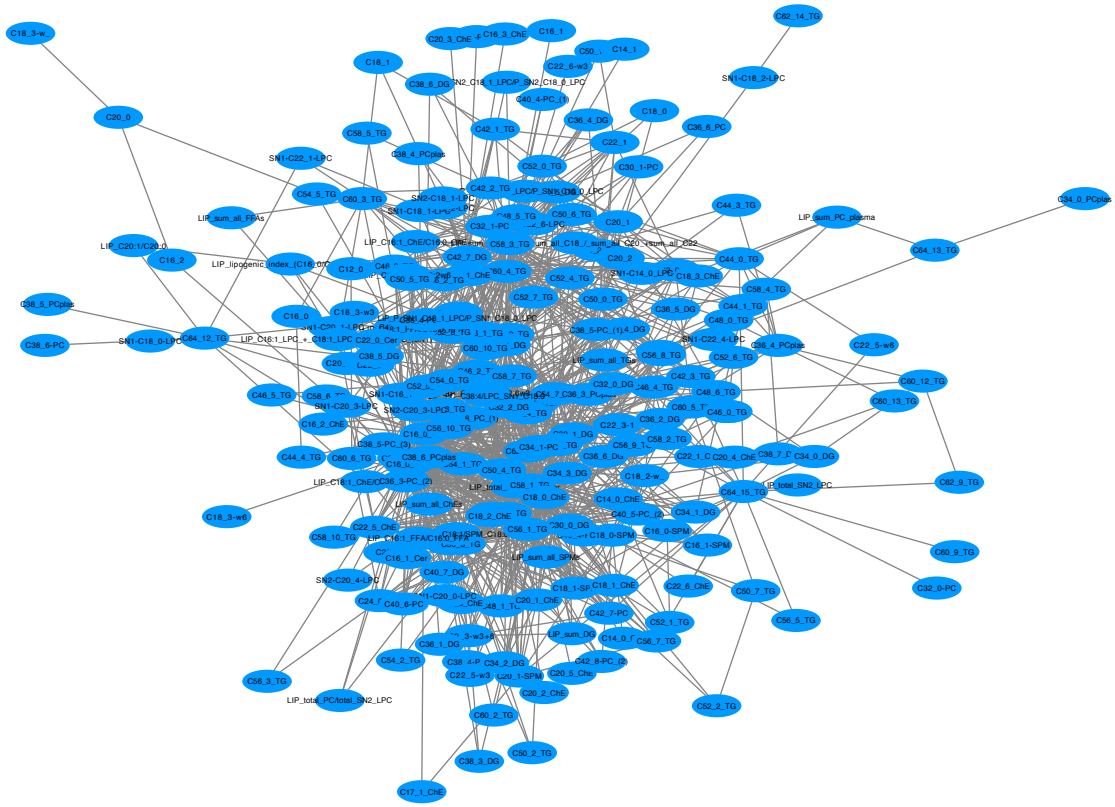


Fig C.1: FuNeL network generated using the 6.5 years lipidomics data using the ACR criteria definition for the incidence of knee OA.

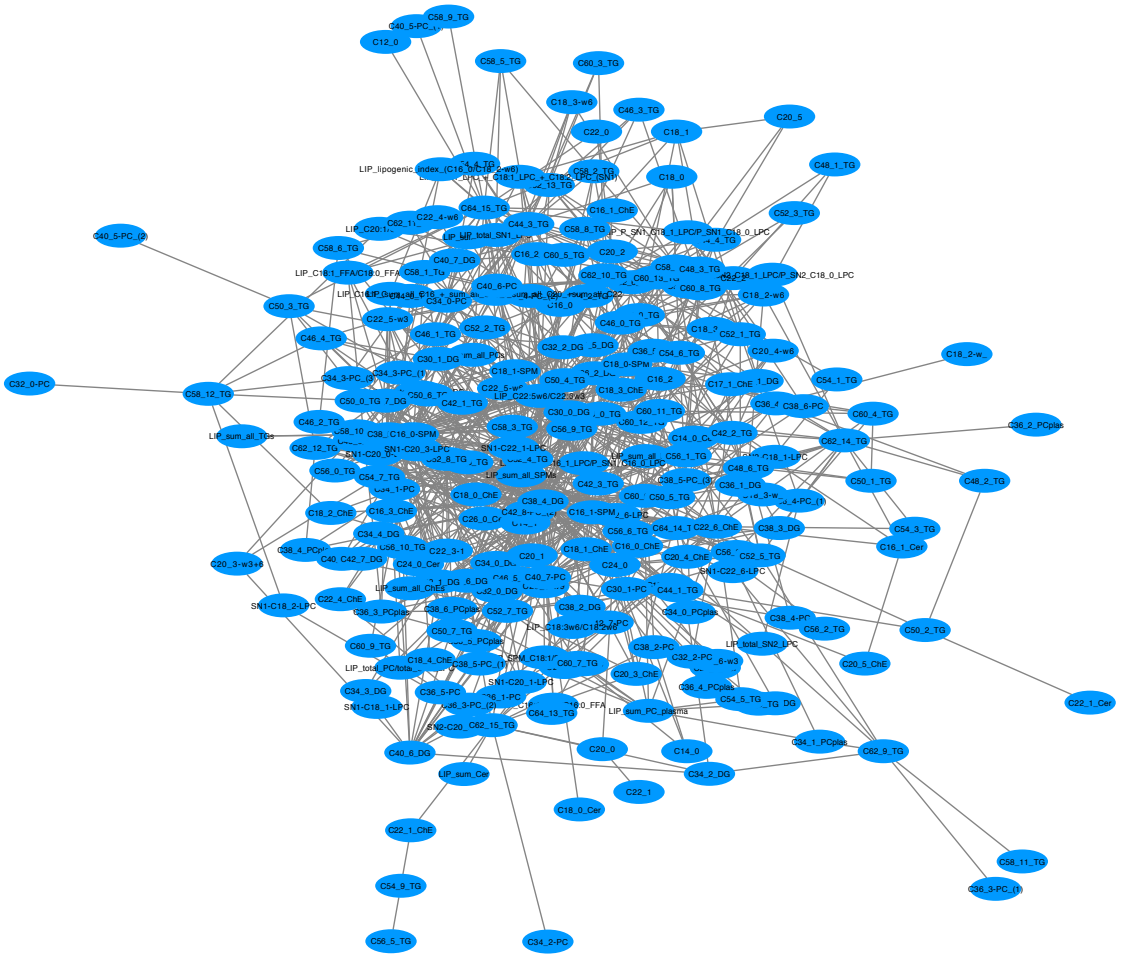


Fig C.2: FuNeL network generated using the 6.5 years lipidomics data using the Knee pain criteria definition for the incidence of knee OA.

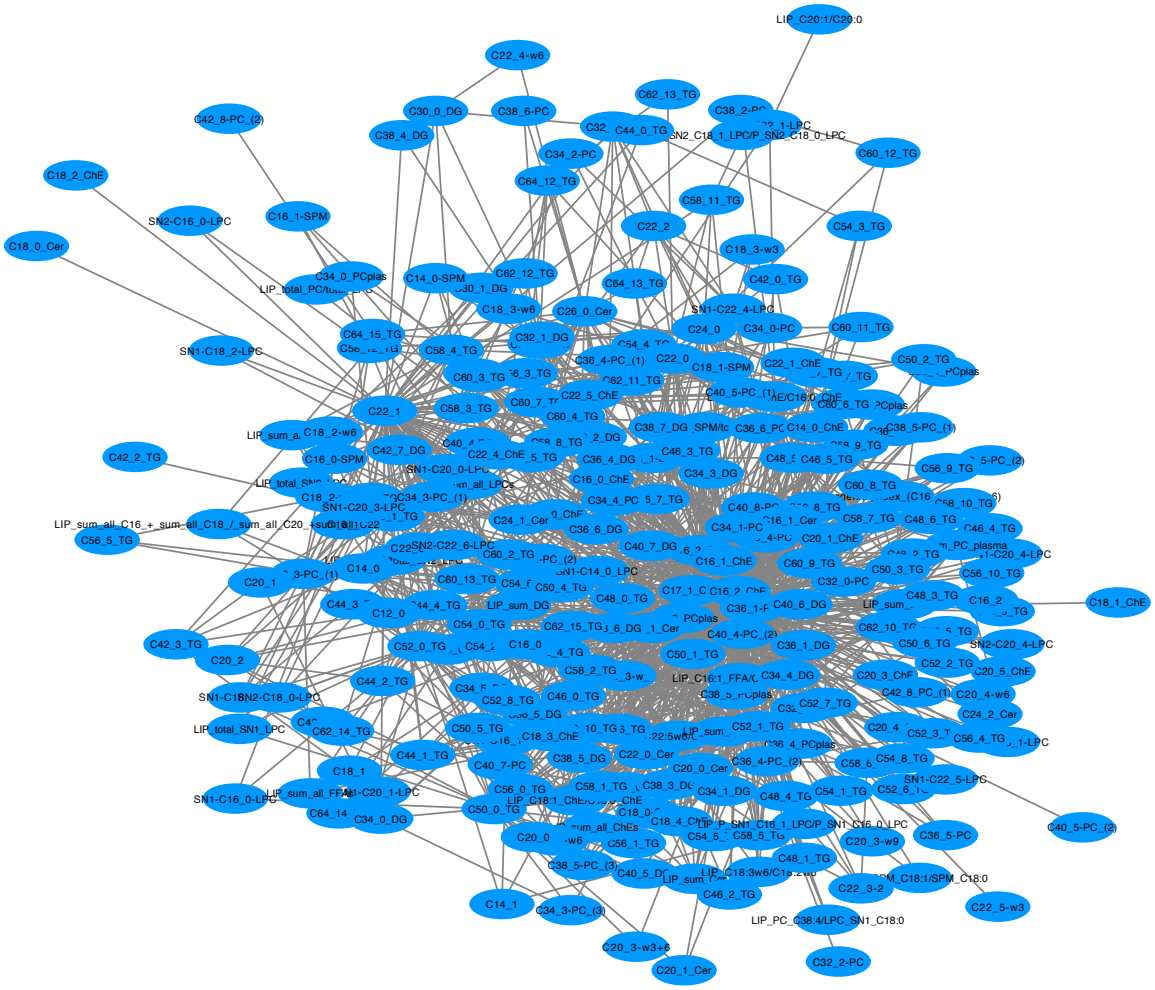


Fig C.3: FuNeL network generated using the 6.5 years lipidomics data using the KL score for the incidence of knee OA.

Bibliography

[1] F. Michor, Y. Iwasa, H. Rajagopalan, C. Lengauer, and M. A. Nowak, "Linear model of colon cancer initiation," *Cell cycle*, vol. 3, no. 3, pp. 356–360, 2004.

[2] R. Nussinov, "Advancements and challenges in computational biology," *PLoS computational biology*, vol. 11, no. 1, p. e1004053, 2015.

[3] J. D. Malley, K. G. Malley, and S. Pajevic, *Statistical learning for biomedical data*. Cambridge University Press, 2011.

[4] J. Gillis and P. Pavlidis, ""guilt by association" is the exception rather than the rule in gene networks," *PLoS Comput Biol*, vol. 8, no. 3, p. e1002444, 2012.

[5] K. R. Foster, R. Koprowski, and J. D. Skufca, "Machine learning, medical diagnosis, and biomedical engineering research-commentary," *Biomedical engineering online*, vol. 13, no. 1, p. 94, 2014.

[6] R. C. Deo, "Machine learning in medicine," *Circulation*, vol. 132, no. 20, pp. 1920–1930, 2015.

[7] A. L. Tarca, V. J. Carey, X.-w. Chen, R. Romero, and S. Drăghici, "Machine learning and its applications to biology," *PLoS Comput Biol*, vol. 3, no. 6, p. e116, 2007.

[8] J. B. Mitchell, "Machine learning methods in chemoinformatics," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 4, no. 5, pp. 468–481, 2014.

[9] T. M. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, Inc., 1 ed., 1997.

[10] I. Inza, B. Calvo, R. Armañanzas, E. Bengoetxea, P. Larrañaga, and J. A. Lozano, "Machine learning: an indispensable tool in bioinformatics," *Bioinformatics methods in clinical research*, pp. 25–48, 2010.

[11] A.-L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature reviews genetics*, vol. 5, no. 2, pp. 101–113, 2004.

[12] J. E. McDermott, J. Wang, H. Mitchell, B.-J. Webb-Robertson, R. Hafen, J. Ramey, and K. D. Rodland, "Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data," *Expert opinion on medical diagnostics*, vol. 7, no. 1, pp. 37–51, 2013.

[13] G. W. Beadle and E. L. Tatum, "Genetic control of biochemical reactions in neurospora," *Proceedings of the National Academy of Sciences*, vol. 27, pp. 499–506, Nov 1941.

[14] J. Bacardit, P. Widera, N. Lazzarini, and N. Krasnogor, "Hard data analytics problems make for better data analysis algorithms: bioinformatics as an example," *Big data*, vol. 2, no. 3, pp. 164–176, 2014.

[15] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and structural biotechnology journal*, vol. 13, pp. 8–17, 2015.

[16] F. Eduati, L. M. Mangravite, T. Wang, H. Tang, J. C. Bare, R. Huang, T. Norman, M. Kellen, M. P. Menden, J. Yang, X. Zhan, R. Zhong, G. Xiao, M. Xia, N. Abdo, O. Kosyk, F. Eduati, L. M. Mangravite, J. C. Bare, T. Norman, M. Kellen, M. P. Menden, S. Friend, G. Stolovitzky, A. Dearry, R. R. Tice, R. Huang, M. Xia, A. Simeonov, N. Abdo, O. Kosyk, I. Rusyn, F. A. Wright, T. Wang, H. Tang, X. Zhan, J. Yang, R. Zhong, G. Xiao, Y. Xie, H. Tang, J. Yang, T. Wang, G. Xiao, Y. Xie, S. Alaimo, A. Amadoz, M. A. ud din, C.-A. Azencott, J. Bacardit, P. Barron, E. Bernard, A. Beyer, S. Bin, A. van BÃűmmel, K. Borgwardt, A. M. Brys, B. Caffrey, J. Chang, J. Chang, E. G. Christodoulou, M. Clément-Ziza, T. Cohen, M. Cowherd, S. Demeyer, J. Dopazo, J. D. Elhard, A. O. Falcao, A. Ferro, D. A. Friedenberg, R. Giugno, Y. Gong, J. W. Gorospe, C. A. Granville, D. Grimm, M. Heinig, R. D. Hernansaiz, S. Hochreiter, L.-C. Huang, M. Huska, Y. Jiao, G. Klambauer, M. Kuhn, M. B. Kursa, R. Kutum, N. Lazzarini, I. Lee, M. K. K. Leung, W. K. Lim, C. Liu, F. L. López, A. Mammana, A. Mayr, T. Michoel, M. Mongiovì, J. D. Moore, R. Narasimhan, S. O. Opiyo, G. Pandey, A. L. Peabody, J. Perner, A. Pulvirenti, K. Rawlik, S. Reinhardt, C. G. Riffle, D. Ruderfer, A. J. Sander, R. S. Savage, E. Scornet, P. Sebastian-Leon, R. Sharan, C. J. Simon-Gabriel, V. Stoven, J. Sun, H. Tang, A. L. Teixeira, A. Tenesa, J.-P. Vert, M. Vingron, T. Wang, T. Walter, S. Whalen, Z. Wiśniewska, Y. Wu, G. Xiao, Y. Xie, H. Xu, J. Yang, X. Zhan, S. Zhang, J. Zhao, W. J. Zheng, R. Zhong, D. Ziwei, S. Friend, A. Dearry, A. Simeonov, R. R. Tice, I. Rusyn, F. A. Wright, G. Stolovitzky, Y. Xie, and J. Saez-Rodriguez, "Prediction of human population responses to toxic compounds by a collaborative competition," *Nat Biotechnol*, vol. 33, pp. 933–940, aug 2015.

[17] C. Wang, R. Machiraju, and K. Huang, "Breast cancer patient stratification using a molecular regularized consensus clustering method," *Methods*, vol. 67, no. 3, pp. 304–312, 2014.

[18] C. E. Brodley, "Addressing the selective superiority problem: Automatic algorithm/model class selection," in *Proceedings of the tenth international conference on machine learning*, pp. 17–24, 1993.

[19] J. Rowley, "The wisdom hierarchy: representations of the dikw hierarchy," *Journal of information science*, vol. 33, no. 2, pp. 163–180, 2007.

[20] F. H. Crick, "On protein synthesis," in *Symp Soc Exp Biol*, vol. 12, p. 8, 1958.

[21] Wikipedia, ""lipidomics — wikipedia, the free encyclopedia"," 2017.

[22] A. R. Joyce and B. O. Palsson, "The model organism as a system: integrating'omics' data sets," *Nature Reviews Molecular Cell Biology*, vol. 7, no. 3, pp. 198–210, 2006.

[23] A. L. Swan, A. Mobasheri, D. Allaway, S. Liddell, and J. Bacardit, "Application of machine learning to proteomics data: Classification and biomarker identification in postgenomics biology," *OMICS: A Journal of Integrative Biology*, vol. 17, pp. 595–610, dec 2013.

[24] X. Han, "Neurolipidomics: challenges and developments," *Frontiers in Bioscience*, vol. 12, no. 1, p. 2601, 2007.

[25] I. Mühlberger, J. Wilflingseder, A. Bernthaler, R. Fechete, A. Lukas, and P. Perco, "Computational analysis workflows for omics data interpretation," *Bioinformatics for Omics Data: Methods and Protocols*, pp. 379–397, 2011.

[26] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of research and development*, vol. 3, no. 3, pp. 210–229, 1959.

[27] M. Kubat, R. C. Holte, and S. Matwin, "Machine learning for the detection of oil spills in satellite radar images," *Machine learning*, vol. 30, no. 2-3, pp. 195–215, 1998.

[28] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[29] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, pp. 1137–1145, 1995.

[30] J. G. Moreno-Torres, J. A. Saez, and F. Herrera, "Study on the impact of partition-induced dataset shift on k-fold cross-validation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, pp. 1304–1312, aug 2012.

[31] X. Zeng and T. R. Martinez, "Distribution-balanced stratified cross-validation for accuracy estimation," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 12, pp. 1–12, jan 2000.

[32] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[33] S. Russell, P. Norvig, and A. Intelligence, "A modern approach," *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs*, vol. 25, 1995.

[34] J. R. Quinlan, "C4. 5: Programming for machine learning," *Morgan Kauffmann*, p. 38, 1993.

[35] W. W. Cohen, "Fast effective rule induction," in *Proceedings of the twelfth international conference on machine learning*, pp. 115–123, 1995.

[36] *Generating accurate rule sets without global optimization*, University of Waikato, Department of Computer Science, 1998.

[37] J. HOLLAND, "The possibilities of general-purpose learning algorithms applied to parallel rule-based systems, machine learning," *An Artificial Intelligence Approach*, 1986.

[38] L. Bull, E. Bernadó-Mansilla, and J. Holmes, *Learning classifier systems in data mining*, vol. 125. Springer, 2008.

[39] J. Bacardit, E. K. Burke, and N. Krasnogor, "Improving the scalability of rule-based evolutionary learning," *Memetic Comp.*, vol. 1, pp. 55–67, dec 2008.

[40] V. N. Vladimir and V. Vapnik, "The nature of statistical learning theory," 1995.

[41] M. Law, "A simple introduction to support vector machines," *Lecture for CSE*, vol. 802, 2006.

[42] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[43] M. Cilimkovic, "Neural networks and back propagation algorithm," *Institute of Technology Blanchardstown, Blanchardstown Road North Dublin*, vol. 15, 2015.

[44] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[45] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp. 338–345, Morgan Kaufmann Publishers Inc., 1995.

[46] P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," *Machine learning*, vol. 29, no. 2-3, pp. 103–130, 1997.

[47] K. M. Salama and A. A. Freitas, "Abc-miner: an ant-based bayesian classification algorithm," in *International Conference on Swarm Intelligence*, pp. 13–24, Springer, 2012.

[48] K. M. Salama and A. A. Freitas, "Aco-based bayesian network ensembles for the hierarchical classification of ageing-related proteins," in *European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, pp. 80–91, Springer, 2013.

[49] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[50] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[51] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *European conference on computational learning theory*, pp. 23–37, Springer, 1995.

[52] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297, Oakland, CA, USA., 1967.

[53] E. H. Ruspini, "A new approach to clustering," *Information and control*, vol. 15, no. 1, pp. 22–32, 1969.

[54] R. Agrawal, R. Srikant, *et al.*, "Fast algorithms for mining association rules," in *Proc. 20th int. conf. very large data bases, VLDB*, vol. 1215, pp. 487–499, 1994.

[55] T. A. Kumbhare and S. V. Chobe, "An overview of association rule mining algorithms,"

[56] K. Peason, "On lines and planes of closest fit to systems of point in space," *Philosophical Magazine*, vol. 2, no. 11, pp. 559–572, 1901.

[57] D. Martinus and J. Tax, *One-class classification: Concept-learning in the absence of counterexamples*. PhD thesis, PhD thesis, Delft University of Technology, 2001.

[58] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[59] X. Zhu, M. Gerstein, and M. Snyder, "Getting connected: analysis and principles of biological networks," *Genes & development*, vol. 21, no. 9, pp. 1010–1024, 2007.

[60] M. Bansal, V. Belcastro, A. Ambesi-Impiombato, and D. Di Bernardo, "How to infer gene networks from expression profiles," *Molecular systems biology*, vol. 3, no. 1, p. 78, 2007.

[61] R. L. Nagel, "Epistasis and the genetics of human diseases," *Comptes Rendus Biologies*, vol. 328, pp. 606–615, jul 2005.

[62] M. A. Yıldırım, K.-I. Goh, M. E. Cusick, A.-L. Barabási, and M. Vidal, "Drug-target network," *Nature biotechnology*, vol. 25, no. 10, pp. 1119–1126, 2007.

[63] G. Yu, H. Zhu, C. Domeniconi, and M. Guo, "Integrating multiple networks for protein function prediction," *BMC Systems Biology*, vol. 9, no. 1, p. S3, 2015.

[64] G. Karlebach and R. Shamir, "Modelling and analysis of gene regulatory networks," *Nature Reviews Molecular Cell Biology*, vol. 9, no. 10, pp. 770–780, 2008.

[65] M. Martinez-Ballesteros, I. Nepomuceno-Chamorro, and J. C. Riquelme, "Inferring gene-gene associations from quantitative association rules," in *2011 11th International Conference on Intelligent Systems Design and Applications*, Institute of Electrical & Electronics Engineers (IEEE), nov 2011.

[66] J. G. Rodríguez-Escobedo, C. A. García-Sepúlveda, and J. C. Cuevas-Tello, "KIR genes and patterns given by the a priori algorithm: Immunity for haematological malignancies," *Computational and Mathematical Methods in Medicine*, vol. 2015, pp. 1–11, 2015.

[67] I. A. Nepomuceno-Chamorro, J. S. Aguilar-Ruiz, and J. C. Riquelme, "Inferring gene regression networks with model trees," *BMC Bioinformatics*, vol. 11, no. 1, p. 517, 2010.

[68] I. Nepomuceno-Chamorro, F. Azuaje, Y. Devaux, P. V. Nazarov, A. Muller, J. S. Aguilar-Ruiz, and D. R. Wagner, "Prognostic transcriptional association networks: a new supervised approach based on regression trees," *Bioinformatics*, vol. 27, pp. 252–258, nov 2010.

[69] M. Yoshida and A. Koike, "SNPInterForest: A new method for detecting epistatic interactions," *BMC Bioinformatics*, vol. 12, no. 1, p. 469, 2011.

[70] J. Li, J. D. Malley, A. S. Andrew, M. R. Karagas, and J. H. Moore, "Detecting gene-gene interactions using a permutation-based random forest method," *BioData Mining*, vol. 9, apr 2016.

[71] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts, "Inferring regulatory networks from expression data using tree-based methods," *PLoS ONE*, vol. 5, p. e12776, sep 2010.

[72] S.-H. Chen, J. Sun, L. Dimitrov, A. R. Turner, T. S. Adams, D. A. Meyers, B.-L. Chang, S. L. Zheng, H. Grönberg, J. Xu, *et al.*, "A support vector machine approach for detecting gene-gene interaction," *Genetic epidemiology*, vol. 32, no. 2, pp. 152–167, 2008.

[73] J. Listgarten, S. Damaraju, B. Poulin, L. Cook, J. Dufour, A. Driga, J. Mackey, D. Wishart, R. Greiner, and B. Zanke, "Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms," *Clinical Cancer Research*, vol. 10, no. 8, pp. 2725–2737, 2004.

[74] H. Cho, B. Berger, and J. Peng, "Reconstructing causal biological networks through active learning," *PloS one*, vol. 11, no. 3, p. e0150611, 2016.

[75] M. H. Maathuis, D. Colombo, M. Kalisch, and P. Buhlmann, "Predicting causal effects in large-scale systems from observational data," *Nature Methods*, vol. 7, pp. 247–248, apr 2010.

[76] A. Rau, F. Jaffrézic, and G. Nuel, "Joint estimation of causal effects from observational and intervention gene expression data," *BMC Systems Biology*, vol. 7, no. 1, p. 111, 2013.

[77] R. Li, S. M. Dudek, D. Kim, M. A. Hall, Y. Bradford, P. L. Peissig, M. H. Brilliant, J. G. Linneman, C. A. McCarty, L. Bao, and M. D. Ritchie, "Identification of genetic interaction networks via an evolutionary algorithm evolved bayesian network," *BioData Mining*, vol. 9, may 2016.

[78] K. L. Childs, R. M. Davidson, and C. R. Buell, "Gene coexpression network analysis as a source of functional annotation for rice genes," *PLoS ONE*, vol. 6, p. e22196, jul 2011.

[79] L. Song, P. Langfelder, and S. Horvath, "Comparison of co-expression measures: mutual information, correlation, and model based indices," *BMC bioinformatics*, vol. 13, no. 1, p. 328, 2012.

[80] T. Obayashi and K. Kinoshita, "Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression," *DNA Research*, vol. 16, pp. 249–260, sep 2009.

[81] B. R. Borate, E. J. Chesler, M. A. Langston, A. M. Saxton, and B. H. Voy, "Comparison of threshold selection methods for microarray gene co-expression matrices," *BMC research notes*, vol. 2, no. 1, p. 240, 2009.

[82] P. Langfelder and S. Horvath, "WGCNA: an r package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, no. 1, p. 559, 2008.

[83] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Favera, and A. Califano, "ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC Bioinformatics*, vol. 7, no. Suppl 1, p. S7, 2006.

[84] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner, "Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles," *PLoS biol*, vol. 5, no. 1, p. e8, 2007.

[85] P. E. Meyer, K. Kontos, F. Lafitte, and G. Bontempi, "Information-theoretic inference of large transcriptional regulatory networks," *EURASIP journal on bioinformatics and systems biology*, vol. 2007, no. 1, pp. 1–9, 2007.

[86] A. J. Butte and I. S. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," in *Pac Symp Biocomput*, vol. 5, pp. 418–429, 2000.

[87] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," *Science*, vol. 334, pp. 1518–1524, dec 2011.

[88] D. Albanese, M. Filosi, R. Visintainer, S. Riccadonna, G. Jurman, and C. Furlanello, "minerva and minepy: a c engine for the MINE suite and its r, python and MATLAB wrappers," *Bioinformatics*, vol. 29, pp. 407–408, dec 2012.

[89] C. Huttenhower and O. G. Troyanskaya, "Assessing the functional structure of genomic data," *Bioinformatics*, vol. 24, no. 13, pp. i330–i338, 2008.

[90] I. Lee, S. V. Date, A. T. Adai, and E. M. Marcotte, "A probabilistic functional network of yeast genes," *science*, vol. 306, no. 5701, pp. 1555–1558, 2004.

[91] L. S.J, "Interaction network integration using bayesian data fusion methods," Master's thesis, Newcastle University, 2007.

[92] K. James, A. Wipat, and J. Hallinan, "Integration of full-coverage probabilistic functional networks with relevance to specific biological processes," in *International Workshop on Data Integration in the Life Sciences*, pp. 31–46, Springer, 2009.

[93] J. Weile, K. James, J. Hallinan, S. J. Cockell, P. Lord, A. Wipat, and D. J. Wilkinson, "Bayesian integration of networks without gold standards," *Bioinformatics*, vol. 28, no. 11, pp. 1495–1500, 2012.

[94] T. R. Golub, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531–537, oct 1999.

[95] B. D. W. Group, "Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework," *Clinical Pharmacology & Therapeutics*, vol. 69, pp. 89–95, mar 2001.

[96] P. Xu, *Development of Microarray Genomic Biomarkers and Their Application in Clinical Trials*. ProQuest, 2008.

[97] J. D. Pleil, J. R. Sobus, and D. Smith, "Mathematical and statistical approaches for interpreting biomarker compounds in exhaled human breath," *Volatile Biomarkers: Non-invasive Diagnosis in Physiology and Medicine*, 2013.

[98] D. C. Montgomery, "Design and analysis of experiments," 1991.

[99] E. Robotti, M. Manfredi, and E. Marengo, "Biomarkers discovery through multivariate statistical methods: a review of recently developed methods and applications in proteomics," *Journal of Proteomics & Bioinformatics*, no. S3, p. 1, 2015.

[100] V. Bewick, L. Cheek, and J. Ball, "Statistics review 14: Logistic regression," *Critical Care*, vol. 9, no. 1, p. 112, 2005.

[101] W. S. Noble, "How does multiple testing correction work?," *Nature biotechnology*, vol. 27, no. 12, pp. 1135–1137, 2009.

[102] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[103] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J. Benítez, and F. Herrera, "A review of microarray datasets and applied feature selection methods," *Information Sciences*, vol. 282, pp. 111–135, oct 2014.

[104] M. A. Hall, *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[105] J. Wang, L. Wu, J. Kong, Y. Li, and B. Zhang, "Maximum weight and minimum redundancy: A novel framework for feature subset selection," *Pattern Recognition*, vol. 46, pp. 1616–1627, jun 2013.

[106] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Machine Learning Proceedings 1992*, pp. 249–256, Elsevier BV, 1992.

[107] R. Blanco, P. Larrañaga, I. Inza, and B. Sierra, "Gene selection for cancer classification using wrapper approaches," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 18, no. 08, pp. 1373–1390, 2004.

[108] T. Jirapech-Umpai and S. Aitken, "Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes," *BMC bioinformatics*, vol. 6, no. 1, p. 1, 2005.

[109] L. Li, C. R. Weinberg, T. A. Darden, and L. G. Pedersen, "Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the ga/knn method," *Bioinformatics*, vol. 17, no. 12, pp. 1131–1142, 2001.

[110] R. Ì. N. A. P. B. Maria Fern, Vincent Gardeux, "Ga-kde-bayes: An evolutionary wrapper method based on non-parametric density estimation applied to bioinformatics problems."

[111] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[112] S. Maldonado, R. Weber, and J. Basak, "Simultaneous feature selection and classification using kernel-penalized support vector machines," *Information Sciences*, vol. 181, pp. 115–128, jan 2011.

[113] K.-H. Chen, K.-J. Wang, M.-L. Tsai, K.-M. Wang, A. M. Adrian, W.-C. Cheng, T.-S. Yang, N.-C. Teng, K.-P. Tan, and K.-S. Chang, "Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm," *BMC bioinformatics*, vol. 15, no. 1, p. 1, 2014.

[114] E. Glaab, "Using prior knowledge from cellular pathways and molecular networks for diagnostic specimen classification," *Brief Bioinform*, vol. 17, pp. 440–452, jul 2015.

[115] N. Vlassis and E. Glaab, "GenePEN: analysis of network activity alterations in complex diseases via the pairwise elastic net," *Statistical Applications in Genetics and Molecular Biology*, vol. 14, jan 2015.

[116] E. Lee, H.-Y. Chuang, J.-W. Kim, T. Ideker, and D. Lee, "Inferring pathway activity toward precise disease classification," *PLoS Comput Biol*, vol. 4, p. e1000217, nov 2008.

[117] S. Kim, M. Kon, and C. DeLisi, "Pathway-based classification of cancer subtypes," *Biology Direct*, vol. 7, no. 1, p. 21, 2012.

[118] D. Petrochilos, A. Shojaie, J. Gennari, and N. Abernethy, "Using random walks to identify cancer-associated modules in expression data," *BioData Mining*, vol. 6, oct 2013.

[119] S. Gama-Castro, V. Jiménez-Jacinto, M. Peralta-Gil, A. Santos-Zavaleta, M. I. Peñaloza-Spinola, B. Contreras-Moreira, J. Segura-Salazar, L. Muñiz-Rascado, I. Martínez-Flores, H. Salgado, *et al.*, "Regulondb (version 6.0): gene regulation model of escherichia coli k-12 beyond transcription, active (experimental) annotated promoters and textpresso navigation," *Nucleic acids research*, vol. 36, no. suppl 1, pp. D120–D124, 2008.

[120] T. Schaffter, D. Marbach, and D. Floreano, "Genenetweaver: in silico benchmark generation and performance profiling of network inference methods," *Bioinformatics*, vol. 27, no. 16, pp. 2263–2270, 2011.

[121] G. O. Consortium *et al.*, "Gene ontology consortium: going forward," *Nucleic acids research*, vol. 43, no. D1, pp. D1049–D1056, 2015.

[122] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "Kegg as a reference resource for gene and protein annotation," *Nucleic acids research*, p. gkv1070, 2015.

[123] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using david bioinformatics resources," *Nature protocols*, vol. 4, no. 1, pp. 44–57, 2009.

[124] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic acids research*, vol. 37, no. 1, pp. 1–13, 2009.

[125] J. Reimand, T. Arak, P. Adler, L. Kolberg, S. Reisberg, H. Peterson, and J. Vilo, "g:profiler—a web server for functional interpretation of gene lists (2016 update)," *Nucleic Acids Res*, vol. 44, pp. W83–W89, apr 2016.

[126] H. Mi, A. Muruganujan, J. T. Casagrande, and P. D. Thomas, "Large-scale gene function analysis with the PANTHER classification system," *Nat Protoc*, vol. 8, pp. 1551–1566, jul 2013.

[127] M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, M. G. McDermott, C. D. Monteiro, G. W. Gundersen, and A. Mayan, "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update," *Nucleic Acids Res*, vol. 44, pp. W90–W97, may 2016.

[128] E. Glaab, A. Baudot, N. Krasnogor, R. Schneider, and A. Valencia, "EnrichNet: network-based gene set enrichment analysis," *Bioinformatics*, vol. 28, pp. i451–i457, sep 2012.

[129] M. J. Cowley, M. Pinese, K. S. Kassahn, N. Waddell, J. V. Pearson, S. M. Grimmond, A. V. Biankin, S. Hautaniemi, and J. Wu, "Pina v2. 0: mining interactome modules," *Nucleic acids research*, p. gkr967, 2011.

[130] F. Bastian, G. Parmentier, J. Roux, S. Moretti, V. Laudet, and M. Robinson-Rechavi, "Bgee: Integrating and comparing heterogeneous transcriptome data

among species," in *Lecture Notes in Computer Science* (T, ed.), pp. 124–131, Springer Science Business Media.

[131] G. D. Bader and C. W. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," *BMC bioinformatics*, vol. 4, no. 1, p. 1, 2003.

[132] K. Mitra, A.-R. Carvunis, S. K. Ramesh, and T. Ideker, "Integrative approaches for finding modular structure in biological networks," *Nature Reviews Genetics*, vol. 14, no. 10, pp. 719–732, 2013.

[133] A. J. Cornish and F. Markowetz, "Santa: quantifying the functional content of molecular networks," *PLoS Comput Biol*, vol. 10, no. 9, p. e1003808, 2014.

[134] L. Wadi, M. Meyer, J. Weiser, L. D. Stein, and J. Reimand, "Impact of outdated gene annotations on pathway enrichment analysis," *Nature Methods*, vol. 13, pp. 705–706, aug 2016.

[135] R. J. Xavier and J. D. Rioux, "Genome-wide association studies: a new window into immune-mediated diseases," *Nature Reviews Immunology*, vol. 8, no. 8, pp. 631–643, 2008.

[136] J. Mullen, S. J. Cockell, P. Woollard, and A. Wipat, "An integrated data driven approach to drug repositioning using gene-disease associations," *PloS one*, vol. 11, no. 5, p. e0155811, 2016.

[137] R.-L. Liu and C.-C. Shih, "Identification of highly related references about gene-disease association," *BMC bioinformatics*, vol. 15, no. 1, p. 1, 2014.

[138] À. Bravo, J. Piñero, N. Queralt-Rosinach, M. Rautschka, and L. I. Furlong, "Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research," *BMC bioinformatics*, vol. 16, no. 1, p. 1, 2015.

[139] N. Rappaport, M. Twik, N. Nativ, G. Stelzer, I. Bahir, T. I. Stein, M. Safran, and D. Lancet, "Malacards: A comprehensive automatically-mined database of human diseases," *Current Protocols in Bioinformatics*, pp. 1–24, 2014.

[140] "Opentargets." `http://www.opentargets.org`, 2016.

[141] D. Smedley, A. Oellrich, S. Kohler, B. Ruef, M. Westerfield, P. Robinson, S. Lewis, and C. Mungall, "PhenoDigm: analyzing curated annotations to associate animal models with human diseases," *Database*, vol. 2013, pp. bat025–bat025, may 2013.

[142] N. Rosenthal and S. Brown, "The mouse ascending: perspectives for human-disease models," *Nature cell biology*, vol. 9, no. 9, pp. 993–999, 2007.

[143] A. Hamosh, "Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Research*, vol. 33, pp. D514–D517, dec 2004.

[144] A. P. Davis, C. J. Grondin, K. Lennon-Hopkins, C. Saraceni-Richards, D. Sciaky, B. L. King, T. C. Wiegers, and C. J. Mattingly, "The comparative toxicogenomics databases 10th year anniversary: update 2015," *Nucleic Acids Research*, vol. 43, pp. D914–D920, oct 2014.

[145] M. Magrane and U. Consortium, "UniProt knowledgebase: a hub of integrated protein data," *Database*, vol. 2011, pp. bar009–bar009, mar 2011.

[146] INSERM, "Orphanet: an online database of rare diseases and orphan drugs." `http://www.orpha.net`, 1997.

[147] "Semrep." `https://skr3.nlm.nih.gov`, 2016.

[148] B. Barzel and A.-L. Barabási, "Network link prediction by global silencing of indirect correlations," *Nat Biotechnol*, vol. 31, pp. 720–725, jul 2013.

[149] A. P. Presson, E. M. Sobel, J. C. Papp, C. J. Suarez, T. Whistler, M. S. Rajeevan, S. D. Vernon, and S. Horvath, "Integrated weighted gene co-expression network analysis with an application to chronic fatigue syndrome," *BMC Systems Biology*, vol. 2, no. 1, p. 95, 2008.

[150] M. Ray, J. Ruan, and W. Zhang, "Variations in the transcriptome of alzheimer's disease reveal molecular networks involved in cardiovascular diseases," *Genome Biol*, vol. 9, no. 10, p. R148, 2008.

[151] V. Ransbotyn, E. Yeger-Lotem, O. Basha, T. Acuna, C. Verduyn, M. Gordon, V. Chalifa-Caspi, M. A. Hannah, and S. Barak, "A combination of gene expression ranking and co-expression network analysis increases discovery rate in large-scale mutant screens for novelArabidopsis thalianaabiotic stress genes," *Plant Biotechnology Journal*, vol. 13, pp. 501–513, nov 2014.

[152] Y. Yang, L. Han, Y. Yuan, J. Li, N. Hei, and H. Liang, "Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types," *Nature Communications*, vol. 5, feb 2014.

[153] A. T. Silva, P. A. Ribone, R. L. Chan, W. Ligterink, and H. W. Hilhorst, "A predictive coexpression network identifies novel genes controlling the seed-to-seedling phase transition inArabidopsis thaliana," *Plant Physiology*, vol. 170, pp. 2218–2231, feb 2016.

[154] A. Kommadath, H. Bao, A. S. Arantes, G. S. Plastow, C. K. Tuggle, S. M. Bearson, L. Guan, and P. Stothard, "Gene co-expression network analysis identifies porcine genes associated with variation in salmonella shedding," *BMC Genomics*, vol. 15, no. 1, p. 452, 2014.

[155] S. Uygun, C. Peng, M. D. Lehti-Shiu, R. L. Last, and S.-H. Shiu, "Utility and limitations of using gene expression data to identify functional associations," *PLOS Computational Biology*, vol. 12, no. 12, p. e1005244, 2016.

[156] F. Mordelet and J.-P. Vert, "Prodige: Prioritization of disease genes with multitask machine learning from positive and unlabelled examples," *BMC bioinformatics*, vol. 12, no. 1, p. 1, 2011.

[157] E. Glaab, J. Bacardit, J. M. Garibaldi, and N. Krasnogor, "Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data," *PLoS ONE*, vol. 7, p. e39932, jul 2012.

[158] A. L. Swan, K. L. Hillier, J. R. Smith, D. Allaway, S. Liddell, J. Bacardit, and A. Mobasheri, "Analysis of mass spectrometry data from the secretome of an explant model of articular cartilage exposed to pro-inflammatory and anti-inflammatory stimuli using machine learning," *BMC Musculoskeletal Disorders*, vol. 14, dec 2013.

[159] H. P. Fainberg, K. Bodley, J. Bacardit, D. Li, F. Wessely, N. P. Mongan, M. E. Symonds, L. Clarke, and A. Mostyn, "Reduced neonatal mortality in meishan piglets: A role for hepatic fatty acids?," *PLoS ONE*, vol. 7, p. e49101, nov 2012.

[160] J. Bacardit, P. Widera, A. Marquez-Chamorro, F. Divina, J. S. Aguilar-Ruiz, and N. Krasnogor, "Contact map prediction using a large-scale ensemble of rule sets and the fusion of multiple predicted structural features," *Bioinformatics*, vol. 28, pp. 2441–2448, jul 2012.

[161] G. W. Bassel, E. Glaab, J. Marquez, M. J. Holdsworth, and J. Bacardit, "Functional network construction in arabidopsis using rule-based machine learning on large-scale data sets," *THE PLANT CELL ONLINE*, vol. 23, pp. 3101–3116, sep 2011.

[162] R. Urbanowicz, A. Granizo-Mackenzie, and J. Moore, "An analysis pipeline with statistical and visualization-guided knowledge discovery for michigan-style learning classifier systems," *IEEE Comput. Intell. Mag.*, vol. 7, pp. 35–45, nov 2012.

[163] R. J. Urbanowicz, A. S. Andrew, M. R. Karagas, and J. H. Moore, "Role of genetic heterogeneity and epistasis in bladder cancer susceptibility and outcome: a learning classifier system approach," *Journal of the American Medical Informatics Association*, vol. 20, pp. 603–612, jul 2013.

[164] R. J. Urbanowicz, J. Kiralis, N. A. Sinnott-Armstrong, T. Heberling, J. M. Fisher, and J. H. Moore, "GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures," *BioData Mining*, vol. 5, oct 2012.

[165] D. Baron, A. Bihouee, R. Teusan, E. Dubois, F. Savagner, M. Steenman, R. Houlgatte, and G. Ramstein, "MADGene: retrieval and processing of gene identifier lists for the analysis of heterogeneous microarray datasets," *Bioinformatics*, vol. 27, pp. 725–726, jan 2011.

[166] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, T. S. Ray, M. A. Koval, K. W. Last, A. Norton, T. A. Lister, J. Mesirov, D. S. Neuberg, E. S. Lander, J. C. Aster, and T. R. Golub, "Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nature Medicine*, vol. 8, pp. 68–74, jan 2002.

[167] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. H. Kim, L. C. Goumnerova, P. M. Black, C. Lau, J. C. Allen, D. Zagzag, J. M. Olson, T. Curran, C. Wetmore, J. A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D. N. Louis, J. P. Mesirov, E. S. Lander, and T. R. Golub, "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 415, pp. 436–442, jan 2002.

[168] D. G. Beer, S. L. Kardia, C.-C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. Chen, T. G. Gharib, D. G. Thomas, M. L. Lizyness, R. Kuick, S. Hayasaka, J. M. Taylor, M. D. Iannettoni, M. B. Orringer, and S. Hanash, "Gene-expression profiles predict survival of patients with lung adenocarcinoma," *Nature Medicine*, jul 2002.

[169] L.-l. H. S. R. G. J. E. B. S. R. W. G. R. D. J. S. R. B. Gavin J. Gordon, Roderick V. Jensen, "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," *Cancer Res*, 2002.

[170] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub, and W. R. Sellers, "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, pp. 203–209, mar 2002.

[171] T. Yagi, "Identification of a gene expression signature associated with pediatric AML prognosis," *Blood*, vol. 102, pp. 1849–1856, sep 2003.

[172] D. Chowdary, J. Lathrop, J. Skelton, K. Curtin, T. Briggs, Y. Zhang, J. Yu, Y. Wang, and A. Mazumder, "Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative," *The Journal of Molecular Diagnostics*, vol. 8, pp. 31–39, feb 2006.

[173] E. Jones, T. Oliphant, P. Peterson, *et al.*, "SciPy: Open source scientific tools for Python," 2001–.

[174] P. E. Meyer, F. Lafitte, and G. Bontempi, "minet: A r/bioconductor package for inferring large transcriptional networks using mutual information," *BMC Bioinformatics*, vol. 9, no. 1, p. 461, 2008.

[175] P. D. Thomas, "PANTHER: A library of protein families and subfamilies indexed by function," *Genome Research*, vol. 13, pp. 2129–2141, sep 2003.

[176] N. H. Tran, K. P. Choi, and L. Zhang, "Counting motifs in the human interactome," *Nature Communications*, vol. 4, aug 2013.

[177] P. Shannon, "Cytoscape: A software environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, pp. 2498–2504, nov 2003.

[178] Z. Chen and W. Lu, "Roles of ubiquitination and SUMOylation on prostate cancer: Mechanisms and clinical implications," *IJMS*, vol. 16, pp. 4560–4580, feb 2015.

[179] Y. Gan, C. Shi, L. Inge, M. Hibner, J. Balducci, and Y. Huang, "Differential roles of ERK and akt pathways in regulation of EGFR-mediated signaling and motility in prostate cancer cells," *Oncogene*, vol. 29, pp. 4947–4958, jun 2010.

[180] G. R. Monteith, "Prostate cancer cells alter the nature of their calcium influx to promote growth and acquire apoptotic resistance," *Cancer Cell*, vol. 26, pp. 1–2, jul 2014.

[181] M. Flourakis and N. Prevarskaya, "Insights into ca2+ homeostasis of advanced prostate cancer cells," *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, vol. 1793, pp. 1105–1109, jun 2009.

[182] B. E. Barton, J. G. Karras, T. F. Murphy, A. Barton, and H. F.-S. Huang, "Signal transducer and activator of transcription 3 (stat3) activation in prostate cancer: Direct stat3 inhibition induces apoptosis in prostate cancer lines," *Molecular Cancer Therapeutics*, vol. 3, no. 1, pp. 11–20, 2004.

[183] E. M. Kwon, S. K. Holt, R. Fu, S. Kolb, G. Williams, J. L. Stanford, and E. A. Ostrander, "Androgen metabolism and JAK/STAT pathway genes and prostate cancer risk," *Cancer Epidemiology*, vol. 36, pp. 347–353, aug 2012.

[184] A. Minelli, I. Bellezza, C. Conte, and Z. Culig, "Oxidative stress-related aging: A role for prostate cancer?," *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, vol. 1795, pp. 83–91, apr 2009.

[185] L. Khandrika, B. Kumar, S. Koul, P. Maroni, and H. K. Koul, "Oxidative stress in prostate cancer," *Cancer Letters*, vol. 282, pp. 125–136, sep 2009.

[186] T. Drewa, Z. Wolski, Z. Skok, R. Czajkowski, and H. Wisniewska, "The fas-related apoptosis signaling pathway in the prostate intraepithelial neoplasia and cancer lesions," *Acta poloniae pharmaceutica*, vol. 63, no. 4, pp. 311–315, 2006.

[187] A. G. DiLella, T. J. Toner, C. P. Austin, and B. M. Connolly, "Identification of genes differentially expressed in benign prostatic hyperplasia," *Journal of Histochemistry & Cytochemistry*, vol. 49, pp. 669–670, may 2001.

[188] J. Luo, T. A. Dunn, C. M. Ewing, P. C. Walsh, and W. B. Isaacs, "Decreased gene expression of steroid 5 alpha-reductase 2 in human prostate cancer: Implications for finasteride therapy of prostate carcinoma," *The Prostate*, vol. 57, pp. 134–139, aug 2003.

[189] R. Ribeiro, C. Monteiro, R. Silvestre, A. Castela, H. Coutinho, A. Fraga, P. Principe, C. Lobato, C. Costa, A. C. da Silva, J. M. Lopes, C. Lopes, and R. Medeiros, "Human periprostatic white adipose tissue is rich in stromal progenitor cells and a potential source of prostate tumor stroma," *Experimental Biology and Medicine*, vol. 237, pp. 1155–1162, oct 2012.

[190] V. C. Thompson, T. K. Day, T. Bianco-Miotto, L. A. Selth, G. Han, M. Thomas, G. Buchanan, H. I. Scher, C. C. Nelson, N. M. Greenberg, L. M. Butler, and W. D. Tilley, "A gene signature identified using a mouse model of androgen receptor-dependent prostate cancer predicts biochemical relapse in human disease," *International Journal of Cancer*, vol. 131, pp. 662–672, jan 2012.

[191] N. Sampson, C. Ruiz, C. Zenzmaier, L. Bubendorf, and P. Berger, "PAGE4 positivity is associated with attenuated AR signaling and predicts patient survival in hormone-naive prostate cancer," *The American Journal of Pathology*, vol. 181, pp. 1443–1454, oct 2012.

[192] T. Shiraishi, N. Terada, Y. Zeng, T. Suyama, J. Luo, B. Trock, P. Kulkarni, and R. H. Getzenberg, "Cancer/testis antigens as potential predictors of biochemical recurrence of prostate cancer following radical prostatectomy," *Journal of Translational Medicine*, vol. 9, no. 1, p. 153, 2011.

[193] S. Larsen, T. Yokochi, E. Isogai, Y. Nakamura, T. Ozaki, and A. Nakagawara, "LMO3 interacts with p53 and inhibits its transcriptional activity," *Biochemical and Biophysical Research Communications*, vol. 392, pp. 252–257, feb 2010.

[194] B. S. Taylor, N. Schultz, H. Hieronymus, A. Gopalan, Y. Xiao, B. S. Carver, V. K. Arora, P. Kaushik, E. Cerami, B. Reva, Y. Antipin, N. Mitsiades, T. Landers, I. Dolgalev, J. E. Major, M. Wilson, N. D. Socci, A. E. Lash, A. Heguy, J. A. Eastham, H. I. Scher, V. E. Reuter, P. T. Scardino, C. Sander, C. L. Sawyers, and W. L. Gerald, "Integrative genomic profiling of human prostate cancer," *Cancer Cell*, vol. 18, pp. 11–22, jul 2010.

[195] E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson, Y. Antipin, B. Reva, A. P. Goldberg, C. Sander, and N. Schultz, "The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data," *Cancer Discovery*, vol. 2, pp. 401–404, may 2012.

[196] C. Liu, F. Rohart, P. T. Simpson, K. K. Khanna, M. A. Ragan, and K.-A. L. Cao, "Integrating multi-omics data to dissect mechanisms of DNA repair dysregulation in breast cancer," *Scientific Reports*, vol. 6, p. 34000, sep 2016.

[197] I. Merelli, H. Pérez-Sánchez, S. Gesing, and D. D'Agostino, "Managing, analysing, and integrating big data in medical bioinformatics: Open problems and future perspectives," *BioMed Research International*, vol. 2014, pp. 1–13, 2014.

[198] D. Gomez-Cabrero, I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, E. Ballestar, E. Bongcam-Rudloff, A. Conesa, and J. Tegnér, "Data integration in the era of omics: current and future challenges," *BMC Systems Biology*, vol. 8, no. Suppl 2, p. I1, 2014.

[199] C. E. Cook, M. T. Bergman, R. D. Finn, G. Cochrane, E. Birney, and R. Apweiler, "The european bioinformatics institute in 2016: data growth and integration," *Nucleic acids research*, vol. 44, no. D1, pp. D20–D26, 2016.

[200] R. Bellman, *Dynamic Programming*. Princeton University Press, 1957.

[201] E. Keogh and A. Mueen, "Curse of dimensionality," in *Encyclopedia of Machine Learning*, pp. 257–258, Springer, 2011.

[202] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods," *Bioinformatics*, vol. 26, no. 3, pp. 392–398, 2010.

[203] E. Putin, P. Mamoshina, A. Aliper, M. Korzinkin, A. Moskalev, A. Kolosov, A. Ostrovskiy, C. Cantor, J. Vijg, and A. Zhavoronkov, "Deep biomarkers of human aging: application of deep neural networks to biomarker development," *Aging (Albany NY)*, vol. 8, no. 5, p. 1021, 2016.

[204] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1226–1238, aug 2005.

[205] H. Pang, S. L. George, K. Hui, and T. Tong, "Gene selection using iterative feature elimination random forests for survival outcomes," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, pp. 1422–1431, sep 2012.

[206] J. Bedo, C. Sanderson, and A. Kowalczyk, "An efficient alternative to SVM based recursive feature elimination with applications in natural language processing and bioinformatics," in *Lecture Notes in Computer Science*, pp. 170–180, Springer Science Business Media, 2006.

[207] M. Yousef, S. Jung, L. C. Showe, and M. K. Showe, "Recursive cluster elimination (RCE) for classification and feature selection from gene expression data," *BMC Bioinformatics*, vol. 8, no. 1, p. 144, 2007.

[208] A. L. Swan, D. J. Stekel, C. Hodgman, D. Allaway, M. H. Alqahtani, A. Mobasheri, and J. Bacardit, "A machine learning heuristic to identify biologically relevant and minimal biomarker panels from omics data," *BMC Genomics*, vol. 16, no. Suppl 1, p. S2, 2015.

[209] R. Díaz-Uriarte and S. A. de Andrés *BMC Bioinformatics*, vol. 7, no. 1, p. 3, 2006.

[210] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, October 2011.

[211] S. O'Hara, K. Wang, R. A. Slayden, A. R. Schenkel, G. Huber, C. S. O'Hern, M. D. Shattuck, and M. Kirby, "Iterative feature removal yields highly discriminative pathways," *BMC Genomics*, vol. 14, no. 1, p. 832, 2013.

[212] I. Kononenko, "Estimating attributes: analysis and extensions of relief," in *European conference on machine learning*, pp. 171–182, Springer, 1994.

[213] H. Liu and R. Setiono, "Chi2: feature selection and discretization of numeric attributes," in *Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence*, Institute of Electrical and Electronics Engineers (IEEE).

[214] A. Jaiantilal, G. Grudic, H. Liu, H. Motoda, R. Setiono, and Z. Zhao, "Increasing feature selection accuracy for l1 regularized linear models in large datasets."

[215] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software," *ACM SIGKDD Explorations Newsletter*, vol. 11, p. 10, nov 2009.

[216] G. Kim, Y. Kim, H. Lim, and H. Kim, "An mlp-based feature subset selection for hiv-1 protease cleavage site analysis," *Artificial intelligence in medicine*, vol. 48, no. 2, pp. 83–89, 2010.

[217] S. B. Thrun, J. Bala, E. Bloedorn, I. Bratko, B. Cestnik, J. Cheng, K. De Jong, S. Dzeroski, S. E. Fahlman, D. Fisher, *et al.*, "The monk's problems a performance comparison of different learning algorithms," 1991.

[218] R. Díaz-Uriarte and S. A. De Andres, "Gene selection and classification of microarray data using random forest," *BMC bioinformatics*, vol. 7, no. 1, p. 3, 2006.

[219] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature extraction: foundations and applications*, vol. 207. Springer, 2008.

[220] D. Dembélé, "A flexible microarray data simulation model," *Microarrays*, vol. 2, no. 2, pp. 115–130, 2013.

[221] A. Sboner, F. Demichelis, S. Calza, Y. Pawitan, S. R. Setlur, Y. Hoshida, S. Perner, H.-O. Adami, K. Fall, L. A. Mucci, P. W. Kantoff, M. Stampfer, S.-O. Andersson, E. Varenhorst, J.-E. Johansson, M. B. Gerstein, T. R. Golub, M. A. Rubin, and O. Andrén, "Molecular sampling of prostate cancer: a dilemma for predicting disease progression," *BMC Med Genomics*, vol. 3, mar 2010.

[222] W.-J. Kim, E.-J. Kim, S.-K. Kim, Y.-J. Kim, Y.-S. Ha, P. Jeong, M.-J. Kim, S.-J. Yun, K. Lee, S.-K. Moon, S.-C. Lee, E.-J. Cha, and S.-C. Bae, "Predictive value of progression-related gene classifier in primary non-muscle invasive bladder cancer," *Molecular Cancer*, vol. 9, no. 1, p. 3, 2010.

[223] H. O. Habashy, D. G. Powe, E. Glaab, G. Ball, I. Spiteri, N. Krasnogor, J. M. Garibaldi, E. A. Rakha, A. R. Green, C. Caldas, and I. O. Ellis, "RERG (ras-like, oestrogen-regulated, growth-inhibitor) expression in breast cancer: a marker of ER-positive luminal-like subtype," *Breast Cancer Research and Treatment*, vol. 128, pp. 315–326, aug 2010.

[224] L. Badea, V. Herlea, S. O. Dima, T. Dumitrascu, and I. Popescu, "Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia-the authors reported a combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia," *Hepato-gastroenterology*, vol. 55, no. 88, pp. 2015–2026, 2008.

[225] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, pp. 1–30, Jan. 2006.

[226] A. G. Banerjee, J. Liu, Y. Yuan, V. K. Gopalakrishnan, S. L. Johansson, A. K. Dinda, N. P. Gupta, L. Trevino, and J. K. Vishwanatha, "Expression of biomarkers modulating prostate cancer angiogenesis: Differential expression of annexin ii in prostate carcinomas from india and usa," *Molecular Cancer*, vol. 2, no. 1, p. 34, 2003.

[227] L. Walker, A. C. Millena, N. Strong, and S. A. Khan, "Expression of tgfß and its effects on migratory and invasive behavior of prostate cancer cells: involvement of PI3-kinase/AKT signaling pathway," *Clinical & Experimental Metastasis*, vol. 30, pp. 13–23, jun 2012.

[228] D. M. Altintas, N. Allioli, M. Decaussin, S. de Bernard, A. Ruffion, J. Samarut, and V. Vlaeminck-Guillem, "Differentially expressed androgen-regulated genes in androgen-sensitive tissues reveal potential biomarkers of early prostate cancer," *PLoS ONE*, vol. 8, p. e66278, jun 2013.

[229] I. Guyon, H. Fritsche, P. Choppa, L.-Y. Yang, and S. Barnhill, "A four-gene expression signature for prostate cancer cells consisting of UAP1, PDLIM5, IMPDH2, and HSPD1," *UroToday International Journal*, vol. 02, no. 04, 2009.

[230] D. B. Bernkopf and E. D. Williams, "Potential role of EPB41l3 (protein 4.1B/dal-1) as a target for treatment of advanced prostate cancer," *Expert Opinion on Therapeutic Targets*, vol. 12, pp. 845–853, jun 2008.

[231] P. Kelly, "A role for the g12 Family of heterotrimeric g proteins in prostate cancer invasion," *Journal of Biological Chemistry*, vol. 281, pp. 26483–26490, jun 2006.

[232] Y. Daaka, "G proteins in cancer: The prostate cancer paradigm," *Science Signaling*, vol. 2004, pp. re2–re2, jan 2004.

[233] M. Ammirante, J.-L. Luo, S. Grivennikov, S. Nedospasov, and M. Karin, "B-cell-derived lymphotoxin promotes castration-resistant prostate cancer," *Nature*, vol. 464, pp. 302–305, mar 2010.

[234] J. R. Woo, M. A. Liss, M. T. Muldong, K. Palazzi, A. Strasner, M. Ammirante, N. Varki, A. Shabaik, S. Howell, C. J. Kane, M. Karin, and C. A. Jamieson, "Tumor infiltrating b-cells are increased in prostate cancer tissue," *Journal of Translational Medicine*, vol. 12, no. 1, p. 30, 2014.

[235] V. Hillerdal and M. Essand, "Chimeric antigen receptor-engineered t cells for the treatment of metastatic prostate cancer," *BioDrugs*, vol. 29, pp. 75–89, apr 2015.

[236] G. Bindea, B. Mlecnik, H. Hackl, P. Charoentong, M. Tosolini, A. Kirilovsky, W.-H. Fridman, F. Pages, Z. Trajanoski, and J. Galon, "ClueGO: a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks," *Bioinformatics*, vol. 25, pp. 1091–1093, feb 2009.

[237] G. Rodríguez-Berriguete, B. Fraile, P. Martínez-Onsurbe, G. Olmedilla, R. Paniagua, and M. Royuela, "MAP kinases and prostate cancer," *Journal of Signal Transduction*, vol. 2012, pp. 1–9, 2012.

[238] V. Svetnik, A. Liaw, C. Tong, and T. Wang, "Application of breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules," in *Multiple Classifier Systems*, pp. 334–343, Springer Science and Business Media, 2004.

[239] M. Cross, E. Smith, D. Hoy, S. Nolte, I. Ackerman, M. Fransen, L. Bridgett, S. Williams, F. Guillemin, C. L. Hill, L. L. Laslett, G. Jones, F. Cicuttini, R. Osborne, T. Vos, R. Buchbinder, A. Woolf, and L. March, "The global burden of hip and knee osteoarthritis: estimates from the global burden of disease 2010 study," *Annals of the Rheumatic Diseases*, vol. 73, pp. 1323–1330, feb 2014.

[240] Y. Zhang and J. M. Jordan, "Epidemiology of osteoarthritis," *Clinics in Geriatric Medicine*, vol. 26, pp. 355–369, aug 2010.

[241] R. W. Wright, R. H. Boyce, T. Michener, Y. Shyr, E. C. McCarty, and K. P. Spindler, "Radiographs are not useful in detecting arthroscopically confirmed mild chondral damage," *Clinical Orthopaedics and Related Research*, vol. 442, pp. 245–251, jan 2006.

[242] S. Kluzek, A.-C. Bay-Jensen, A. Judge, M. A. Karsdal, M. Shorthose, T. Spector, D. Hart, J. L. Newton, and N. K. Arden, "Serum cartilage oligomeric matrix protein and development of radiographic and painful knee osteoarthritis. a community-based cohort of middle-aged women," *Biomarkers*, vol. 20, pp. 557–564, nov 2015.

[243] J. Runhaar, C. Sanchez, S. Taralla, Y. Henrotin, and S. Bierma-Zeinstra, "Fibulin-3 fragments are prognostic biomarkers of osteoarthritis incidence in overweight and obese women," *Osteoarthritis and Cartilage*, vol. 24, pp. 672–678, apr 2016.

[244] M. Landsmeer, J. Runhaar, Y. Henrotin, M. M. van, E. Oei, D. Vroegindeweij, M. Reijman, G. van Osch, B. Koes, P. Bindels, and S. Bierma-Zeinstra, "Association of urinary biomarker COLL2-1No2 with incident clinical and radiographic knee OA in overweight and obese women," *Osteoarthritis and Cartilage*, vol. 23, pp. 1398–1404, aug 2015.

[245] S. Kluzek, N. K. Arden, and J. Newton, "Adipokines as potential prognostic biomarkers in patients with acute knee injury," *Biomarkers*, pp. 1–7, may 2015.

[246] A. Mobasheri and Y. Henrotin, "Biomarkers of (osteo)arthritis," *Biomarkers*, vol. 20, pp. 513–518, nov 2015.

[247] U. Ahmed, A. Anwar, R. S. Savage, M. L. Costa, N. Mackay, A. Filer, K. Raza, R. A. Watts, P. G. Winyard, J. Tarr, R. C. Haigh, P. J. Thornalley, and N. Rabbani, "Biomarkers of early stage osteoarthritis, rheumatoid arthritis and musculoskeletal health," *Sci. Rep.*, vol. 5, p. 9259, mar 2015.

[248] B. Ashinsky, C. Coletta, M. Bouhrara, V. Lukas, J. Boyle, D. Reiter, C. Neu, I. Goldberg, and R. Spencer, "Machine learning classification of OARSI-scored human articular cartilage using magnetic resonance imaging," *Osteoarthritis and Cartilage*, vol. 23, pp. 1704–1712, oct 2015.

[249] B. J. Heard, J. M. Rosvold, M. J. Fritzler, H. El-Gabalawy, J. P. Wiley, and R. J. Krawetz, "A computational method to differentiate normal individuals, osteoarthritis and rheumatoid arthritis patients using serum biomarkers," *Journal of The Royal Society Interface*, vol. 11, pp. 20140428–20140428, jun 2014.

[250] J. Runhaar, M. van Middelkoop, M. Reijman, S. Willemsen, E. H. Oei, D. Vroegindeweij, G. van Osch, B. Koes, and S. M. Bierma-Zeinstra, "Prevention of knee osteoarthritis in overweight females: The first preventive randomized controlled trial in osteoarthritis," *The American Journal of Medicine*, vol. 128, pp. 888–895.e4, aug 2015.

[251] R. Altman, G. Alarcon, D. Appelrouth, D. Bloch, D. Borenstein, K. Brandt, C. Brown, T. Cooke, W. Daniel, D. Feldman, *et al.*, "The american college of rheumatology criteria for the classification and reporting of osteoarthritis of the hip," *Arthritis & Rheumatology*, vol. 34, no. 5, pp. 505–514, 1991.

[252] J. Kellgren and J. Lawrence, "Radiological assessment of osteo-arthrosis," *Annals of the rheumatic diseases*, vol. 16, no. 4, p. 494, 1957.

[253] D. Hunter, A. Guermazi, G. Lo, A. Grainger, P. Conaghan, R. Boudreau, and F. Roemer, "Evolution of semi-quantitative whole joint assessment of knee OA: MOAKS (MRI osteoarthritis knee score)," *Osteoarthritis and Cartilage*, vol. 19, pp. 990–1002, aug 2011.

[254] D. Hunter, N. Arden, P. Conaghan, F. Eckstein, G. Gold, A. Grainger, A. Guermazi, W. Harvey, G. Jones, M. H. L. Graverand, J. Laredo, G. Lo, E. Losina, T. Mosher, F. Roemer, and W. Zhang, "Definition of osteoarthritis on MRI: results of a delphi exercise," *Osteoarthritis and Cartilage*, vol. 19, pp. 963–969, aug 2011.

[255] A. Siebuhr, K. Petersen, L. Arendt-Nielsen, L. Egsgaard, T. Eskehave, C. Christiansen, O. Simonsen, H. Hoeck, M. Karsdal, and A. Bay-Jensen, "Identification and characterisation of osteoarthritis patients with inflammation derived tissue turnover," *Osteoarthritis and Cartilage*, vol. 22, pp. 44–50, jan 2014.

[256] L. Nanni, S. Brahnam, N. Lazzarini, and C. Fantozzi, "Heterogeneous ensembles for the missing feature problem," in *2013 annual meeting of the northeast decision sciences institute*, 2013.

[257] J. Alcala-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "KEEL: a software tool to assess evolutionary algorithms for data mining problems," *Soft Computing*, vol. 13, pp. 307–318, may 2008.

[258] K. Napierała, J. Stefanowski, and S. Wilk, "Learning from imbalanced data in presence of noisy and borderline examples," in *Rough Sets and Current Trends in Computing*, pp. 158–167, Springer Science, Business Media, 2010.

[259] D. Li, J. Deogun, W. Spaulding, and B. Shuart, "Towards missing data imputation: A study of fuzzy k-means clustering method," in *Rough Sets and Current Trends in Computing*, pp. 573–579, Springer Science Business Media, 2004.

[260] J. F. Trevor Hastie, Robert Tibshirani, *The Elements of Statistical Learning*. Springer New York, 2009.

[261] A.-C. Bay-Jensen, Q. Liu, I. Byrjalsen, Y. Li, J. Wang, C. Pedersen, D. J. Leeming, E. B. Dam, Q. Zheng, P. Qvist, *et al.*, "Enzyme-linked immunosorbent assay (elisas) for metalloproteinase derived type ii collagen neoepitope, ciim–increased serum ciim in subjects with severe radiographic osteoarthritis," *Clinical biochemistry*, vol. 44, no. 5, pp. 423–429, 2011.

[262] H. J. M. Kerkhof, S. M. A. Bierma-Zeinstra, N. K. Arden, S. Metrustry, M. Castano-Betancourt, D. J. Hart, A. Hofman, F. Rivadeneira, E. H. G. Oei, T. D. Spector, A. G. Uitterlinden, A. C. J. W. Janssens, A. M. Valdes, and J. B. J. van Meurs, "Prediction model for knee osteoarthritis incidence, including clinical, genetic and biochemical risk factors," *Annals of the Rheumatic Diseases*, vol. 73, pp. 2116–2121, aug 2013.

[263] W. Zhang, D. F. McWilliams, S. L. Ingham, S. A. Doherty, S. Muthuri, K. R. Muir, and M. Doherty, "Nottingham knee osteoarthritis risk prediction models," *Annals of the Rheumatic Diseases*, vol. 70, pp. 1599–1604, may 2011.

[264] M. Kinds, A. Marijnissen, K. Vincken, M. Viergever, K. Drossaers-Bakker, J. Bijlsma, S. Bierma-Zeinstra, P. Welsing, and F. Lafeber, "Evaluation of separate quantitative radiographic features adds to the prediction of incident radiographic osteoarthritis in individuals with recent onset of knee pain: 5-year follow-up in the CHECK cohort," *Osteoarthritis and Cartilage*, vol. 20, no. 6, pp. 548 – 556, 2012.

[265] J. I. Galván-Tejada, J. M. Celaya-Padilla, V. Treviño, and J. G. Tamez-Peña, "Multivariate radiological-based models for the prediction of future knee pain: Data from the OAI," *Computational and Mathematical Methods in Medicine*, vol. 2015, pp. 1–10, 2015.

[266] J. Huang, J. J. Burston, L. Li, S. Ashraf, P. I. Mapp, A. J. Bennett, S. Ravipati, P. Pousinis, D. A. Barrett, B. E. Scammell, *et al.*, "Targeting the d-series resolvin receptor system for the treatment of osteoarthritic pain," *Arthritis & Rheumatology*, 2016.

[267] L. Nanni, C. Fantozzi, and N. Lazzarini, "Coupling different methods for overcoming the class imbalance problem," *Neurocomputing*, vol. 158, pp. 48 – 61, 2015.

[268] A. Allahyar and J. de Ridder, "FERAL: network-based classifier with application to breast cancer outcome prediction," *Bioinformatics*, vol. 31, pp. i311–i319, jun 2015.

[269] S. Baron, N. Lazzarini, and J. Bacardit, "Characterising the influence of rule-based knowledge representations in biological knowledge extraction from transcriptomics data," in *EvoBio 2017, Evolutionary Computation, Machine Learning and Data Mining for Biology and Medicine*, 2017.

[270] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, Association for Computing Machinery (ACM), 2016.

[271] H. R. Varian, "Big data: New tricks for econometrics," *Journal of Economic Perspectives*, vol. 28, pp. 3–28, may 2014.